Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and its Relationship to Discourse Comprehension

Daniel C. Richardson (richardson@psych.stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305, USA

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

Abstract

While their eye movements were being recorded, participants spoke extemporaneously about a TV show whose cast members they were viewing. Later, other participants listened to these speeches while their eyes were tracked. Within this naturalistic paradigm using spontaneous speech, a number of results linking eye movements to speech comprehension, speech production and memory were replicated. More importantly, a cross-recurrence analysis demonstrated that speaker and listener eye movements were coupled, and that the strength of this relationship positively correlated with listeners' comprehension. Just as the mental state of a single person can be reflected in patterns of eye movements, the commonality of mental states that is brought about by successful communication is mirrored in a similarity between speaker and listener's eye movements.

Introduction

Imagine standing in front of a painting, discussing it with a friend. As you talk, your eyes will scan across the image, moving approximately three times a second. They will be drawn by characteristics of the image itself, areas of contrast or detail, as well as features of the objects or people portrayed. Eye movements are driven both by properties of the visual world and processes in a person's mind. Your gaze might also be influenced by what your friend is saying, what you say in reply, what is thought but not said, and where you agree and disagree. If this is so, what is the relationship between your eye movements and those of your friend? How is that relationship related to the flow of conversation between you?

Language use often occurs within rich visual contexts such as this, and the interplay between linguistic processes and visual perception is of increasing interest to psycholinguists and vision researchers (Henderson & Ferreira, 2004). As yet, however, such processes have been limited to experiments that examine the eye movements of the speaker or the listener in isolation. Language use, more often than not, occurs within a richer social context as well.

Direct eye contact between conversants plays an interesting, crucial role in coordinating a conversation (Bavelas, Coates, & Johnson, 2002), and in conveying various attitudes or social roles (Argyle & Cook, 1976). The focus of the current experiment, however, is cases such as those introduced at the outset, where conversants are not looking at each other, but at some visual scene that is the topic of the conversation. More common examples might be discussing a diagram drawn on a whiteboard, figuring out together how to do something on a computer, or talking during a movie.

Uniquely poised between perception and cognition, eye movements can reveal cognitive processes such as speech planning, language comprehension, memory, mental imagery and decision making. The current experiment investigates whether the eye movements of a speaker and a listener to a visual common ground can provide insight into a discourse.

Eye movement Research

Eye movements of a speaker

If a speaker is asked to describe a simple scene, they will fixate the objects in the order in which they are mentioned, around 900ms before naming them (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). Since such pictures can be identified rapidly, it is argued that during this time speakers are not just retrieving words but selecting and planning which to use.

Eye movements of a listener

Eye movement research has shown that there is a tight interdependence between speech recognition and visual perception. Eye movements to potential referents for a word can provide evidence for a lexical item being recognized before the word is finished being spoken. The link between visual and linguistic processing can also be seen in eye movements that disambiguate syntactic structures (Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995) and anticipate the future agents of actions (Kamide, Altmann, & Haywood, 2003). Recent studies of the eye-movements of a participant engaged in a conversation with another naïve participant reveal a remarkable sensitivity to the referential domains established by the task, the visual context and the preceding conversation (Brown-Schmidt, Campana, & Tanenhaus, 2004). Qualitatively, eye movement research reveals a very close, time-locked integration between visual and linguistic processing (Tanenhaus, Magnuson, Dahan, & Chambers, 2000). Although fixation times are heavily modulated by context, as a very rough quantitative guide, research suggests that listeners will fixate an object around 400-800ms after the name onset.

Eye movements of a thinker

Since participants will make systematic eye movements to entirely empty and uninformative regions of space when retrieving information from memory (Richardson & Kirkham, in press; Richardson & Spivey, 2000) or listening to a story (Spivey & Geng, 2001) it is clear that they can be governed by cognitive as well as perceptual processes. Influencing how the eye moves across an image can have profound effects on mental processes. Researchers have recorded the eye movements of participants interpreting an ambiguous picture in a particular way, or solving a difficult deductive problem from a diagram. Using low level visual cues, a second set of participants were then influenced to attend to the same regions of the picture. The second set of participants were more like to form the same interpretation of the ambiguous picture (Pomplun, Ritter, & Velichkovsky, 1996), and remarkably, were more likely to solve the deductive problem (Grant & Spivey, 2003). If forced similarity between participants' eye movements can result in similar cognitive states, then will the similar cognitive states that are brought about by successful verbal communication result in similar eye movement patterns between speaker and listener?

