Grounding Dialogue: Eye movements reveal the coordination of attention during conversation and the effects of common ground

Daniel C. Richardson (dcr@ucsc.edu)
Department of Psychology, University of Caifornia, Santa Cruz
273 Social Sciences 2, Santa Cruz, CA 95064

Rick Dale (rad28@cornell.edu)
Department of Psychology, Cornell University,
211 Uris Hall, Ithaca, NY 14853

Abstract

When two people discuss something in front of them, what is the relationship between their eye movements? In Richardson and Dale's (2005) study, participants talked extemporaneously about a TV show while viewing pictures of its cast members. Later, other participants listened to these monologues while viewing the same screen. Cross-recurrence analysis revealed that the coupling between speaker and listener eye-movements predicted how well the listener understood what was said. In our current research, we extended these findings by studying the eye movements of two conversants engaged in a live, spontaneous dialog. The participants talked to each other over the telephone while viewing identical visual displays, and we tracked the eye movements of both conversants simultaneously. In our first study, we found the conversants' eye movements were coupled across several seconds. In the second study we showed that this coupling increases if participants both heard the same background information prior to their conversation. Our results highlight the central role of grounding utterances in the visual context.

Introduction

Coordinating attention across a visual common ground is essential for successful communication (Clark, 1996; Clark & Brennan, 1991; Schober, 1993). In collaborative tasks, conversants readily use gestures, actions and pointing to manipulate each other's attention (Bangerter, 2004; Clark, 2003; Clark & Krych, 2004), and the ability to manipulate joint attention is thought to emerge prelinguistically (Baldwin, 1995). A burgeoning research area has demonstrated that eye movements are tightly linked to the time course of language comprehension (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Brown-Schmidt, Campana, & Tanenhaus, 2004; Hanna & Tanenhaus, 2004; Hanna, Tanenhaus, & Trueswell, 2003; Henderson & Ferreira, 2004; Kamide, Altmann, & Haywood, 2003; Matlock & Richardson, in press; Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995) and language production (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). In the current studies, we used eye movements as a fine-grained index of how two conversants deployed their attention within a visual 'common ground'. This allowed us to investigate the temporal coupling between conversants' eye movements and to examine how this coupling relates to communication.

Monologues and visual common ground

Richardson and Dale (2005) focused on cases in which conversational partners are looking at a visual scene that is the topic of the discussion. The situation is analogous to two people discussing a diagram on a whiteboard, figuring out a route on a map, or talking during a movie. In the first study, the speech and eye movements of one set of participants were recorded as they looked at pictures of six cast members of a TV sitcom (either 'Friends' or 'The Simpsons'). They spoke spontaneously about their favourite episode and characters. One-minute segments were chosen and then played back unedited to a separate set of participants. The listeners looked at the same visual display of the cast members, and their eye movements were recorded as they listened to the segments of speech. They then answered a series of comprehension questions.

Listener and speaker eye movements were coded as to which of the six cast members was being fixated during every 33ms time slice. Cross-recurrence analysis (Zbilut, Giuliani, & Webber, 1998) quantified the degree to which speaker and listener eye positions overlapped at successive time lags (see below for a brief explanation). This speaker X listener distribution of fixations was compared to a speaker X randomized-listener distribution, produced by shuffling the temporal order of each listener's eye movement sequence and then calculating the cross recurrence with the speakers they had heard. This randomized series serves as a baseline of looking 'at chance' at any given point in time, but with the same overall distribution of looks to each picture as the real listeners (see Figure 1).

From the moment a speaker looks at a picture, and for the following six seconds, a listener was more likely than chance to be looking at that same picture. The breadth of this timeframe suggests that speakers and listeners may keep track of a subset of the depicted people who are relevant moment-by-moment (Brown-Schmidt et al., 2004). The overlap between speaker and listener eye movements peaked at about 2000ms. In other words, two seconds after the speaker looked at a cast member, the listener was most likely to be looking at the same cast member. The timing of this peak roughly corresponds to results in the speech production and comprehension literatures. Speakers will fixate objects 800-1000ms (Griffin & Bock, 2000) before naming them, and listeners will typically take 500-1000ms to fixate an object from the word onset (Allopenna et al., 1998). The coupling between speaker and listener eye movements was pervasive, suggesting that planning diverse

