Cues to turn boundary prediction in adults and preschoolers

Marisa Casillas
Stanford University
Department of Linguistics
middyp@stanford.edu

Michael C. Frank
Stanford University
Department of Psychology
mcfrank@stanford.edu

Abstract

Conversational turns often proceed with very brief pauses between speakers. In order to maintain "no gap, no overlap" turn-taking, we must be able to anticipate when an ongoing utterance will end, tracking the current speaker for upcoming points of potential floor exchange. The precise set of cues that listeners use for turn-end boundary anticipation is not yet established. We used an eyetracking paradigm to measure adults' and children's online turn processing as they watched videos of conversations in their native language (English) and a range of other languages they did not speak. Both adults and children anticipated speaker transitions effectively. In addition, we observed evidence of turn-boundary anticipation for questions even in languages that were unknown to participants, suggesting that listeners' success in turn-end anticipation does not rely solely on lexical information.

1 Introduction

Turn-taking in human communication is efficient: we usually switch between speakers with brief pauses. Though there is a wide distribution of gap lengths in everyday conversation, the median gap between conversational turns is close to zero milliseconds, and maintaining brief inter-speaker junctions may be universal to human languages (de Ruiter et al., 2006; Heldner and Edlund, 2010; Stivers et al., 2009). These gaps, though brief, result in minimal overlap, and beg the question of how we manage to come in with such precise timing.

Sacks, Schegloff, and Jefferson (1974) noted that inter-speaker gaps are too brief for listeners to be relying on turn-end silences before starting up their response. They suggested that instead we track ongoing turns for cues to their eventual end, using linguistic information about syntactic, propositional, and intonational structure. Using these cues, listeners should be able to predict the moment at which a speaker will stop speaking with high accuracy. This insight was important, but they did not further investigate which cues—whether linguistic or non-linguistic–listeners track.

More recent research has addressed this question, investigating which linguistic cues might be most informative in anticipating the close to an ongoing turn. Corpus study of available cues has yielded somewhat inconclusive results since so many linguistic boundaries co-occur (Caspers, 2003; Ford and Thompson, 1996). Even if reliable turn-end cues were apparent, we could not be confident that listeners actually attended to them to in conversation without experimentally manipulating them and measuring their effects on listeners.

De Ruiter and colleagues (2006) created an experimental paradigm to measure turn boundary anticipation while also beginning to test which cues were most informative in this process. They extracted utterances from a recording of a spontaneous conversation and presented them to participants over headphones. Participants were asked to press a button at the moment they anticipated the speaker would stop speaking. Participants were extremely accurate in identifying the moment before a turn was about to end. To test the effects of different cues on anticipation, they separately controlled for the presence of intonation and lexical information. There was no significant differ-

ence between participants' accuracy when intonation was present and when it was omitted from the stimulus. When lexical information was taken away, however, participants' accuracy declined significantly. De Ruiter et al. thus suggested that word-level information is of primary importance in turn-boundary anticipation.

Although the de Ruiter study was carefully controlled, the button-pressing task was explicit, and might easily have focused participants' attention on words and word-level information more so than they would have been otherwise, especially since the instructions asked for precisely-timed responses. If this were the case, their results would reflect a use of linguistic cues under somewhat unnatural conditions. In addition, de Ruiter et al. (2006) did not control for all prosodic cues—duration was left unmodified in their stimuli. This information might have accounted for some of their accuracy effects in the condition without intonation.

Many people have the intuition that intonation and rhythm are part of the prediction process, but may be more important prior to the end of the turn, at which point lexical information may be most informative. Carlson, Hirschberg, and Swerts (2005) showed that listeners can use prosodic cues to predict the strength of upcoming prosodic breaks. The estimation of upcoming prosodic breaks can help listeners determine when a speaker-switch will be appropriate, even without lexical information (Carlson et al., 2005; Heldner et al., 2006). These experiments were run on "offline" judgments, unlike those in the de Ruiter et al. (2006) study—which found no prosodic effects. Could prosodic effects emerge during online speech processing under different experimental circumstances?

Our current work uses eye-tracking as an implicit measure of turn boundary anticipation. This method allows us to study both adults and children and to systematically manipulate the content of the videos we track.

