Cognitive Science 29 (2005) 39–54

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

Daniel C. Richardson[a], Rick Dale[b]

[a]*Psychology Department, Stanford University*
[b]*Department of Psychology, Cornell University*

Received 1 November 2004; received in revised form 4 April 2005; accepted 13 April 2005

Abstract

We investigated the coupling between a speaker's and a listener's eye movements. Some participants talked extemporaneously about a television show whose cast members they were viewing on a screen in front of them. Later, other participants listened to these monologues while viewing the same screen. Eye movements were recorded for all speakers and listeners. According to cross-recurrence analysis, a listener's eye movements most closely matched a speaker's eye movements at a delay of 2 sec. Indeed, the more closely a listener's eye movements were coupled with a speaker's, the better the listener did on a comprehension test. In a second experiment, low-level visual cues were used to manipulate the listeners' eye movements, and these, in turn, influenced their latencies to comprehension questions. 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.

Keywords: Psychology; Attention; Communication; Discourse; Language understanding; Perception; Situated cognition; Human experimentation; Eye movements

1. Introduction

Imagine standing in front of a painting, discussing it with a friend. As you talk, you scan the image, your eyes moving three or four times a second. Your eyes 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 influenced both by properties of the visual world and processes in a person's mind. Your gaze will also be affected 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

Requests for reprints should be sent to Daniel C. Richardson, Psychology Department, Jordan Hall, Stanford University, Stanford, CA 94305. E-mail: richardson@psych.stanford.edu

relation between your eye movements and those of your friend? How is that relation related to the flow of conversation between you?

In these studies, we used eye movements as a fine-grained index of how speakers and listeners deployed their attention within a visual "common ground." This allowed us to investigate the temporal coupling between conversants' eye movements and to examine whether this coupling is helpful to the success of the discourse.

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). Although eye contact between conversants plays a crucial role in coordinating a conversation (Bavelas, Coates, & Johnson, 2002) and in conveying various attitudes or social roles (Argyle & Cook, 1976), these studies focus on cases in which two partners are looking not at each other, but at a visual scene that is the topic of the conversation. 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.

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; Matlock & Richardson, 2004). Much is known about the eye movements of speakers and listeners in isolation. For example, when speakers are asked to describe a simple scene, they fixate the objects in the order in which they are mentioned and roughly 800 to 1,000 msec before naming them (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). Likewise, when listeners view a scene containing referents for what they are hearing, their eye movements show that they can recognize a word before hearing all of it (Allopenna, Magnuson, & Tanenhaus, 1998), use visual information to disambiguate syntactic structures (Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995), and anticipate agents of actions (Kamide, Altmann, & Haywood, 2003). Participants engaged in a collaborative task reveal a remarkable sensitivity to the referential domains established by the task, the visual context, and the preceding conversation (Brown-Schmidt, Campana, & Tanenhaus, 2004; Clark & Krych, 2004; Hanna & Tanenhaus, 2004; Hanna, Tanenhaus, & Trueswell, 2003). Although fixation times are heavily modulated by context, research suggests that listeners will fixate an object roughly 500 to 1,000 msec after the onset of the spoken name, which includes the 100 to 200 msec needed to plan and execute an eye movement (for a review, see Fischer, 1998).

Eye movements have provided insight into many mental processes (Just & Carpenter, 1976; Liversedge & Findlay, 2000; Richardson & Spivey, 2004). For example, adult and infant participants will make systematic eye movements to particular empty regions of space when retrieving information from memory (Richardson & Kirkham, 2004; Richardson & Spivey, 2000). Moreover, influencing how the eyes move across a scene affects mental processes. Researchers have recorded the eye movements of participants interpreting an ambiguous picture 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 likely 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 brought about by successful dialogue result in similar eye-movement patterns between speaker and listener?

The relation between speaker and listener eye movements was investigated here in one observational study and one experiment. In our paradigm, participants either talked, or listened to talk, about television shows. All participants looked at the same pictures of the cast members. The literature reviewed previously suggests that speakers' eye movements will be systemically related to their production and that listeners' eye movements will be systematically related to their comprehension. Therefore, one might expect some sort of relation between the speaker's and listener's eye movements. But these previous experiments typically concerned single sentences that either described simple scenes (e.g., "The horse is kicking the donkey") or directed the listener toward objects in an array (e.g., "Put the apple on the towel in the box"). In contrast, our participants are involved in extended discussions that not only refer to the pictures before them, but also concern various events, narratives, and opinions that are not depicted. Despite the wider scope of our participants' discourse, and the extended, spontaneous nature of their speech, we hypothesized that the visual common ground will still play a central role in the verbal interaction (Clark, 1996). Therefore, our first prediction was that speaker and listener eye movements around the common ground will be closely coupled in time. Because the success of a linguistic interaction is often dependent on a successful coordination of attention (Clark & Krych, 2004), our second prediction is that the degree of the eye-movement coupling will reflect the degree to which the listener understood the speaker. Given the evidence that patterns of eye movements can determine how a stimulus is interpreted (Pomplun et al., 1996) or a problem is solved (Grant & Spivey, 2003), our final prediction was that the relation between eye movements and comprehension would be causal: that if we manipulated a listener's eye movements we would influence his or her understanding.

