Turn-boundary projection: Looking ahead

Marisa Tice (middyp@stanford.edu)
Margaret Jacks Hall, Stanford University
Stanford, CA 94305-2150 USA

Tania Henetz (thenetz@stanford.edu)
Jordan Hall, Stanford University
Stanford, CA 94305-2150 USA

Abstract

Coordinating with others is hard; and yet we accomplish this every day when we take turns in a conversation. How do we do this? The present study introduces a new method of measuring turn-boundary projection that enables researchers to achieve more valid, flexible, and temporally informative data on online turn projection: tracking an observer's gaze from the current speaker to the next speaker. In this preliminary investigation, participants consistently looked at the current speaker during their turn. Additionally, they looked to the next speaker before her turn began, and sometimes even before the current speaker finished speaking. This suggests that observer gaze is closely aligned with perceptual processes of turn-boundary projection, and thus may equip the field with the tools to explore how we manage to take turns.

Keywords: Turn-taking; Social cognition; Eye tracking; Coordination; Timing; Conversation; Interaction

Introduction

Interacting with others requires us to make a number of complex linguistic, social, and cognitive decisions. Despite this, most conversations proceed smoothly, allowing us to take for granted the intricate processes taking place in getting the timing of our actions right on cue. Turn-taking during conversation is one phenomenon that exemplifies these issues. Intuitively we seem to wait for the current speaker to stop talking before we start conjuring up a response, with each turn preceded and followed by orderly pauses or 'gaps' in speaking. But this is not the case: not only do we not 'wait', but there are often no gaps between speakers at all! Speakers are extremely adept at taking turns efficiently, averaging 0.2-0.4 second gaps in face-to-face conversation (Brady, 1968; Stivers et al., 2009) and 0.7 second gaps over the phone (Jaffe & Feldstein, 1970), often with less than 5% overlap (Levinson, 1983). This general pattern has been observed across many cultures, leading researchers to conclude that interlocutors adhere to standards of no-gap-no-overlap in transferring turns from one speaker to the next (de Ruiter et al., 2006; Sacks et al., 1974; Stivers et al., 2009). To accomplish this no-gap-no-overlap timing, listeners must be able to actively **project** the end of the current speaker's turn (hereafter, turn-end boundary), while simultaneously starting to plan their response.

The prevailing method of investigating how projection takes place is to use corpora to identify linguistic cues that coincide with turn-end boundaries (e.g. prosodic, syntactic, and pragmatic boundaries; (Ford & Thompson, 1996; Caspers, 2003)). But these cues often co-occur, making it difficult to interpret relative cue importance. Additionally, some of these cues might come too late for listeners to make use of them, for instance, lengthening of the final word in an utterance happens nearly at the end of the turn (de Ruiter et al., 2006, but see Gravano & Hirschberg, 2011). Addressing turn projection experimentally, de Ruiter and colleagues (2006) asked Dutch speakers to listen to fragments of spontaneous speech and press a button at the moment they anticipated the speaker would be finished speaking. The stimuli were manipulated phonetically, controlling for potential projection cues such as intonation, lexicosyntactic information, and rhythm. Their results suggest that speakers rely primarily on lexical information (which also provides syntactic cues) to identify upcoming turn-end boundaries.

The experimental approach introduced by de Ruiter et al. (2006) is an significant step forward in research on boundary projection, but there is still much to be addressed. Specifically, we do not know how to account for boundary projection as the turn is unfolding. Listeners have access to only that information which has already been spoken, and so their use of cues may differ over the course of a turn. For example, it could be the case that listeners track intonation as a primary cue to the beginning of a turn's denouement, and then increase their reliance on lexical information to precisely identify the end of the upcoming syntactic clause. The information that listeners use to track upcoming turn boundaries should reflect their integrated knowledge of all the cues available to them as the turn is unfolding.

The button-press methodology gives us a single point in the turn at which to test a manipulation. It is incapable of tracking listeners' ongoing certainty level about upcoming turn-end boundaries; especially in cases where there is a possible, or even probable, but not realized turn-end (e.g. "Did I ever tell you about my Aunt Millie? She was a wild one."). In listening to this signal, participants in a button-press experiment are likely to enter their response after "Millie" or after "one." In the case that they were not fooled by the first potential turn-end place, the single button press could not tell us about their ongoing projection: how close they came to thinking of it as a turn-end boundary, what cues were important at the time, et cetera. The button-press also adds an "input" requirement to the task, which might be sensitive to the task instructions. An ideal measure of anticipation would not require explicit instructions, easing the cognitive load on participants that might arise from the specific task.