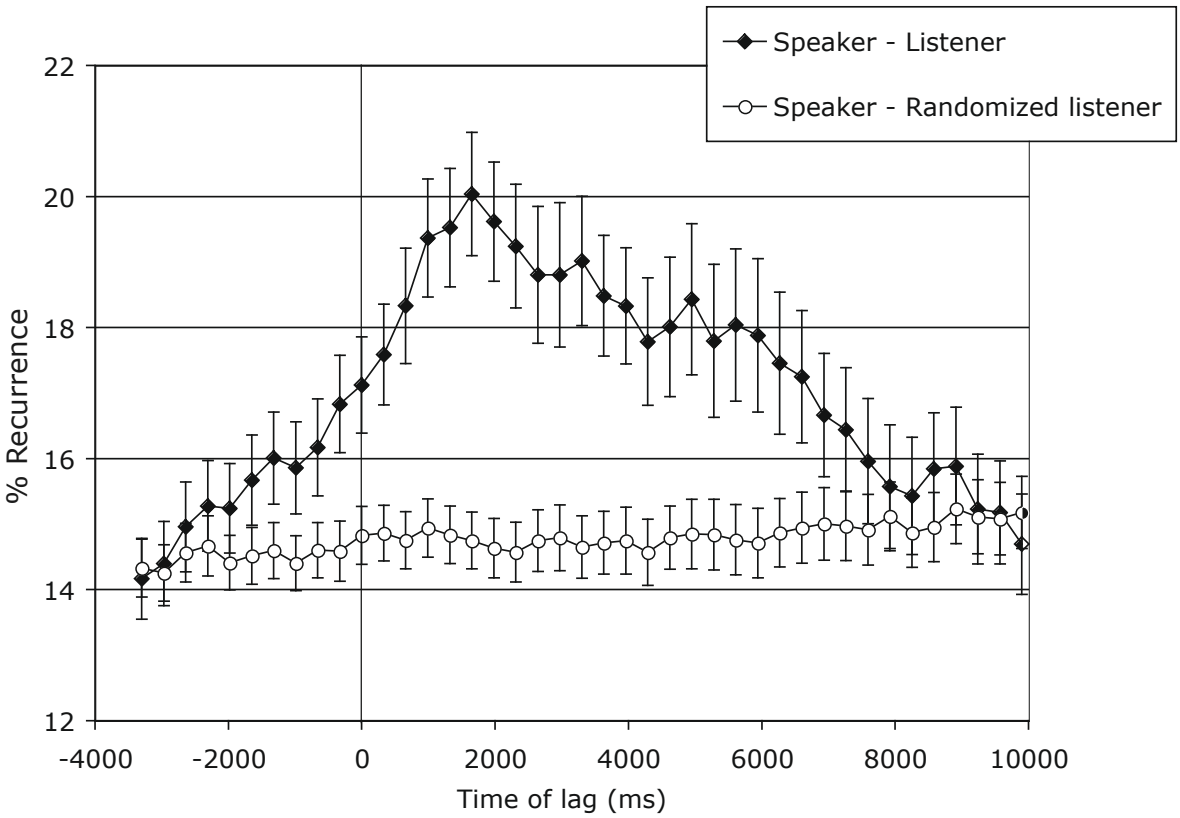

Figure 1. Richardson and Dale (2005). Eye movement recurrence at different time lag intervals in a monologue

types of speech will influence the speaker's eye movements, and a few seconds later, hearing them will influence the listener's eye movements.

Importantly, this coupling of eye-movements between speaker and listener was not merely an epiphenomenal by-product of conversation. It played a functional role in comprehension. When the overall proportion of cross-recurrence between individual speaker-listener pairs was quantified, the strength of the relationship between speaker and listener eye-movement patterns reliably predicted how many of the comprehension questions the listener answered correctly. This correlation was supported by a follow-up study that experimentally manipulated the relationship between speaker and listener eye movements. Examples in visual perception and problem solving (Grant & Spivey, 2003; Pomplun, Ritter, & Velichkovsky, 1996) show that a low-level perceptual cue can cause one person's eye movements to look like another's, and as a consequence, affect their cognitive state. We found that by flashing the pictures in time with the speakers' fixations (or a randomized version) we caused the listeners' eye movements look more (or less) like the speakers', and influenced the listeners' performance on comprehension questions.