Tice & Henetz (2011) explored eyetracking as a possible alternative method for measuring online turn processing, which they call Observer Gaze. They seated participants in front of a large screen, under which was tucked a small digital video camera tilted toward participants' faces. While viewing a one-minute dyadic, split-screen conversation in English, participants consistently tracked the current speaker with their gaze. In addition, they anticipated the ends of turn boundaries by looking at the next speaker on question-answer pairs. Observer Gaze is founded upon natural looking behavior—observers tend to look at the current speaker during his or her turn (Kendon, 1967; Bavelas et al., 2002). It requires little or no instruction and allows experimenters to collect high temporal resolution looking data over the course of a conversation. Thus, this method provides a measure of turn-boundary anticipation that we can use to investigate the cues that contribute to this ability.

Since it is a passive method founded on natural, spontaneous behavior, Observer Gaze can be used with both child and adult participants to begin exploring the developmental trajectory of turn-end boundary prediction. We are interested in comparing adult turn-end prediction skills with those of children because of the protracted development of turn-taking. By age five, children's turn-taking skills are still not up to the timing standards of adults. Even in adjacency pairs, when the response is often restricted and the context makes clear who the next speaker is, children's responses are still delayed. It has been proposed that their delay is due to complexity and predictability level of responding to the question at hand (Garvey and Berninger, 1981; Casillas et al., in preparation), but we do not yet know whether children's delay is due to the need to formulate a response or a slowly developing ability to predict turn-end boundaries. The eye-tracking method described above makes it possible to compare adults and children directly, allowing for investigation of this question in our study.

In the current study, we introduce a simple method for controlling word-level information in the speech signal: we show participants videos of languages that they do not speak. Though the non-lexical signals in the videos (e.g., intonation, prosody, gaze, gesture) are foreign to the participants, the information may still be robust enough to support online turn-tracking. Because the linguistic cues are foreign, eye gaze behavior while watching a foreign language (which has similar, but not identical cues) is a stringent test of the use of non-lexical cues in online-turn-processing. To keep the stimuli engaging for children, we used child-oriented speech (as described below) in the video stimuli.

2 Methods

2.1 Participants

Seventy-two pre-school aged children (19 three-year-olds, 32 four-year-olds, and 21 five-year-olds) and 11 adults participated in the study. All were native speakers of English who had little to no language experience with the four non-native languages used in the stimuli (see Procedure below).

2.2 Materials

The video segments were recorded in a sound-attenuated booth by two native speakers of each language (all non-native English speakers enrolled in graduate study in the U.S.) Each person was audio recorded from a lapel microphone (one on the right channel and one on the left) feeding into a Marantz PMD 660 solid state field recorder. Participants were video-recorded from the iSight of a MacBook. Pairs of speakers were selected by native language, and ranged from acquainted individuals to good friends. They were asked to speak on four topics for 20 minutes (five minutes each on favorite foods, entertainment, hometown layout, and pets). Following this recording they were asked to choose a topic relevant to young children (e.g., riding a bike, eating breakfast, siblings) and improvise on that topic as if they were on a children's television show until they had at least 30 seconds of continuous material. Most pairs took less than three minutes to record these "child-friendly" improvised conversations, and the resulting recordings remained natural but engaging for both young children and adults. The audio and video recordings were aligned afterward using video editing software.

The child-friendly videos were then edited to include 30 seconds from each language with maximal turn activity and were wedged between entertaining filler videos (e.g., running puppies, singing muppets, flying bugs) for an experimental duration of approximately six minutes long. The order of the non-English videos (videos 2–5) was varied in four versions of the experiment so that no consistent order effects might skew the data. The first and last videos in English (videos 1 and 6) were always kept the same.

2.3 Procedure

Participants were seated in front of an SMI 120Hz corneal reflection eye tracker and a large screen with speakers placed on a table at each side of the screen. The eye-tracker is mounted beneath a flat-panel display; the display is in turn mounted on an ergonomic arm so that it can be positioned at a comfortable height approximately 60cm (an adult arm's length) from the participant. After being seated, participants were told that they would hear videos in a number of different languages. We then asked each participant what languages they could speak. We used a 5-point calibration routine in which participants followed a point on the screen with their eyes. For purposes of engaging children, Elmo (an animated puppet) was used as the calibration image.