We propose a new method of investigating turn-projection behavior: tracking observer gaze. In the utterance about Aunt Millie, gaze might reveal a robust effect of the initial probable turn-end point: a gradient increase and then decrease in transition-related looks as the utterance continues. Button-pressing, in contrast, indicates the point in time when the observer felt they had sufficient evidence to respond to an utter-

ance[1], which is somewhat analogous to "jumping in" to actually take a turn. Producing a response is an essential behavior in conversation, but is not the same phenomenon as active turn-end boundary projection, which we may do at all times without ever intending to jump in. Tracking observer gaze allows us to measure how listeners track upcoming boundaries without adding in the complexities of what is required to *respond*.

Observer gaze

During face-to-face conversation, listeners tend to look at the current speaker. This behavior has been documented through naturalistic observation (Kendon, 1967) and replicated in the laboratory (Bavelas et al., 2002) and in studies of human-computer interaction (Jokinen et al., 2009). In each of these studies, eye gaze has been shown to be an effective turn-taking cue. Eye gaze has also been tied to predictive linguistic processes in other contexts (Brown-Schmidt & Tanenhaus, 2006; Griffin & Bock, 2000; Richardson & Dale, 2005). It is possible, then, that when an ongoing conversation nears a point of turn transition, third-party observers will look to the next speaker *anticipatorily* as the current speaker's turn is coming to a close—that is, before the current speaker has stopped speaking. Note that we don't mean to suggest that observer gaze plays the same role in face-to-face conversation as it would in our task, only that there is precedence for this looking tendency.

A third-party observer's eye movements over the transition period from current to next speaker could provide a continuous and naturalistic measure of turn-end boundary projection. This methodology retains the ability to control phonetic and other linguistic factors in the presentation of video stimuli, while permitting the examination of non-linguistic factors in the accompanying visual scene. Furthermore, with minimal changes, it could lend itself well to developmental work since eye-tracking is an effective online measure for young children (Fernald et al., 2010; Gredebäck et al., 2009; Kidd et al., 2011).

This study is an initial investigation of observer gaze as a measure of turn-boundary projection. If observer gaze is a reliable measure of anticipatory turn-taking behavior, we may harness it to investigate the processes and cues used for online turn projection. In this study, we ask the following questions: (1) Do third-party observers reliably track current speakers with their gaze? and (2) Do third-party observers anticipate transitions to the next speaker?

Methods

Participants

The seventeen participants in this study were all current members of the Stanford Linguistics or Psychology Departments (*females* = 9). The participants were volunteers who were not paid for their participation and were unaware of the purpose of the study.

[1]Though in this case, the response is simple.

Materials

To assess observer gaze as a measure of turn-boundary projection, we recorded the eye movements of participants as they watched two short film clips of dialogue. To optimize the ease of coding participants' eye gaze, the video materials also needed to display each speaker in relatively isolated and static positions on screen. In this study we rely on the film device known as the "split-screen" telephone conversation (see Figure 1). During a typical split-screen conversation, the screen is partitioned to simultaneously show two or more speakers as they converse over the phone. This medium satisfies the constraint of conversational interactivity required to expect turn-transitional looks, but also keeps the speakers in distinct enough regions of the screen to make gaze coding feasible. We chose two "split-screen" telephone conversations from the relatively recent film *Mean Girls* (Paramount Pictures, 2004).

Figure 1: Frame of a split-screen scene in the film *Mean Girls*. Speaker 1 is on the left, Speaker 2 on the right.

Procedure

Participants were asked to watch two short scenes from the film and then answer questions about each scene immediately after it finished playing. For the duration of the experiment, participants were seated at a small desk in front of a large display screen. A small video camera was tucked beneath the screen, pointing upwards, toward the participant's face and was switched into recording mode for the duration of the experiment. Audio was played aloud over the speakers of a laptop so that it could be captured by the video camera for later coding. No participant reported difficulty hearing the stimuli.

Participants were first asked about their familiarity with the movie. More than half of the participants had seen the film at least once before (N=10). Regardless of familiarity with the film, each participant was briefly familiarized with the characters featured in the clips. At the start of both trials, participants were reminded that they would be asked about the clip after it was over. Then the experimenter asked the participant to focus on a yellow star centered on the screen until the clip began. After the clip was over, the experimenter and participant went through three comprehension questions verbally. The entire experiment took less than five minutes for all participants. During debriefing, participants were asked if they guessed the purpose of the study. One participant reported

awareness that we were measuring his looks to each speaker, so these data were excluded from analyses.