Dialogues and visual common ground

Our current studies concern two participants talking spontaneously over the telephone while looking at the same visual display. Both conversants were eye-tracked simultaneously, and the same cross-recurrence tools were used to quantify eye-movement couplings. Participants were given a number of conversational tasks that allowed us to investigate the relationship between visual attention and discourse processes.

Our first study examined the effect that two way interaction would have on the eye movement couplings Richardson and Dale (2005) found with monologue communication. We presented participants with the same pictures of TV cast members and prompted similar conversations. Would the opportunity to interrupt and query a speaker when misunderstandings arise mean that the listener no longer had a need to ground the speaker's words in the visual display? In a dialogue, a listener can also plan and produce her own utterances. Perhaps the eye movement patterns during this frequent alternation of speaker-listener roles would differ from the eye movement couplings of a mute, obedient listener following the words and the gaze of a speaker.

The alternative view is that communication is fundamentally a joint activity (Clark, 1996). This view suggests that communication takes place on the basis of knowledge in the common ground, which includes the visual context that is shared. Therefore, in our dialogue study we will continue to find eye movement couplings, as conversants ground their understanding in the visual scene they have in common. Our second study investigates a further prediction of this view, that increasing the amount of common ground knowledge participants possess will further increase their eye movement couplings.

Experimental methods

Two studies were carried out during a single session with the same pair of participants. We will explain the methods and data analysis techniques employed throughout, and then present the design and results from each study.

Methods

Participants

Forty Stanford undergraduates participated in exchange for course credit. Participants were randomly assigned to pairs. Four pairs were discarded because of problems calibrating the eye tracker to one of the participants. In study 2, an additional two pairs were excluded due to equipment malfunction and experimenter error.

Apparatus

We employed two eye tracking labs on different floors of a building. In the upstairs lab, an ASL 504 remote eye tracking camera was positioned at the base of a 17" LCD stimulus display. Participants were unrestrained, and sat approximately 30" from the screen. The camera detected pupil and corneal reflection position from the right eye, and the eye tracking PC calculated point-of-gaze in terms of coordinates on the stimulus display. This information was passed every 33ms to a PowerMac G4 which controlled the stimulus presentation and collected looking time data. The downstairs lab used an identical set up, apart from the fact that the display was a 36" x 48" foot screen that was back projected and participants sat 80" away (this lab was designed for infants under a year old).

There was an experimenter in each lab operating the eye tracking PC and the Mac running the experiment. The two experimenters communicated to each other using iChat, an instant messaging application. Participants' communicated to each other using the intercom feature on a set of 2.4Ghz wireless phones. Each wore a hands-free headset with headphones and a small boom microphone. The speech of both participants was recorded by microphones at the base of the displays.

Design

Prior to the experimental session, the two experimenters each ran a 9 point calibration routine on their participants, which typically took 1 or 2 minutes. At the beginning of a study, the experimenters agreed upon a time at which to start. This was entered into the Macs. Since each computer was synchronized with an external time server, this ensured that the study trials and data streams began simultaneously.

In each study, the two participants were presented with exactly the same visual display. Regions of interest (ROIs) were predefined for each image.

Data analysis

Our data consisted of two streams of data specifying which (if any) ROI each participant was fixating every 33ms. Our analyses concerned the degree to which the two participants looking at the same thing at the same time. We quantified this question by generating categorical cross-recurrence plots between the speaker and listener time series of fixations (Dale & Spivey, in press; Richardson & Dale, 2005). These plots permit visualization and quantification of recurrent patterns of states between two time series (Shockley et al., 2003, Eckmann et al., 1987; Zbilut & Webber, 1992).

Points of recurrence are simply the times at which the two data streams have the same value; in our case, this means that the two participants' gaze is overlapping and they are fixating the same ROI. For a pair of time series, we can add up all the points recurrence and divide by the total number of possible to get a percentage. In our cross recurrence analysis, one of the data streams is then lagged, so that 0ms on one data stream is aligned with 33ms on the other. Again, all the points of recurrence are calculated. This represents the degree to which one participant is looking at the same thing as the other participant 33ms later. A full cross recurrence analysis consists of calculating the recurrence for all possible alignments, or lag times, of the two data series.