Experiment

Speech production and speech comprehension have previously been studied in separate eye tracking paradigms. Yet if both are indeed closely linked to eye movements, the eye movement patterns of two people engaged in a natural, unscripted conversation may bare some relationship to each other. Moreover, it raises the intriguing possibility that the strength of the relationship between conversants' eye movements will parallel the success of their linguistic relationship.

The current experiment approximates a conversation between naïve conversants by asking participants to speak spontaneously, with neither a script nor a rehearsal for an extended period of time about a TV show, whose characters were displayed in front of them. These speeches were then played back to other participants who were looking at the same display. Crucially, both the speakers' and the listeners' eye movements were tracked throughout. The listeners' comprehension was then measured by a series of content questions. Thus in addition to extending various eye movement-language results to natural, spontaneous speech, the current experiment was able to investigate a number of entirely novel hypotheses regarding the linkage between speaker and listener eye movements, and its relation to the listener's comprehension.

Methods

The first four participants recruited to take part in this experiment were designated as speakers, and the remainder were listeners. The methods for both stages will be described below.

Participants

40 Stanford undergraduates took part in the experiment in exchange for course credit.

Apparatus

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 co-ordinates on the stimulus display. This information was passed every 33ms to a PowerMac G4 which controlled the stimulus presentation and collected looking time data. Prior to the experimental session, the participants went through a 9 point calibration routine, which typically took between 2 and 5 minutes.

Speakers' voices were recorded by microphone, and listeners made responses using the two buttons of a mouse held in their lap.

Design – Speakers

The intention was to record participants speaking spontaneously about a TV show while looking at a picture of the cast members. In the first case, a picture of the 6 principal characters of the cast of the TV sitcom *Friends* was used. The characters were shown individual in 6 separate pictures. Potential speakers were asked if they knew they show and would like to talk about it, and two speakers were selected who were knowledgeable and reasonably gregarious. Speakers were instructed to 'Talk about the show for a couple of minutes. You could talk about the relationships between the characters, your opinion of them, or your favourite episode'. In the second case, two participants were shown a 5 minute scene from *The Simpsons* during which they undergo family therapy. These participants were then shown a picture of the five family members and their therapist. The participants were asked to 'Describe what went on in the scene and what you thought about it'.

As they spoke, the speakers' eye movements were tracked and their voices were recorded by microphone. These recordings were trimmed so that they were all roughly one minute long, and the text was transcribed for later analysis.

Design - Listeners

Participants listened while looking at the same picture of the six cast members that had been in front of the speaker. Since there could not be systematic looks to the cast members if the participant did not recognize any of them, participants were first asked if they were familiar with either show. On this basis, the listeners were presented with one or both of the *Friends* and *Simpsons* stimuli, and were randomly assigned one of the two speakers.

Listeners heard a minute of speech, and then a screen appeared warning them that the question period was about to start. In the four question trials, participants saw six solid grey circles or squares in the locations where pictures of the individual cast members had previously appeared. After a 1000ms pause, they heard a question and responded yes or no using the two mouse buttons. There followed a 2000ms ISI during which the screen was blank.

The questions were recorded by the experimenter and were of the form, “Did the speaker say…?”. The questions were designed such that they could not be answered on the basis of knowledge about *Friends* or *The Simpsons* alone, but were specific to the information mentioned (or not) by that particular speaker. The correct answer to half the questions was yes and half no.

Data Coding

Roughly half of our listeners were familiar with both TV shows and half knew the characters from only one. All analyses are based on 49 usable listener-speaker dyads. A further 9 cases were dropped due to problems with the equipment or the calibration procedure.

The eye movements of the speaker and of the listener during the minute of speech were analyzed in exactly the same way. The eye movement data relayed which, if any, of the six pictures were being fixated every 33ms. The data were cleaned for blinks and saccades across a picture - only stable fixations longer than 99ms were analyzed – and then expressed in terms of a sequences of gaze onsets and offsets in the six pictures.

The speakers’ recordings were transcribed with onset times for each word spoken. In addition, words were flagged if they were names of any of the six characters pictured. Listener responses to the questions were coded for accuracy, and their looking times to each of the pictures while answering were calculated.