Question	Answer
1. Hello?	I know your secret.
2. Secret? What are you saying about?	Gretchen told me that you like A.S.
3. Is that bad?	But if you like him, whatever.
4. Really? You would do that? I mean nothing embarrassing though, right?	Oh no, trust me, I know exactly how to play it
5. Aren't you so mad at Gretchen for telling me?	No.

Table 1: Question-answer pairs included in the analyses drawn from the one-minute dialogue. See Figure 2 for individual gaze trajectories.

Eye-gaze coding

Here we report eye gaze data from the first video clip (comprehension score average: 95.8%). We omitted data from the second video before running any analyses. It included a shifting four-way screen split that made direction of gaze impossible to code reliably[2]. There are 15 total transitions in the one-minute video clip. Before analyzing the data, we selected all of the five question-answer pairs as the target of our analyses. Question-answer sequences present a reasonable example case for testing this method since they are reliable as adjacency pairs (i.e. they usually elicit a response), but still provide a diverse sample of Speaker1-Speaker2 sequences.

The video recording of each participant's gaze during the clip was analyzed by at least two coders: one of the authors and one trained coder naive to the purpose of the study. Direction of gaze was coded for each 50 ms interval of the one-minute film clip as 'right', 'left', 'center' and 'not codeable'. These were recoded to numerical values for averaging across coders, replacing uncodeable values with the average of the values directly preceding and directly following. Intercoder agreement was high (96%)[3].

Results

Do observers look at the current speaker? Figure 2a displays the average gaze trajectory for all 16 participants across the entire one-minute film clip. It is clear that observers reliably look at the current speaker. This was confirmed by unpaired Wilcoxon signed rank tests on the proportion of looks to the *current speaker* during each turn in the minute-long dialogue. While Speaker 1 was talking, observers were looking at her 79.6% of the time, and while Speaker 2 was talking, observers gazed at Speaker 1 only 25.2% of the time. When neither speaker was talking, gaze was divided between speakers, with 58.1% looking to Speaker 1. Each of these differences is significant overall ($p<0.001$) and for 15 of the 16 participants ($p<0.05$). This is strong evidence that observers are looking at the current speakers during their turns, replicating previous work by Bavelas, Coates, & Johnson (2002).

This pattern also means that observers must be reliably shifting their gaze between speakers when the floor is transferred. So do observers reliably *anticipate* the next speaker's turn with their gaze, as they do when participating in everyday conversation? Figures 2b-f show the average gaze trajectories for each question-answer pair, and Figure 2g shows the trajectory averaged across Q-A pairs. The wide variation in trajectories is partially due to shorter and longer questions and answers, which include transitional gazes from previous and following discourse.

In each Q-A pair, observers' gaze shifts from the person asking the question (the *Questioner*) to the person responding to the question (the *Responder*). Observers might do this by shifting their gaze only after the Responder has begun speaking. Alternatively, they could anticipate the beginning of the next turn so that observers are already looking at the Responder *as* her turn begins. This would align with the listening behavior of interlocutors who are actually participating in conversation.

To compare the *reaction* or *anticipation* accounts, we first identified critical time windows during the question-answer sequence for statistical comparison. Since we know that observers reliably gaze at the current speaker during their turn, our windows of interest need to include the region between turns. One way to assess whether observers have shifted gaze toward the Responder is to compare the proportion of looks to the Responder in the beginning of the gap with the proportion at the end of the gap. To account for the planning and execution of eye-movements, we extended this region by 200 ms on either side of the gap since the measurement reflects shifts that were planned before the change in gaze (Fischer & Ramsperger, 1984; Griffin & Bock, 2000). If observers anticipate turn beginnings with their gaze, then their looks to the Responder should increase between the window 200 ms before the gap and 200 ms after the gap.

A visual analysis of the gaze trajectories in Figures 2b-f demonstrates that observers generally shift their gaze to the Responder before she begins speaking. Furthermore, some of these shifts are happening at the very beginning of the gap (or even earlier), indicating that observers may also make anticipatory looks to the next speaker while the current speaker is finishing her turn.