Richardson and Dale (2005) employed this technique on their monologue data to find out exactly what temporal lag between the listener and the speaker would produce the greatest degree of recurrence, or overlap, between the eye movement patterns. Figure 1 shows the average recurrence for 49 dyads at different lag times. As discussed above, this plot reveals that speaker and listener eye movements are coupled at above chance levels from when they are synchronous, up to when the listeners' lag 6000ms behind the speakers'.

Study 1

In the first of our studies, we investigated how the difference between a one way monologue and an interactive dialogue would play out in the eye movement couplings of the participants. The task and stimuli were identical to Study 1, Richardson and Dale (2005). Participants saw a picture of six cast members from the sitcom *Friends* or *The Simpsons.* (Figure 2) The participants were asked to discuss their favourite characters or episodes from the show. These were the same prompts used to elicit monologues from the speakers in Richardson and Dale (2005). The participants were allowed to say as much as they liked, but typically, conversations lasted for 1 to 5 minutes.

In the original monologue study there was a peak of recurrence when the listeners' eye movements followed the speakers at a lag of roughly 2000ms. We hypothesized that in this dialogue study there would be a similar peak in eye movement recurrence, reflecting a similar process of grounding language in the visual context. We predicted that this peak would be centered around 0ms on average, since this would reflect the that fact the participants would take turns in speaking, and consequently, in leading the eye movement coordination.

Results and discussion

Figure 2 shows the average recurrence between participants' eye movements at different time lags, averaged over 16 dyads. As in Figure 1, the randomized baseline provides a comparison of looks that are distributed equally to participants' eye movement, but have had the temporal structure removed. And as in Figure 1, there is a window of roughly six seconds in which participants eye movements

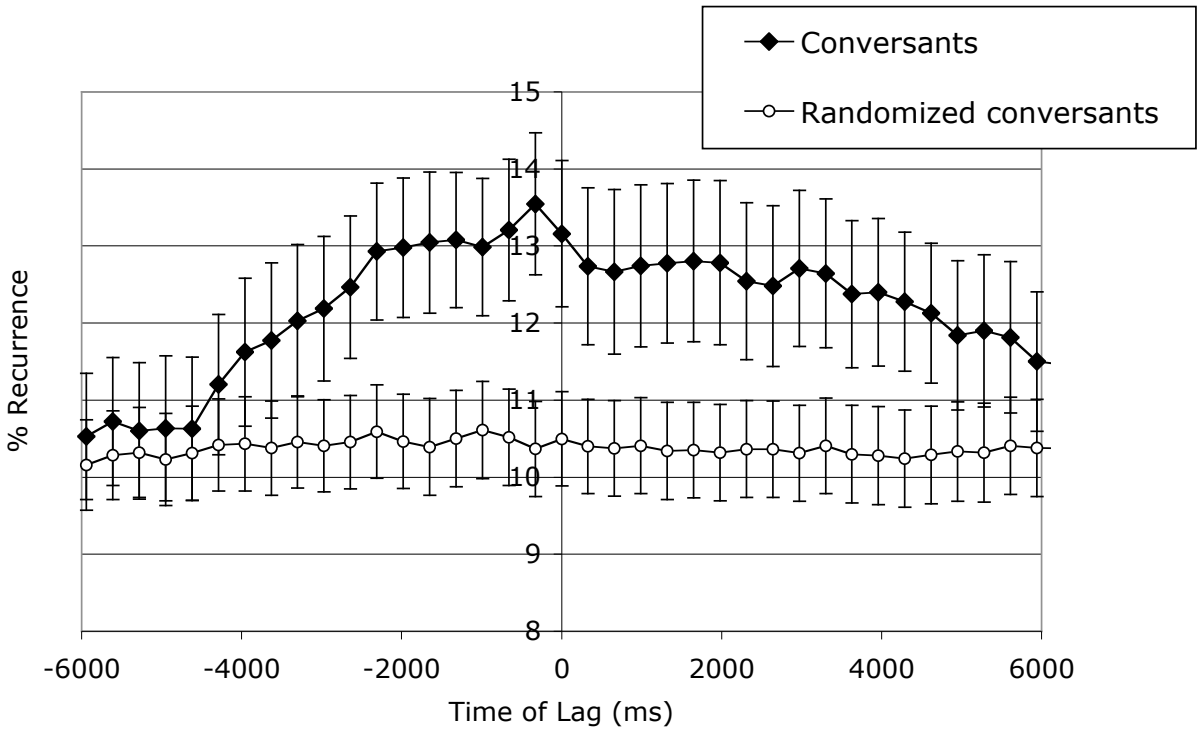

Figure 2. Eye movement recurrence at different time lag intervals in a dialogue, Study 1.

are clearly coupled at above chance levels. Unlike the monologue results though, in this dialogue data the peak in recurrence occurs at around 0ms.