2. Study 1

This study examines the eye movements of speakers and listeners looking at the same visual scene. One set of participants (called speakers) talked spontaneously about a television show whose characters were displayed in front of them. Audio recordings of their speech were then played to a second set of participants (called listeners) who were looking at the same display. Afterward, we measured the listeners' comprehension by a series of questions. Speakers' and listeners' eye movements were tracked throughout.

2.1. Methods

2.1.1. Participants

Forty Stanford University undergraduates took part in the study in exchange for course credit. The first 4 participants were designated as speakers, and the other 36 were listeners.

2.1.2. Apparatus

An ASL 504 remote eye-tracking camera (Applied Science Laboratories, Bedford, Massachusetts) was positioned at the base of a 17-in. LCD stimulus display. Participants sat unrestrained approximately 30 in. from the screen. The camera and eye-tracking PC detected pupil and corneal reflections from the right eye. This information was passed every 33 msec to a PowerMac G4, which controlled the stimulus presentation and recorded which, if any, of the stimuli were currently under the participant's point of gaze. The participants went through a 9-point calibration routine, which typically took between 2 and 5 min.

2.1.3. Design: Speakers

The 4 speakers were asked to talk about one of two popular television shows, Friends and The Simpsons, while looking at an array of pictures of the six principal cast members. The array subtended approximately 26 × 19° of visual angle, and each cast member's picture subtended approximately 8°.

For Friends, the speakers were asked to "talk about the relationships between the characters, your opinion of them, or your favorite episode." For The Simpsons, the speakers were shown a 5-min scene and were asked to "Describe what went on in the scene and what you thought about it." We tracked the speakers' eye movements and recorded their voices as they spoke. We extracted an unedited 55- to 60-sec segment of their speech for use in the second phase of the study. Segments were chosen so that they contained a sufficient amount of factual content for a comprehension test.

2.1.4. Design: Listeners

The 36 listeners listened to a segment of speech while looking at the same picture of the six cast members that had been in front of the speaker. Because there could not be systematic looks to the cast members if they did not recognize anyone, we asked if they were familiar with either show. On this basis, they were presented with one or both of the Friends and The Simpsons stimuli and were randomly assigned one of the 2 speakers for that show.

After listening to the speech, the listeners answered four spoken questions using the buttons of a mouse. The questions, recorded by DCR, were of the form, "Did the speaker say … ?" The questions could not be answered on the basis of knowledge about the television shows but were specific to the information mentioned by that particular speaker. For example, the correct answer to "Did the speaker say that Bart electrocuted Homer?" was "No." Even though the event may have happened, that particular speaker did not convey that information. The correct answer to half the questions was "yes" and half "no."

2.1.5. Data coding

Roughly half of the listeners were familiar with both television shows and half knew the characters from only one. Nine cases were dropped due to problems with calibration, leaving 49 usable speaker–listener pairings. The eye-movement data were cleaned for blinks and saccades across a picture—gaze durations of less than 100 msec in any single region were discarded. The speakers' recordings were manually transcribed with onset times using waveform-editing software. On average, speakers spoke 160 words, only 12 of which were the names of the characters depicted. These monologues were not edited for content and include

all the deviations, hesitations, and repetitions that are typical of just a minute of normal, spontaneous speech.

2.2. Results

An excerpt from a monologue, and eye movements made while producing it and listening to it can be seen at http://www.cognitivesciencesociety.org/supplements/. Movie 1 is a composite of three eye-tracking recordings. The speaker's eye movements are shown as a crosshair, and two listeners' eye movements are shown as dots. The listener shown as a light gray dot subsequently answered all four comprehension questions correctly; the other listener only answered one correctly.

In addition to the video record, this study provided precise timing information for speech and eye movements. This information can be depicted graphically in what we call a "scarf plot." The left hand side of Fig. 1 shows a 7-sec segment of a scarf plot for one speaker–listener pair. 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.

For each occasion that a speaker named character X, his or her eye-movement data were consulted to find the point at which X was last previously fixated. On average, a character was