These general observations were confirmed by comparing the average proportion of looks to the Responder in the 200 ms window before and after the gap using paired Wilcoxon signed rank tests. Since our hypothesis makes a strong directional prediction that looks to the Responder will increase before that speaker's turn, we report one-tailed p values. Averaged across Q-A pairs, looks to the Responder increased from 24% in the 200 ms window before the gap to 66% in the 200 ms window after the gap (p=.03). This comparison was significant at p=.03 for each of the five Q-A pairs except for

[2]Some participants also found the second clip confusing (comprehension score average = 87.5% overall, but only 72.2% for those who had not seen the movie before).

[3]91% of disagreements were between 'center'/'unclear' and 'right'/'left' codes (not 'right' vs. 'left').

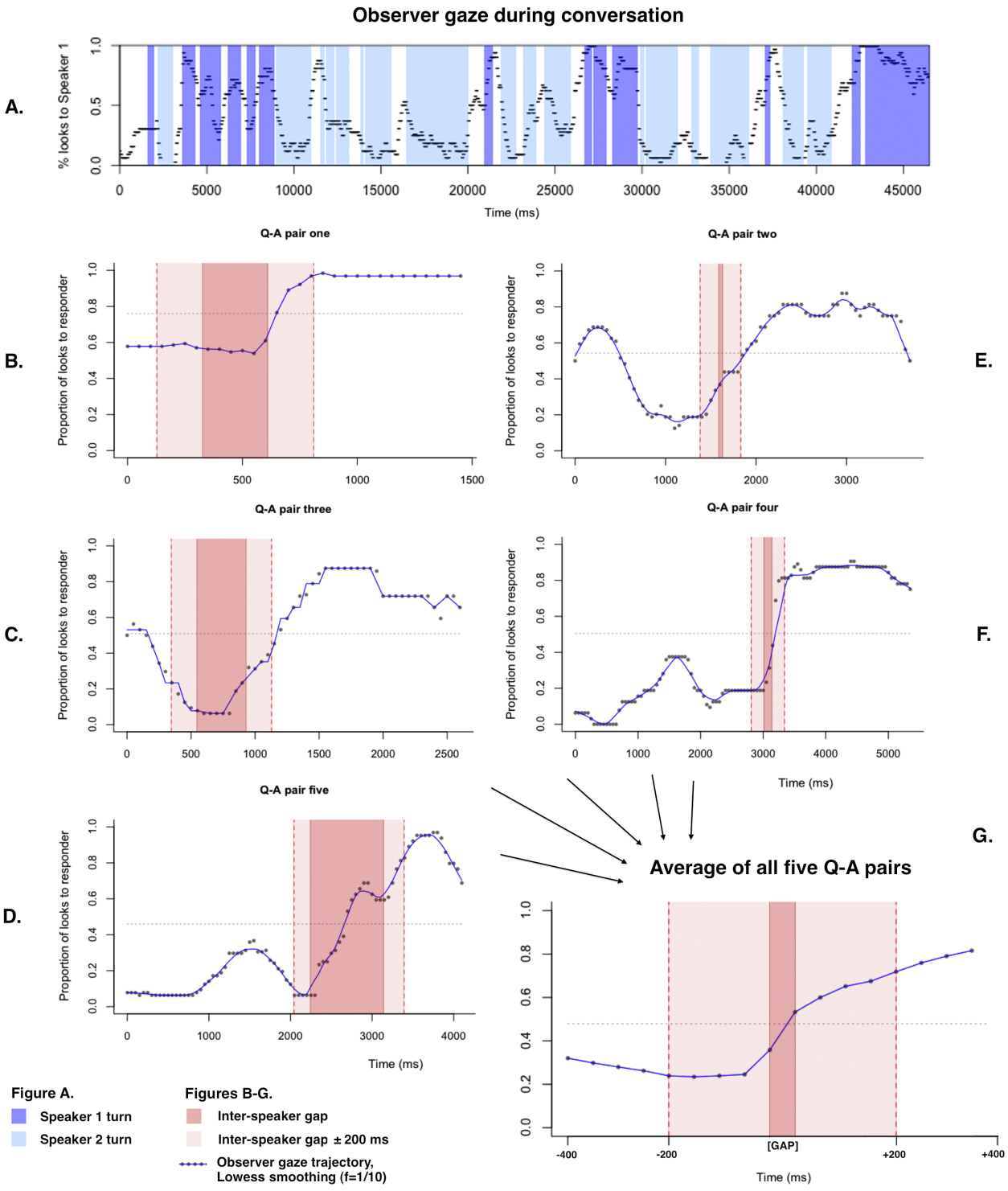

Figure 2: Gaze trajectories averaged across participants for (a) the entire conversation, (b)-(f) each Q-A pair, and (g) an average of the five Q-A pairs. In (a), each speaker turn is highlighted by light and dark shaded regions. Figures (b)-(f) plot proportion of looks to the Responder at each 50 ms interval, with shading over the duration of the gap. 200 ms before and after the gap are lightly shaded, and indicate time windows on which analyses were made. In (g), the gaze trajectory is for all five pairs. The gap in this figure has been collapsed into a single, average duration.

pair one, which was marginal (p=.07).

These analyses suggest that observers do, indeed, anticipate the beginning of the Responder's turn. However, there is still the question of whether the data show any evidence that observers are not *only* anticipating the beginning of the Responder's turn, but *also* anticipating the end of the Questioner's turn. If observers are anticipating the end of the Questioner's turn, then we would expect looks to the Responder to increase within the first 200 ms of the inter-speaker gap, since their eye movements must be planned in advance. Since three of our Q-A pairs (one, two, and four) have gaps of length less than 200 ms, our previous analysis already provides suggestive evidence that this anticipation is taking place (the difference between the pre- and post- gap windows was significant for pairs two and four, p=.03). For the two pairs with gaps longer than 200 ms, we can compare the first 200 ms of the gap with the previous 200 ms. These Q-A pairs do not provide evidence that observers anticipated the end of the Questioner's turn: for pair five, the difference trends in the right direction (p=.09) and for pair three, the looks to the Responder actually decreased significantly, (p=.03).

Discussion

In this preliminary analysis of observer gaze, we were interested in two primary questions: (1) Do third-party observers reliably track current speakers with their gaze? (2) Do third-party observers anticipate transitions to the next speaker?