The differences between the dialogue and Richardson and Dale's (2005) monologue data were demonstrated by analyzing the results from the two experiments together. A 2 (monologue/dialogue) x 41 (lag times) mixed-effects ANOVA (lag as a repeated-measures factor) revealed a significant main effect of experiment ($F(1,87)=20.5$, $p<.001$) and a main effect of lag ($F(40,3480)=8.3$, $p<.001$). Most importantly, there was a significant interaction between the factors ($F(40,3480)=4.2$, $p<.001$), showing that the two way interaction in the dialogue experiment changed the temporal structure of the eye movement coupling.

Though perhaps not surprising, the results from this first study support our hypothesis that the eye movement coupling found in monologue communication extends to dialogues. Even though in this case participants were able to verbally interact with each other, and could make use of all the common verbal back channels in communication that signal assent, understanding, or a need for more information (Clark, 1996), participants were still visually coordinating their attention as they conversed.

Study 2

The term 'common ground' refers to much more than the visual context shared by conversational participants. It also describes the many beliefs, opinions and facts that conversants share (Clark, 1996; Lee 2001). In the second study we tested the hypothesis that manipulating the amount of common ground in knowledge between participants would affect their ability to coordinate their attention in the visual common ground.

Participants were required to talk about a painting by Salvador Dali (Figure 3). Prior to their conversation, they were told that they would each hear a short discussion of Dali's art. They were informed that they would either be hearing the same information, or that they would each hear different information. Accordingly, the participants then listened to 90 second passages that discussed either the history, content and meaning of the specific painting (e.g., "the still life objects in the original canvas have separated from the table and float in the air, and even the particles of paint have broken loose from the canvas"), or Dali's personality and theory (e.g. "the paranoiac critical method entailed the creation of a visionary reality from elements of dreams, memories and psychological or pathological distortions. At times Dali would stand on his head to induce hallucinations."). As we discuss below, the conditions varied in that participants both believed they heard same/different information, and actually heard same/different information.

Once more, the participants were allowed to talk for as long as they required, during which time their gaze was recorded. ROIs were defined on Dali's painting which corresponded to six of the main objects or elements. Our prediction was that pairs of participants who had heard the same information about Dali would have a higher recurrence between their eye movements than those who heard different passages.

Results and discussion

For each of our dyads, we quantified the amount of recurrence within a window of +/- 3000ms. In other words, we looked at the overlap between participants' eye movements when they lagged each other by up to 3000ms. A window of this size was chosen because in Richardson and Dale (2005), study 1 and study 2 above, participants' eye movements were coupled at above chance levels in a roughly six second window. By restricting our analysis to this window, we focus on times when the eye movements are indeed coupled, and look specifically at the effects of the common ground manipulation.

A one way ANOVA was performed on the average recurrence in each dyad within a window of +/- 3000ms. There was a significant effect of common ground condition ($FI(1,12)=4.9$. $p<.05$), such that dyads who heard the same information had recurrence levels over a third higher than those who heard different information (Figure 4). We conclude that a simple manipulation changing the information participants share about a painting directly affects the coordination of their visual attention.

Conclusion

In spontaneous, natural dialogue relating to a common visual scene, conversants' visual attention is tightly coupled. This conclusion was suggested by Richardson and Dale's (2005) experiments on the causal role of eye movement couplings during communication between a speaker and listener. Their paradigm, however, excluded one of the most important features of verbal communication: two-way interaction. The present studies provide a demonstration and quantification of the eye movement couplings during interactive verbal communications. The recurrence, or overlap, between eye movement series was greatest when the series were aligned at 0ms, but was at above chance with a lag of +/- 3000ms. In other words, the conversants were most likely to be looking at the same thing when one examines the same point in time in both their eye movement recordings. However, if one picked any two points in their eye movement recordings that were within 3000ms, then they would be more like than chance to be looking at the same thing.