In the body of the study, participants watched a six-minute video containing six 30 second dyadic conversations with 15–30 second filler videos between them. The first and last conversations (numbers 1 and 6) were in American English and the intervening conversations (2–5) were recorded in Hebrew, Japanese, German, and Korean. After each conversation, adult participants were asked if they understood any part of the speech to make a second check for any lexical access during the non-English videos.

3 Results and discussion

Child and adult observers in both the English and non-English videos were more likely to keep their eyes on a speaker when that person was speaking rather than when they were silent (Table 1), though they also glanced back at silent participants between 15 and 20% of the time. Children were less likely than adults to keep their eyes on the current speaker while watching the non-English videos, but still showed a reliable difference in gaze to a speaker during speech and during silence. This result indicates that participants were performing basic turn-tracking with their gaze while viewing the stimuli (Kendon, 1967). When point-of-gaze is averaged across the entire recording in this way, there do not appear to be large developmental differences between children and adults in their ability to track the current speaker, though the adults were slightly more consistent.

We next turn to the question of the quick, an-

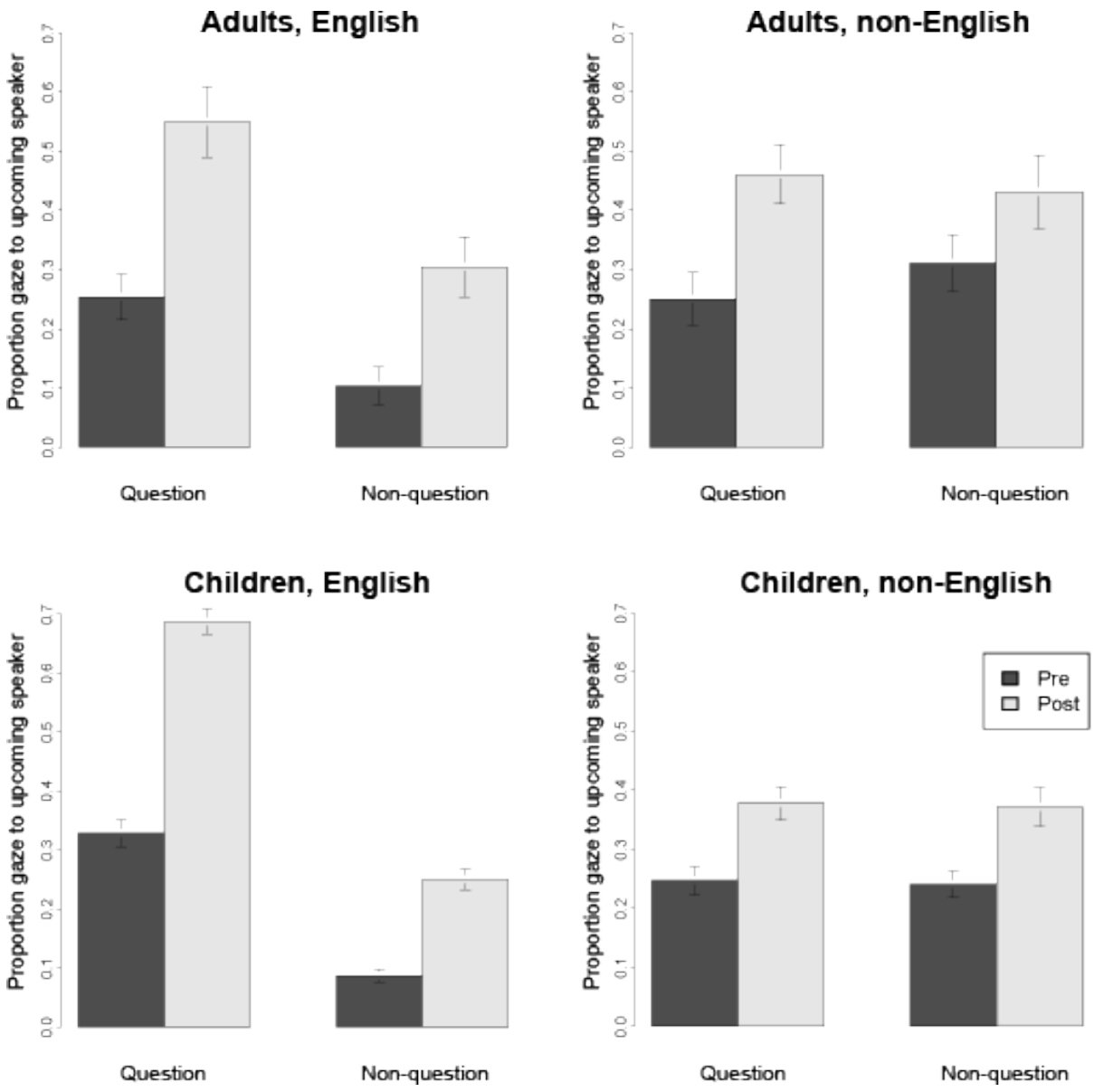

Figure 2: Children and adults' gaze to the upcoming speaker during pre- and post- gap 200 ms windows of speaker switches. Error bars show standard error of the mean across participants.

Group	Language	Current	Non-current
Children	English	0.64	0.17
	Non-Eng	0.48	0.19
Adults	English	0.63	0.16
	Non-Eng	0.61	0.21

Table 1: Average proportion of gaze during speech segments to the current and non-current speaker. Child and adult observers look to the non-current speaker 16–20% of the time the current speaker is talking, and look at neither speaker 18–33% of the time the current speaker is talking. Children watching non-English videos were least likely to be looking at the current speaker during his or her speech.

ticipatory eye-movements around conversational turns observed in previous work (Tice and Henetz, 2011). We test for the presence of turn-end anticipation by measuring shifts in gaze near the inter-speaker gap. Using the average direction of gaze (between previous and upcoming speakers), we compare the 200 ms window prior to the onset of an inter-speaker gap and the 200 ms window following the offset of that inter-speaker gap. Since it would take adults and children at least 