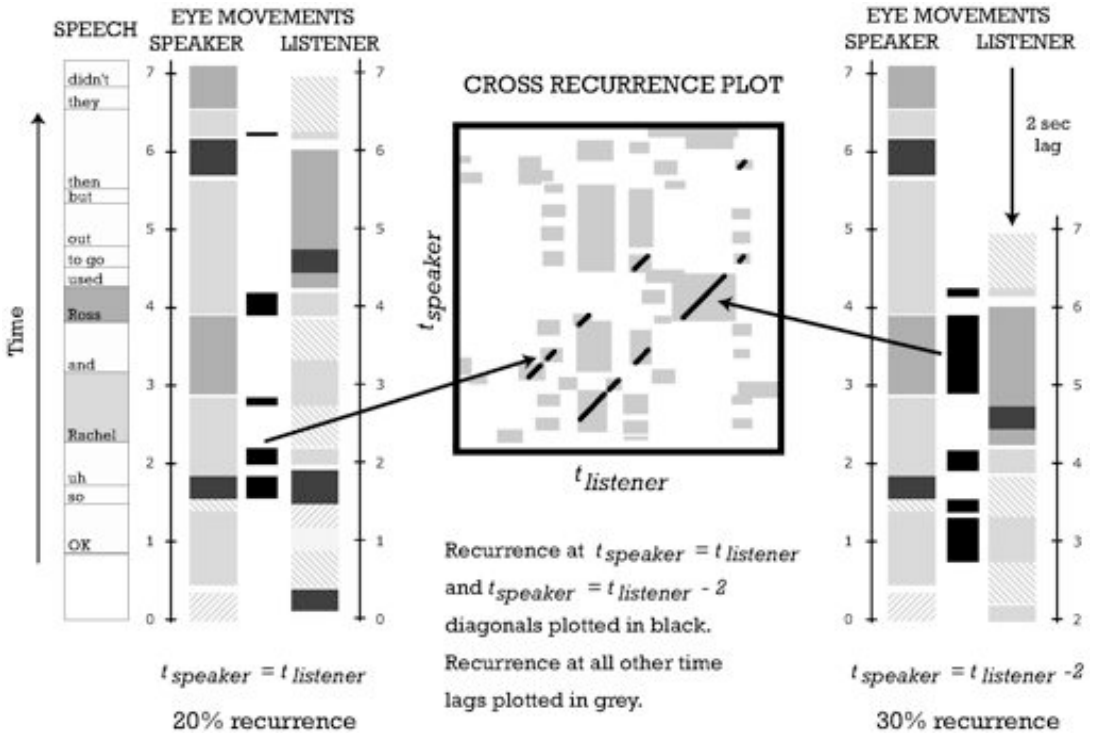

Fig. 1. Scarf plot and explanation of cross-recurrence analysis.

fixated 860 msec prior to being named. This lag is well within the range reported in experiments in which participants are instructed to describe a simple picture (see Griffin & Bock, 2000). Here we have found a lag of the same length in natural speech, when speakers are describing not what is in front of them, but things that are not depicted—stories, opinions, relationships—that relate to the characters shown.

2.2.1. Cross-recurrence analysis

What is the relation between the eye movements of the speaker and of the listener? We addressed this question using recurrence analysis (Eckmann, Kamphorst, & Ruelle, 1987; Zbilut & Webber, 1992). Cross-recurrence plots permit visualization and quantification of recurrent patterns of states between two time series (Zbilut, Giuliani, & Webber, 1998). Such analyses are useful because they can reveal the temporal dynamics of a data set without making assumptions about its statistical nature. For example, in a study by Shockley, Santana, and Fowler (2003) recurrence analysis was used to demonstrate the coordination of body sway by two people in conversation. And in analysis by Dale and Spivey (2005) it was used to uncover the temporal properties of children's developing language structure.

Fig. 1 offers a graphical intuition into how we have used cross-recurrence plots to analyze the relation between speaker and listener eye movements. Each diagonal on a cross-recurrence plot corresponds to a particular alignment of the speaker's and listener's eye-movement data with a particular lag time between them. A point is plotted along that diagonal whenever the speaker and listener's eye movements are recurrent—whenever their eyes are fixating the same object. Note that if the speaker and listener are not looking at any object at the same time (they were looking at blank spaces or off the screen or were blinking) this is not counted as recurrence. On the left side of Fig. 1, the speaker and listener eye-movement scarf plots for a 7-sec segment are aligned with no time lag. In between them, the periods of recurrence are shown in black. In total, these areas of recurrence account for 20% of the time series. These points are plotted along the tspeaker = tlistener line on the cross-recurrence plot. On the right side of Fig. 1, the speaker and listener eye movements are aligned with the listener lagging behind the speaker by 2 sec. The recurrence between these series is plotted along the tspeaker = tlistener × 2 line. At this time lag there is 30% recurrence. A full recurrence plot for a speaker–listener dyad is formed by calculating the recurrence between all such alignments at all possible lag times. These points are shown in gray in Fig. 1. In the full analyses that follow, the whole minute of data was processed and the lag times were incremented at 33-msec intervals.

Cross-recurrence plots allow temporal structure to be visualized and differences between speaker and listener pairs to be quantified. Fig. 2 highlights such differences by showing example cross-recurrence plots between a speaker and a good listener who answered all comprehension questions correctly, a bad listener who answered few correctly, and a listener with his or her eye-movement data shuffled in a random order. At first glance, the good listener's plot reveals more points of recurrence and also appears more dense and clustered. Both of the real listeners have a higher density in the region on and below the tspeaker = tlistener diagonal. This indicates that the speaker' and listeners' eye positions overlapped more when the listeners' eye-position time series was shifted behind the speakers' in time.