Results and Discussion

This experiment provided precise timing information about speakers’ speech and gaze onsets, and listeners’ gaze onsets. This information can depicted graphically in what we call a ‘scarf plot’, which represents a transcript of the speech together with the timing of word onsets and the eye movements of both speaker and listen. Figure 1 shows an nine second segment of a scarf plot for one speaker-listener dyad. Such eye movement data can be statistically analyzed and compared with the objective measure of the listeners’ understanding of the speech provided by their performance answering four comprehension questions.

Before the detailed inferential analyses begin, it is useful to get a rough sense of the behavior being studied. On average, speakers used 160 words, only 12 of which were the names of the characters depicted. It is important to note that the speeches were not edited for content, and include all the deviations, hesitations and repetitions that are typical of just a minute of normal, spontaneous speech.

Speakers and listeners switched their gaze between pictures around 120 times. For each occasion, they spent about 500ms looking at the picture. Since the average eye fixation lasts 200-300ms, it is reasonable to assume that this represents two fixations within the same picture.

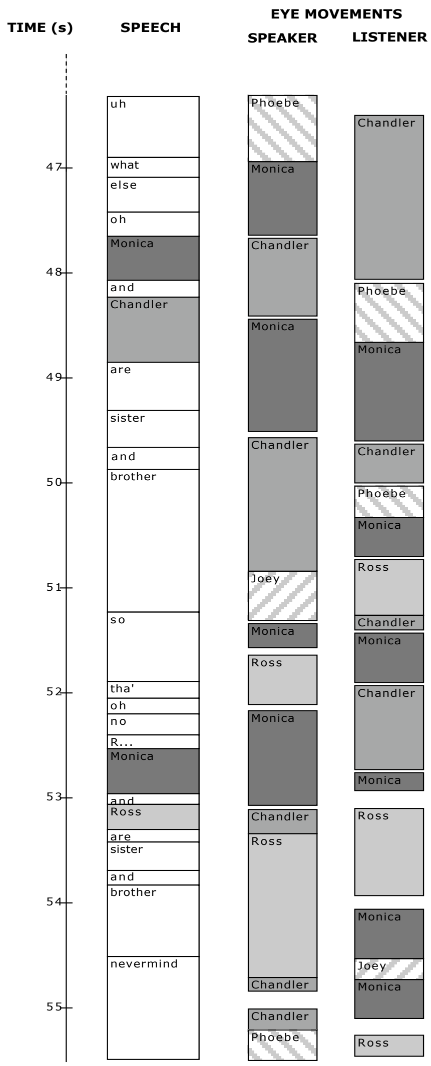

Figure 1. Scarf Plot of a 9 second segment of one dyad. The speaker’s words are shown on the left, with nouns highlighted. The speaker’s and listener’s eye movements are shown in the middle and right columns respectively. Time is on the y axis, increasing down the page.

Speaker Fixations Prior To Naming

For each occasion that the speaker named character X, their eye movement data were consulted to find the point at which X was last previously fixated. The difference between the gaze onset and the name onset was computed for every name used by every speaker. On average, a character was fixated 860 ms prior to being named.

This lag is exactly in the range reported by the speech production literature (Griffin & Bock, 2000), where typically participants are explicitly instructed to describe a simple picture. We have found a lag of the same magnitude with spontaneous, natural speech, when participants are describing not what is front of them per se, but things that are not depicted - stories, opinions, relationships – that relate to those characters.

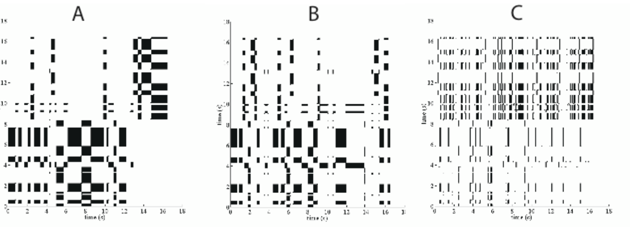

Figure 2. Example CRPs

Relationship Between Speaker And Listener Eye Movements

To what degree were speaker and listener 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 submission). These plots permit visualization and quantification of recurrent patterns of states between two time series (see Shockley, Santana & Fowler, 2003, for a fuller introduction; see Eckmann, Kamphorst & Ruelle, 1987; Zbilut & Webber, 1992 for foundational treatises). In our case, the cross-recurrence plot portrays the extent to which dyad fixations are overlapping temporally.