Our analyses suggest that observers *do* reliably track current speakers with their gaze, and that they often do so *before* the Responder begins talking, and even sometimes before the Questioner finishes talking, though evidence for this is more mixed. These preliminary data are promising given the limited, naturalistic stimuli and the minimal task. Our result confirms previous work showing that listeners actively project upcoming turns, but it does not **yet** build on this methodology to investigate relatively untouched aspects of turn-boundary projection (e.g. continuous cue usage, developmental study, et cetera.) But, the methodology we demonstrate here may bring these results sooner and more accurately than present button-pressing methodologies can.

There are hints in our present data that such insights can be gained. The gaze patterns on a few Q-A pairs show deviations that we might have expected *a priori*, given previous research. Q-A pair four is structured analogously to the "Aunt Millie" example above, since a number of cues (intonational, syntactic, semantic, and more) could lead participants to 'false alarm' and project that the turn will end earlier than it actually does. The resulting gaze behavior is exactly what we expected: less than a second into the turn, observers start looking to the Responder, recovering as the Questioner (unexpectedly) continues with her turn. Similarly, the questions in pairs one and three could be interpreted as alternatives to a full turn (i.e. a ritualized greeting and a backchannel[4], respectively). These questions contain little information, and we might suspect that the lower proportion of looks to the Questioner is due to the observers' preference for looking to the current speaker. Finally, the gap in Q-A pair five is extremely long (949 ms), and observers appear to actually start looking back to the first speaker midway through the gap. This supports work showing that silences over a second may be perceived as "trouble" in a conversation (Jefferson, 1989; Brennan & Williams, 1995).

We did not design or select our Q-A pairs to satisfy these properties and so the previous description is post hoc. But, the existence of these looking patterns increases our confidence that observer gaze is a robust measure of online turn projection. Observer gaze could be manipulated by *controlling* some of these properties, but it will take further research to determine which ones are most important.

Conversations on video

There are two issues one might have with our stimuli. The first is that the video stimuli are from a scripted dialogue in a film which is meant to be entertaining, and might conceivably exaggerate conversational cues. In future versions of this work, we plan to replicate these effects without the "director" effect of using a Hollywood film[5]. We are in the process of collecting naturalistic conversations between strangers, both in split-screen and co-present conversational situations to use as stimuli.