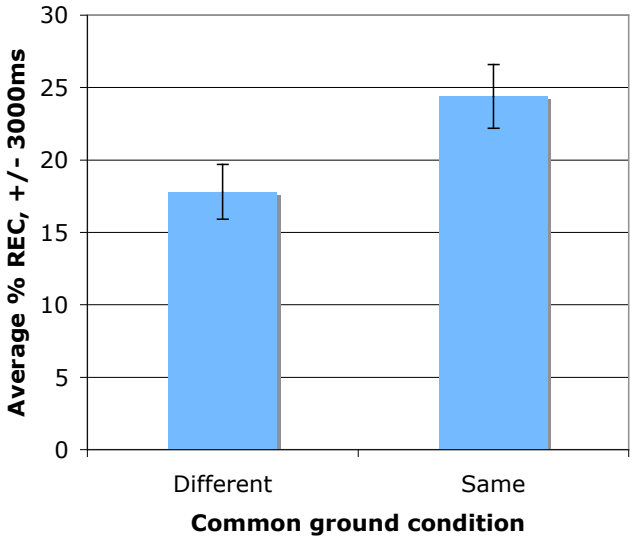

Figure 4. Average recurrence by common ground, Study 2

Figure 3. *Nature Morte Vivante* by Salvador Dali

Interestingly, this eye movement coupling is sensitive to the knowledge that conversants have prior to their conversation. If they each hear the same background information, rather than two different passages, then their subsequent eye movements have a significantly tighter coupling with each other. This result provokes several interesting hypotheses which are the subject of our ongoing research. Firstly, it could be that the shared information given to subjects supplies a vocabulary, or way in which participants can refer to elements of the picture. Further experiments are addressing this issue by drawing on the notion of 'conceptual pacts' (Clark & Brennan, 1991) and eye tracking participants during tasks where they generate novel referring expressions. Secondly, is the advantage of our same condition due solely to the fact that participants actually know the same information, or is it also important that they know that they each know the same information? Clark (1996) would suggest the latter, and since the current study conflates these these two possibilities, they will be contrasted in future experiments.

In all of our studies, eye movement couplings reveal an intimate relationship between discourse processes and visual attention. Just as eye movements reflect the mental state of an individual, the coupling between a speaker's and a listener's eye movements reflects the success of their communication. We conclude that looking around the common ground in step with each other is part of the process of mutual understanding.

Acknowledgments

The authors are indebted to Herbert H. Clark, Natasha Kirkham, Michael Ramscar and Michael Spivey for many inspiring discussions, and the members of the Kirkham Learning Lab who donated their time and talents to eye tracking: Lisa Smythe, Carl Dambkowski, Sasha Filippova, Debbie Kim, Lauren Rimoin and Rosemary Reidy. Rick Dale was supported by a Paller-Dallenbach Fellowship from Cornell University.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language.* 38(4), 419-439.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: its origins and role in development.* Hillsdale, NJ: Lawrence Erlbaum.

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

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2004). Real-time reference resolution by naïve participants during a task-based unscripted conversation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *World-situated language processing: Bridging the language as product and language as action traditions.* Cambridge: MIT Press.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243-268). Mahwah, NJ: Lawrence Erlbaum.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. J*ournal of Memory & Language, 50(1),* 62-81.

Dale, R. & Spivey, M.J. (in press). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning.*

Eckmann, J. P., Kamphorst, S. O., & Ruelle, D. (1987). Recurrence lots of dynamical systems. *Europhysics Letters,* 5, 973-977.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5), 462-466.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science,* 11(4), 274-279.

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

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

Henderson, J. M., & Ferreira, F. (Eds.). (2004). The integration of language, vision, and action: *Eye movements and the visual world*. New York: Psychology Press.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory & Language, 49(1)*, 133-156.

Lee, B. H. P. (2001) Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground, *Journal of Pragmatics*, 33, pp 21-44.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science,* 4 (6-14).

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2), B25-B33.

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25(8), 931-948.

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, 1045–1060.

Richardson, D.C & Matlock, T. (in press). The integration of figurative language and static depictions: An eye movement study of fictive motion. *Cognition*

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(5217), 1632-1634.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, 47(1), 1-24.

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

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

Zbilut, J. P., & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters* A, 171, 199-203.