To begin, windows of a given length are moved along each time series, forming individual windows at every time index. The windows of each time series are then compared to *all* those of the other time series (comparing *every* time index). At time index *i* for the first time series and *j* for the second, if their windows are sufficiently similar, a point *(i, j)* is recorded on a two-dimensional plot. By comparing every window in the first to the second time series, we can generate a full plot of points in which the two time series are close to each other – a cross-recurrence plot.

For simplicity, we used a window size of 1 for our analysis. By using a categorical metric (see Dale & Spivey, in submission, for details), we have the criterion that dyad fixations are recurrent if falling on the same object for 33ms. We generated plots using this metric between every speaker-listener pair. Figure 2 shows example cross recurrence plots between a speaker and (a) a listener who answered all comprehension questions correctly (b) a listener who answered few correctly, and (c) a listener with their eye movement data placed in a random order. There are three things to notice here. Firstly, the good listener has higher density in their plot, indicating more points of recurrence with the speaker. Secondly, both listeners have more structured plots compared to the randomized series. Lastly, one can see that for the two real listeners there is a higher density in the region on and below the *i=j* diagonal. This indicates that the speaker and listeners' eye movements overlapped more when the listeners' eye movements lagged behind the speakers.

We employed a further analysis to find out exactly what temporal lag between the listener and the speaker would produce the greatest degree of recurrence, or overlap, between their eye movement patterns. Listener time series were successively lagged by 330ms. On the line defined by *i = j* in the plot (the *line of incidence*), any point indicates that *in the same temporal context* fixations are recurrent. Thus, by lagging the listeners' time series, and recording maximal recurrence along the line of incidence within each lag, we get a measure of the extent to which dyads' eye movements are related. Though our chosen window size is small, the results are quite compelling.

Figure 3 shows the degree of recurrence between speaker and listener at different time lags, averaged across all 49 dyads. We also randomized listeners' eye movement data and calculated its recurrence with the speakers'. 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.

A 2 (listeners/randomized listener) x 40 (lag times) ANOVA revealed a significant main effect of listener type (F(1,45)=785.5, p<.0001) and a main effect of lag (F(40,1800)=25.2, p<.001). Moreover, there was a significant interaction between the factors (F(40,1800)=24.7, p<.001).

Clearly, the real listeners are not looking around these displays randomly. Rather their eye movements are linked to the speakers', and this relationship has a temporal character. More precisely, the maximum recurrence between the speakers and listeners, the lag time at which their eye movements overlap the most, is at 1650ms

These results are exactly what one would expect from the combination of the speech production and speech comprehension eye movement literature. Typically, speakers will fixate an item 900ms before naming it and listeners will fixate an object around 800ms after the name onset. Very roughly this would suggest we would find a lag of 900+800=1700ms between speaker gaze onsets and listeners'. This derived value corresponds both to the exact lag that produces a maximum recurrence value, 1650ms, and the general region of higher recurrence in the 1000-2000ms range.

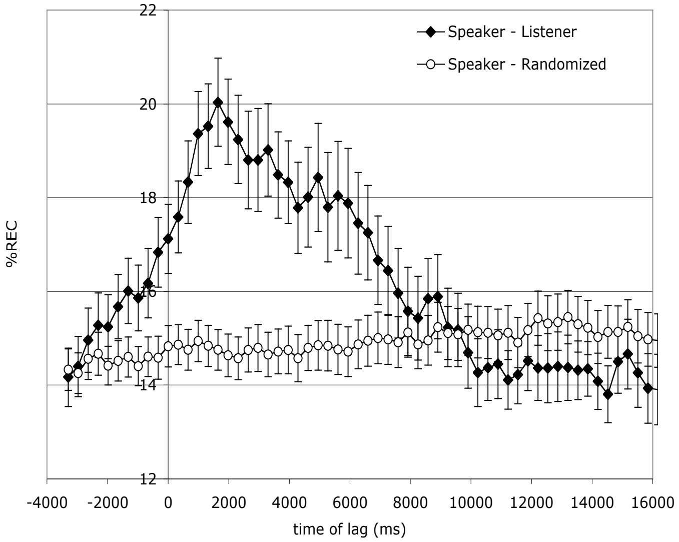

Figure 3. Cross recurrence at different time lags

The speech production and comprehension literatures, however, deal with cases where an object or person is explicitly named. Perhaps it is the case that the differences between critical and non-critical gaze onset lag distributions observed here are due mainly to the occasions when the speaker planned and spoke out loud a name of one of the characters pictured.