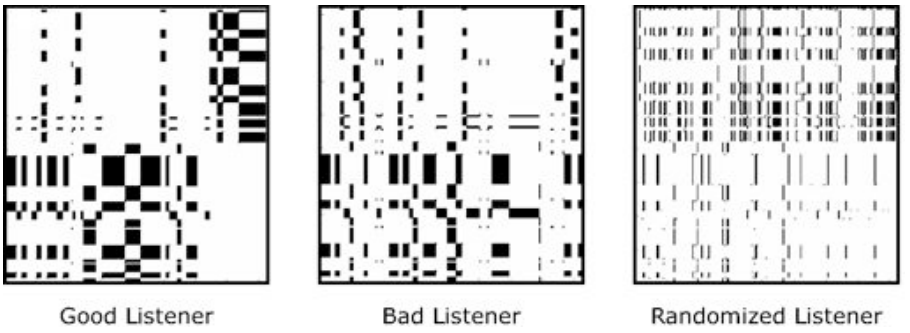

Fig. 2. Example cross-recurrence plots of the eye movements of a speaker and (a) a good listener, (b) a bad listener (c) a shuffled listener.

2.2.2. Relation between speaker and listener eye movements

What lag time produced the most recurrence between the speaker and listener eye movements? We answered this question by generating cross-recurrence plots for each of the 49 speaker–listener pairs. Fig. 3 shows the average percentage of recurrence in these pairings for each time lag. This distribution can be compared to two different baseline conditions. To produce the "speaker–randomized listener" distribution we shuffled the temporal order of each listener's eye-movement sequence and calculated its recurrence with the speaker. 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. To produce the "speaker–mismatched listener" distribution we calculated the recurrence between a speaker and a listener who was looking at the same picture but listening to a different speaker. This provides a baseline of recurrence between people who share the same visual information but different verbal information.

As shown in Fig. 3, the real listeners are not looking around these displays randomly. Compared to the randomized listeners, their eye movements are linked to the speakers' within a particular temporal window. Between 0 and 6,000 msec after the speaker has looked at something, the listeners are looking at the same thing at above-chance levels. The maximum recurrence between the speakers and listeners, the lag time at which their eye movements overlap the most, is around 2,000 msec. The differences between the speaker–listeners and speaker–randomized listeners were supported by a 2 (listeners–randomized listeners) × 41 (lag times) mixed-effects analysis of variance (ANOVA) (lag as a repeated measures factor) that revealed a significant main effect of listener type, $F(1, 48) = 40.6$, $p < .0001$, and a main effect of lag, $F(40, 1920) = 7.4$, $p < .0001$. And there was a significant interaction between the factors, $F(40, 1920) = 7.7$, $p < .0001$. To assess the variance in our estimation of the point of maximum recurrence, we carried out a resampling analysis (Lunneborg, 2000). We generated 1,000 samples that fit the mean and standard distribution in each bin of our speaker–listener recurrence data. The average point of maximum recurrence was reached at 1,997 msec with a 95% confidence interval of ± 59 msec.

The coupling between speaker and listener eye movements is clearly not produced by chance, nor is it produced by their common visual stimulus, as shown by our second baseline comparison.

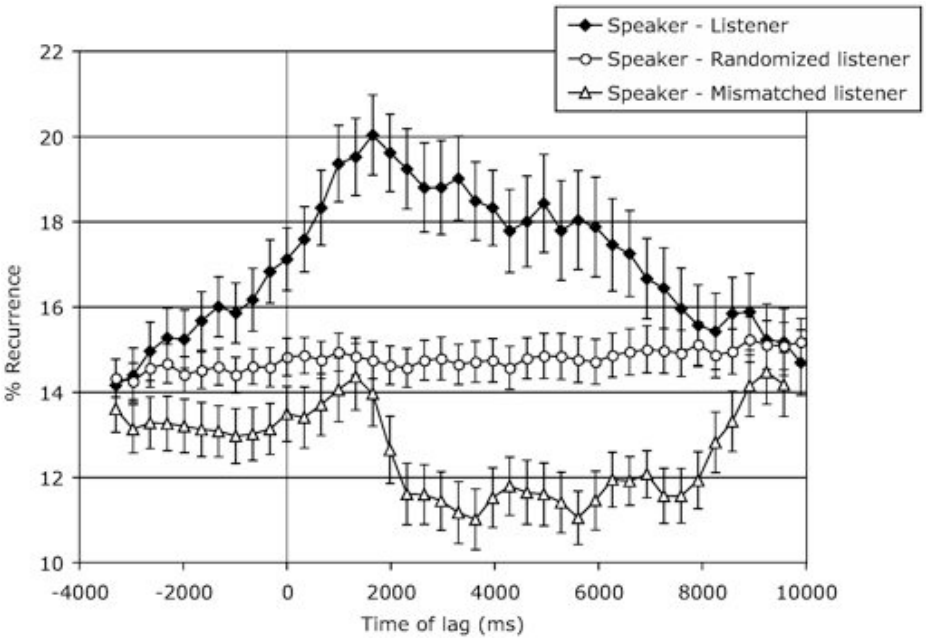

Fig. 3. Average cross-recurrence at different time lags for 49 speaker–listener pairs.