Second, there is some concern about how to interpret looking behavior to a recorded conversation since, though observer gaze is a passive and naturalistic behavior, it does not replicate exactly the experience of being a first-person interactant in a conversation. Not only may a third-party role affect participant engagement in unpredictable ways, but recent work has shown that interactive features of conversational gaze, such as mutual gaze, may affect what information speakers take away from the conversation (Richardson & Dale, 2005). This effect may extend to the information they attend to in projecting upcoming turn-end boundaries. Fortunately, it is increasingly possible to measure observer gaze in interactive first-person experiments thanks to developing technology in minimally intrusive eye-tracking systems.

On the other hand, since we know relatively little about how turn projection is accomplished, it may work to our advantage to leave the complicating factors of first-person dialogue to future work. By this time, our methodology could be well-enough established to make firm predictions about gaze as a continuous measure of certainty about upcoming turn-end boundaries, and the anticipation of upcoming turn-beginnings. Finally, gaze measures may prove to be complementary to button-pressing techniques, since each provides

[4]The analyses presented here do not include a way of separating these 'false alarm' instances from others, but we expect to further develop the analytical tools to do so as we develop the method more generally.

[5]We have collected a second version of this experiment in which observers are given no sound. Observers' eye behavior did not replicate the findings reported here, suggesting that the source of these effects was not the "Hollywood" visual effects and editing.

distinctly different information, but measures behaviors that occur simultaneously in everyday conversation.

Future directions

Observer gaze is a promising new methodology for pinning down the cognitive processes involved in turn-end boundary anticipation. One immediate goal for is to use the naturalistic, spontaneous stimuli that we are currently collecting to replicate the results of this experiment, while adding phonetic manipulations of the sort in de Ruiter et al. (2006). Following much of the previous work on turn-taking, we have focused on question-answer pairs (e.g. Stivers et al., 2009). Questions almost always elicit a response, which make them easier to study than other turn-transitions. With this new methodology we intend to investigate a more diverse set of turn transitions in upcoming work. Finally, the immediate application of observer gaze as studied here is for adult turn-taking behaviors. However, there are several other areas of study, such as child development, second language acquisition, and cross-cultural interaction, that could use this method for investigating turn-taking and other interactional and conversational phenomena.

Observer gaze presents new opportunities to explore how we manage to coordinate with others in interaction—in this case, taking turns in conversation. For decades, the study of conversational timing and turn-taking has been held up for a lack of on-line processing measures. One of this method's most promising features is that it measures a behavior that participants already engage in spontaneously. By capturing this natural behavior in the lab, we may be able to elucidate some of the mechanics of turn processing.

Acknowledgments

This research was supported by an NSF Graduate Research Fellowship to M.T. We thank Eve V. Clark, Herb H. Clark, Paul Thibodeau, Chigusa Kurumada, Michael Frank, and Shawn Tice for their helpful comments, as well as the four anonymous reviewers of our paper. We also thank our coders: Andy Lesser, Laura Yuen, and Armine Pilikian.

References

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

Brady, P. (1968). A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, *47*, 73–91.

Brennan, S., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383–398.

Brown-Schmidt, S., & Tanenhaus, M. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, *31*, 251–276.

de Ruiter, J., Mitterer, H., & Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*, 515–535.

Fernald, A., Thorpe, K., & Marchman, V. (2010). Blue car, red car: Developing efciency in online interpretation of adjectivenoun phrases. *Cognitive Psychology*, *60*, 190–217.

Fischer, B., & Ramsperger, E. (1984). Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, *57*, 191–195.

Ford, C., & Thompson, S. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. A. Schegloff & S. A. Thompson (Eds.), *Interaction and grammar.* Cambridge, MA: Cambridge University Press.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, *25*, 601–634.

Gredebäck, G., Johnson, S., & Hofsten, C. von. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*, 1–19.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *129*, 177–192.

Jaffe, S., & Feldstein, S. (1970). *Rhythms of dialogue*. New York: New York: Academic Press.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: an interdisciplinary perspective.* England: Multilingual Matters Ltd.

Jokinen, K., Nishida, M., & Yamamoto, S. (2009). Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd international universal communication symposium.* New York, NY: ACM.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63.

Kidd, C., White, K., & Aslin, R. (2011). Toddlers use speech disuencies to predict speakers' referential intentions. *Developmental Science*, 1–10.

Levinson, S. (1983). *Pragmatics*. Cambridge, MA: Cambridge University Press.

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

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematic for the organization of turn-taking for conversation. *Language*, *50*, 696–735.

Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *PNAS*, *106*, 10587–10592.