This question was addressed by examining a subset of the data. The name-subset includes only speaker fixations to person X that were immediately prior to the speech onset of name X. As noted previously, since there were on average 12 cases of name use, this constituted about 10% of the 120 fixations the average speaker made.

Figure 4A plots the recurrence at different time lags for the name subset of our data. The 2 (listeners/randomized listener) x 40 (lag times) ANOVA revealed a significant main effect of listener type (F(1,45)=192.3, p<.0001) and a main effect of lag (F(40,1800)=28.1, p<.001). As before, there was a significant interaction between the factors (F(40,1800)=27.5, p<.001).

For the subset of speaker fixations that precede a name, there is a highly pronounced difference between the speaker and the listener and the speaker and randomized looking. Once more, the greatest extent of this difference is just before 2000ms. Again, this would be predicted by the speech production and comprehension eye movement literatures. Is it the case, then, that the current experiment has simply replicated these name-use results using spontaneous speech?

To answer this question the data excluded from the name analysis above were analyzed in isolation. Figure 4B plots the 'non name dataset' that corresponds to the 90% of speaker fixations to person X which were not immediately followed by X being named out loud. The ANOVA showed a similar pattern of results: main effect of listener type (F(1,45)=559, p<.0001), a main effect of lag (F(40,1800)=25.8, p<.001). and a significant interaction between the factors (F(40,1800)=25.0, p<.001). Although subtracting the cases of name use from the full data set appeared to attenuate somewhat the differences between critical and non-critical gaze onset lags, it is certainly the case that these distributions still differ. In other words, it is not just the when the speaker names a character that speaker and listener eye movements are linked. It must be other properties of the discourse (implicit reference, anaphor, topics, agents, for example) which drive the speakers eye movements while they are being planned, and a few seconds later, influence the listener's eye movements once they are spoken.

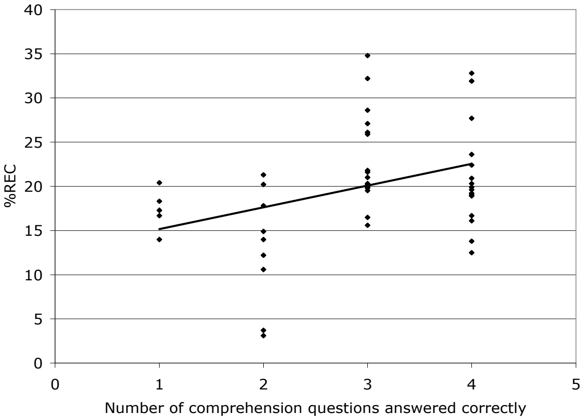

Figure 5. Correlation Between Speaker-Listener Eye Movements Coupling and Listener Comprehension

Speaker-Listener Eye Movement Linkage and Listener Comprehension

The degree to which eye movements were linked in a given speaker-listener dyad were compared with the listener's comprehension of what had been said. For each dyad, we computed the degree of recurrence (REC%) at a lag of 1650ms between speaker and listener. This is the lag that produced the greatest recurrence across our whole data set, and hence serves as a baseline to compare the linkage between individual speaker-listener dyads. The performance of listeners answering four comprehension questions was taken as an objective measure of how well they had comprehended the one minute of speech.

A regression analysis was performed on this data, and found that a linear fit had r^2=0.14. Although it may not account for a large portion of the variance in participants' behaviour, an ANOVA shows that this relationship is significant (F(1,47)=7.39, p<.01).