Compared to the real listeners, the mismatched listeners have much lower recurrence with the speaker, even though they were looking at the same image. This is supported by a 2 (listeners–mismatched listeners) × 41 (lag times) mixed-effects ANOVA that showed a significant main effect of listener type, $F(1, 48) = 18.2$, $p < .0001$, a main effect of lag, $F(40, 1920) = 4.4$, $p < .0001$, and a significant interaction between the factors, $F(40, 1920) = 9.6$, $p < .0001$.

The distribution of the speaker–listener recurrence is what one might expect from the combination of the speech production and speech comprehension eye-movement literature. Typically, speakers fixate an item 800 to 1,000 msec before naming it, and listeners fixate an object 500 to 1,000 msec after the name onset. The sum of these values corresponds to a range of high recurrence in our data set between 1,000 and 3,000 msec. The lag that produced the maximum recurrence is around 2,000 msec. This value is slightly higher than one might expect from the speech production and speech comprehension eye-movement literature, which possibly reflects the relative complexity of the spontaneous speech used in our study.

The speech production and comprehension literatures deal with cases where an object or person is explicitly named. Perhaps the coupling between speaker and listener eye movements observed here is due merely 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 subsets of the data. The name subset included only speaker fixations to person X that were immediately prior to the speech onset of name X. This constituted about 10% of the 120 fixations made by the average speaker.

Fig. 4A plots the recurrence at different time lags for the name subset of our data. Because a subset contains fewer speaker fixations than the full data set, the overall recurrence for listen-

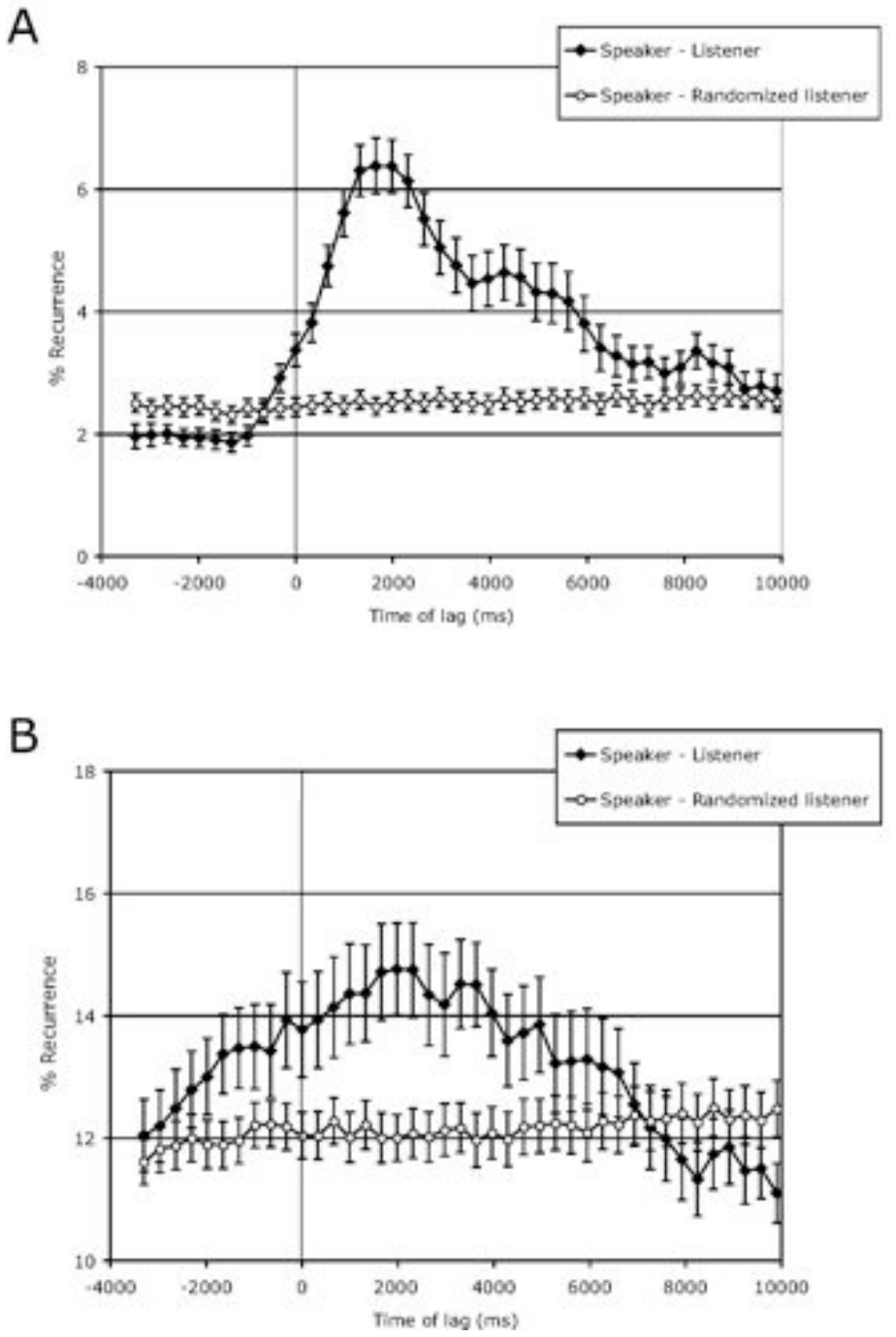

Fig. 4. Cross-recurrence at different time lags for (a) name fixations and (b) non-name fixations.

ers and randomized listeners is lower. Here again, there is a highly pronounced difference between the listeners and the randomized listeners. The 2 (listeners–randomized listener) × 41 (lag times) mixed-effects ANOVA revealed a significant main effect of listener type, $F(1, 48) = 139.2$, $p < .0001$, a main effect of lag, $F(40, 1920) = 24.8$, $p < .0001$, and a significant interaction between the factors, $F(40, 1920) = 24.2$, $p < .0001$. Resampling analyses suggest that the peak of this distribution is at 1,708 msec (± 18) on average. This figure is much closer to that suggested by a sum of the values from the speech production and speech comprehension literatures. Has this study simply replicated these name-use results using spontaneous speech?

Fig. 4B plots the other 90% of the data, the "non-name data set" composed of speaker fixations to person X that were not immediately followed by X being named out loud. The peak of maximum recurrence here is reached at 2,078 msec (± 62 msec) on average in our resampling analyses. In this subset too, there was a large difference between listeners and randomized listeners. The ANOVA showed a main effect of listener type, $F(1, 48) = 36.4$, $p < .0001$, a main effect of lag, $F(40, 1920) = 3.1$, $p < .0001$, and a significant interaction between the factors, $F(40, 1920) = 3.7$, $p < .0001$. Therefore, it is not just when the speaker names a character that speaker and listener eye movements are linked.

Comparisons between the name and non-name data set revealed interesting differences. An ANOVA (Subset × Listener type × Time lag) produced a significant three-way interaction, $F(40, 1920) = 5.82$, $p < .001$. Therefore, if the speaker used the name of one of the characters, it significantly affected the listener eye-movement coupling (compared to randomized looking). In addition, the location of the peaks of maximum recurrence differed significantly between the two subsets, $t(1998) = 9.13$, $p < .0001$), with the name subset producing a maximum 370 msec sooner. More important, although the relation was closer when the speaker named one of the characters, our statistical analyses demonstrate that in both subsets of our data, speaker and listener eye movements were coupled.

2.2.3. Speaker–listener eye-movement linkage and listener comprehension

The accuracy of a listener's comprehension was compared with how closely that listener was following the speaker's eye movements. Listeners were grouped by whether their accuracy was high (3 or 4 correct answers, N = 35) or low (1 or 2 correct, N = 14). Fig. 5 shows the recurrence at different time lags for these groups of listeners. An ANOVA revealed a significant main effect of accuracy group, $F(1, 47) = 14.3$, $p < .0001$, a main effect of lag, $F(39, 1833) = 8.06$, $p < .0001$, and a significant interaction between the factors, $F(39, 1833) = 3.7$, $p < .0001$.

The relation between eye-movement linkage and comprehension was confirmed by two regression analyses. First, for each dyad we computed the degree of recurrence at a lag of 2,000 msec between speaker and listener. This lag produced the greatest recurrence across our whole data set and, hence, serves as a baseline to compare the linkage between individual speaker–listener dyads. There was a correlation of .33 between these values and listener accuracy, $F(1, 47) = 5.91$, $p < .05$. We also calculated the lag time that produced the maximum recurrence for each dyad. There was a correlation of –.34 between the maximum lag and listener accuracy, $F(1, 47) = 6.1$, $p < .05$. In addition to the dichotomous difference between high- and low-accuracy listeners, these regressions indicate that there appears to be a more graded correlation between listener accuracy and various measures of the strength of the speaker–listener eye-movement coupling.