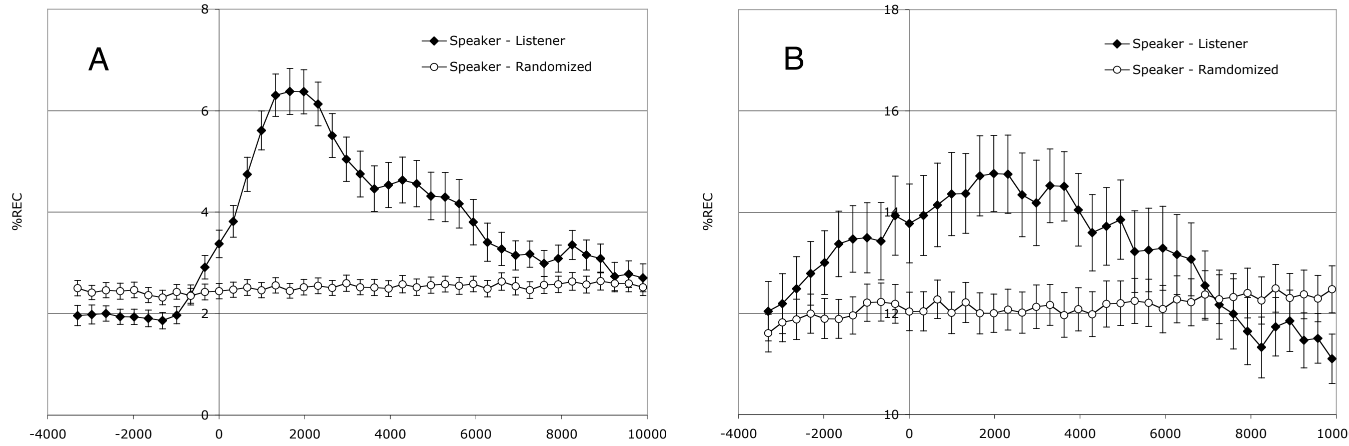

Figure 4. Cross recurrence at different time lags for (a) name fixations, (b) non-name fixations

General Discussion

The current experiment uses a naturalistic paradigm that elicits and presents spontaneous speech. The language-use in this experiment is grounded in the visual items presented on the display, but is not a description of them per se, or an explicit instruction relating to their presence or appearance. Nevertheless, this single paradigm replicates several results obtained in more constrained circumstances concerning the relationship between eye movements, speech production, and speech comprehension.

More importantly, this experiment provides what could be the first demonstration that during the production and comprehension of a spontaneous discourse, the eye movements of a speaker and a listener are coupled. Moreover, this relationship between eye movement patterns is not driven by cases in which the speaker explicitly names people who are depicted. It seems to be that the planning of more diverse types of reference and foregrounding may be influencing the speaker's eye movements, and, a few seconds later via the speech stream, influencing the listener's eye movements. Crucially, the strength of relationship between the speaker's and the listener's eye movements appears to predict the degree to which the listener successfully comprehended the speech.

Instances of new paradigms such as this inevitably raise many questions for future research. Is it the case that a tight coupling between speaker and listener eye movements is an overall indication of listener attentiveness, which also predicts listener comprehension? Or is it that by rapidly bringing their eyes to bear on the same item as the speaker, good listeners receive appropriate visual information that supports the verbal input? Or perhaps it is not so much that moving the eyes closely in step with a speaker brings in visual content, but rather it is an indication (or a cause) that the listener is using spatial information to cognitively structure the information in the same way as the speaker?

The close relationship between speaker and listener eye movements and the success of the discourse clearly aligns with a view of language use as a joint activity (Clark, 1996), in which successful communication is brought about by a successful coordination of information in the common ground. The human eye only receives detailed information from 2º of its visual field: therefore, if the speaker and listener are looking at exactly the same thing, then they are certainly sharing a higher, common ground.

Acknowledgments

The authors are indebted to Natasha Kirkham, Herb Clark and Michael Spivey.

References

Argyle, M., & Cook, M. (1976). Gaze and Mutual Gaze. Cambridge: Cambridge University Press.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52(3), 566-580.

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.

Dale, R. & Spivey, M. J. (submitted). *Data visualization of complex behavioral structure across time.* Manuscript submitted for publication.

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.

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.

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., & Kirkham, N. Z. (in press). Multi-modal events and moving locations: evidence for dynamic spatial indexing in adults and six month olds. Journal of Experimental Psychology: General.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. Cognition, 76, 269-295.

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*, 326-332.

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. Psychological Research/Psychologische Forschung, 65(4), 235-241.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. Journal of Psycholinguistic Research, 29(6), 557-580.

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.

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