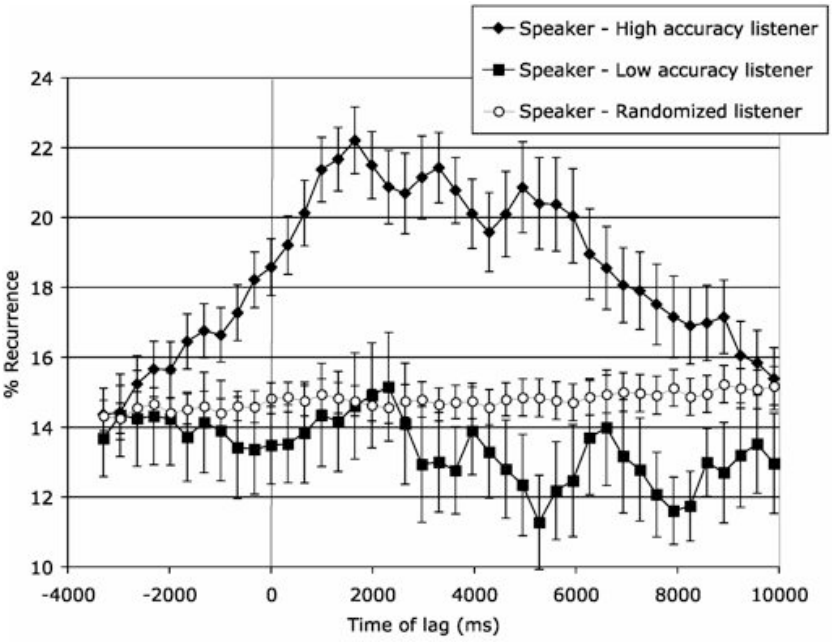

Fig. 5. Cross-recurrence for listeners with high and low comprehension.

2.3. Discussion

The eye movements of speakers and listeners are linked. Although this basic fact is intuitively obvious, it has not been empirically demonstrated before. More interestingly, our results revealed the precise time frame of this coupling. Between 0 and 6 sec after a speaker looks at a picture, the listener is more likely than chance to be looking at it. He or she is most likely to be looking at it about 2 sec after the speaker. This result is not solely due to cases where the speakers explicitly name a person who is depicted, but is found throughout the discourse. Just as participants in Altmann and Kamide's (2004) experiments use their knowledge of verbs to make eye movements toward likely agents and patients, it seems plausible that listeners in our study are establishing reference, using various sources of information in the speech stream.

In an experiment by Brown-Schmidt et al. (2004) a speaker instructed a listener to move various blocks on a grid. Sometimes, the speaker referred to "the red one," even though there were several red blocks in sight. The listener was able to find the correct block, however, because what the speaker had said previously had implicitly identified a smaller set of objects that included only one of the red blocks. This was called "circumscribing the referential domain." Perhaps here, too, speakers and listeners keep track of a smaller set of pictures that are currently relevant to the monologue. This might explain why they are more likely than chance to be looking at the same picture at the same moment and why listeners are more likely to be looking at that picture for a further 6 sec.

How closely a listener is following a speaker's gaze predicts how well he or she will answer comprehension questions. This relation may simply reflect overall listener attentiveness and

interest in or prior knowledge of the sitcoms. In addition, four questions are a relatively coarse measure of comprehension. In Study 2 we sought to make a causal connection between eye-movement linkage and understanding by manipulating eye movements and taking finer measures of comprehension accuracy.

3. Study 2

In this experiment, participants listened to the monologues of Study 1 while pictures of the characters flashed either at the time that the speaker had looked at each picture, or according to a shuffled version of the speakers' eye movements. These bright onsets were assumed to attract the listeners' attention (e.g., Weichselgartner & Sperling, 1987), hence make their eye-movements probabilistically more or less like the speakers'. We hypothesized that manipulating the linkage between speakers and listeners' eye movements would influence the listeners' comprehension.

3.1. Methods

The apparatus, stimuli, and design were identical to the listeners' phase of Study 1, with the following exceptions.

3.1.1. Participants

Thirty-six Stanford University undergraduate students participated in exchange for course credit.

3.1.2. Design

Participants were randomly assigned to one of the four speakers and were counterbalanced between conditions. During presentation of the speaker's monologue, pictures of the six cast members sequentially turned from dimmed to full color. In the synchronized condition, a picture was bright whenever the speaker had been looking at it. In the shuffled condition, its brightness was determined by randomizing the order of the speaker's fixations. Movie 2 (accessible online at http://www.cognitivesciencesociety.org/supplements/) shows examples of the stimuli from each experimental condition. Following the presentation, participants answered eight comprehension questions.

3.2. Results

Four participants were dropped because of failures to calibrate. For the remaining 32 listeners, changes in picture luminance influenced the eye movements as predicted. An ANOVA performed on the cross-recurrence analysis revealed a significant main effect of listener condition (synchronized–shuffled), $F(1, 29) = 11.3$, $p < .005$, a main effect of lag, $F(40, 1160) = 11.2$, $p < .0001$, and a significant interaction between the factors, $F(40, 1160) = 6.5$, $p < .0001$.

Listeners took almost 40% longer to answer questions in the shuffled condition compared to the synchronized condition (synchronized, M = 1,364 msec; shuffled, M = 1,889 msec; F(1,

30) = 5.00, $p < .05$. There was, however, no significant difference in the number of questions answered correctly in the two conditions, F(1, 30) = .04.

3.3. Discussion

The correlational results of Study 1 were confirmed here with an experimental manipulation that affected both listeners' looking behavior and their performance while answering comprehension questions. Although this subtle manipulation did not alter the accuracy of listeners' comprehension, the more sensitive measure of response latencies did reveal a difference between conditions. Following examples in visual perception and problem solving (Grant & Spivey, 2003; Pomplun et al., 1996), this experiment presents a case in language comprehension of a low-level perceptual cue causing one person's eye movements to look like another's and, as a consequence, affecting their cognitive state.

One might argue that all this experiment shows is that listeners in the shuffled condition found the flashing distracting, and hence their comprehension performance suffered. Its distracting quality may not be related to the visual nature of the cue per se and its effect on eye movements, but may simply be because the shuffled flashing does not occur synchronously with the discourse. In the shuffled case the stimuli do not "go together," are harder to process, and hence distracting.

This argument falls down when we consider the other condition, however. When the stimuli flashed according to the speaker's eye movements, it was not the case at all that the flashes were synchronized with the speech stream in any straightforward manner. As can be seen in the scarf plots (Fig. 1) and example stimuli (Movie 2), the lag between the speaker's gaze and name onset (860 msec on average) meant that typically the speaker was not looking at the characters when they spoke their names. Both conditions, therefore, involved flashing stimuli that were asynchronous with the speech stream. Only the shuffled condition proved to be a distraction, however. It is precisely our point that what makes the flashing a distraction or not is its relation to the speaker's eye movements.

4. General Discussion

From the moment a speaker looks at a picture and for the following 6 sec, the listener is more likely than chance to be looking at that same picture. The breadth of this time frame suggests that speakers and listeners may keep track of a subset of the depicted people who are relevant moment by moment. The listener is most likely to be looking at the same thing as the speaker after around 2,000 msec. The timing of this peak corresponds to the timing of eye movements participants make while producing or hearing the names of objects. In our studies, however, speaker and listener eye movements are coupled throughout the discourse, not just while the speaker is naming people who are depicted. The pervasiveness of this coupling suggests that planning diverse types of speech will influence the speaker's eye movements, and a few seconds later, hearing them will influence the listener's eye movements.

Crucially, how closely the speaker's and the listener's eye movements are linked appears to predict how successfully the listener comprehended the speech. However, eye-movement cou-

plings were not merely a cognitive epiphenomenon, simply providing a window onto cognitive activity during communication. This relation is not just correlational but causal: When a low-level perceptual cue made the eye movements of a listener more or less like the speaker's, the listener's performance on comprehension questions was affected.

Why is discourse comprehension related to the dynamic coupling between conversants' eye movements? By rapidly bringing their eyes to bear on the same item as the speaker, do good listeners receive appropriate and timely visual information that supports the verbal input? This might seem unlikely, because no new visual information is presented during the course of the speech. Perhaps it is not that moving the eyes closely in step with a speaker supplies visual data, but that it allows the listener to use spatial structure to organize information in the same way as the speaker.

Previously, it has been observed that conversants coordinate each other's attention by large actions such as pointing, placing, and gesturing (Clark, 2003; Clark & Krych, 2004). Our experiments find support for this notion of conversation as a joint activity, despite the fact that our conversants were separated in time and space. Even though conversants could not interact with each other, their visual attention was coupled at the millisecond resolution of eye movements. We argue that because this coupling determines conversants' comprehension performance, looking around the common ground in step with each other is part of the process of mutual understanding.

Acknowledgments

The authors are indebted to Herb Clark, Michael Spivey, Natasha Kirkham, and Teenie Matlock. We would also like to thank all the participants in this study, and Lisa Smythe and Natalie Ramirez for assistance running experiments.

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,* 419–439.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. M. Henderson & F. Ferreira (Eds.), *Interfacing language, vision, and action* (pp. XX–XX). San Diego, CA: Academic.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze.* Cambridge, England: Cambridge University Press.

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* (pp. XX–XX). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52,* 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* (pp. XX–XX). Cambridge, MA: MIT Press.

Clark, H. H. (1996). *Using language.* Cambridge, England: 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 Associates, Inc.

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: American Psychological Association.

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

Dale, R., & Spivey, M. J. (2005, July). *Categorical recurrence analysis of child language.* Paper presented at the 27th annual meeting of the Cognitive Science Society. Manuscript submitted for publication.

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

Fischer, B. (1998). Attention in saccades. In R. D. Wright (Ed.), *Visual attention* (pp. 289–305). New York: Oxford University Press.

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

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

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8,* 441–480.

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,* 133–156.

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

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Duxbury.

Matlock, T., & Richardson, D. C. (2004). *The integration of figurative language and concrete depictions: An eye movement study of fictive motion.* Manuscript submitted for publication.

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

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

Richardson, D. C., & Kirkham, N. Z. (2004). Multimodal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General, 133*(1), 46–62.

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

Richardson, D. C., & Spivey, M. J. (2004). Eye tracking: Research areas and applications. In G. Wnek & G. Bowlin (Eds.), *Encyclopedia of biomaterials and biomedical engineering* (pp. XX–XX). New York: Marcel Dekker.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47,* 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,* 326–332.

Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995, XXX XX). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634.

Weichselgartner, E., & Sperling, G. (1987, XXXX XX). Dynamics of automatic and controlled visual attention. *Science, 238,* 778–780.

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.