200 ms to plan an eye movement, any significant shift in gaze during the 200 ms post-gap window indicates a movement planned prior to the onset of speech by the second speaker. Using this comparison, we find that while viewing English and non-English stimuli, participants tend to anticipate upcoming turn-end boundaries such that they spontaneously shift between the current and previous speaker before the previous speaker has the opportunity to begin his or her response (Figure 1).

Speaker exchanges in the non-English videos that sounded similar to English question-answer adjacency pairs[1] were coded as "questions" for the analysis. We find that both adult and child observers show divergent performance on question and non-question exchanges during all of the videos. Though their gaze begins to shifts dramatically in nearly every case across the pre- and post- gap windows, participants show an advantage for question-answer pairs such that they are more likely to shift earlier on and already be looking at the answerer when he or she begins to speak (Figure 2).

This behavior indicates spontaneous response anticipation during online processing of the stimuli. The average inter-speaker gap across languages and exchange types was 335 ms. The average for questions across languages was 319 ms and 350 ms for non-questions, though the stimuli contain many cases of sub-200 ms inter-speaker gaps. This means that listeners may still rely on a turn-end pause in some cases. However, if participants were universally reacting to silence, we would not expect the earlier switch in question-answer pairs. More generally, the pattern of reliable performance even with inter-speaker gaps shorter than 200 ms suggests that participants make use of cues that are present in the signal prior to the turn-end silence.

We fit two separate linear mixed-effects models (Gelman and Hill, 2007) to participants' average gaze direction at pre- and post- gap windows: one model for adult data and another for the child data. We used a maximal random effects structure to control for variability between participants on the variables of interest. Model coefficients suggest that the advantage for questions over non-questions was significant or nearly significant for both children and adults (t=-7.03 and -1.76, respectively). For children, there was also a significant effect of language group (English vs. non-English, t=-9.29) and a significant interaction between language group and turn type (question vs. non-question, t=6.27). The effect of language group was also nearly significant in the adult data (t=-1.77), and there was no interaction between language group and turn type.

These statistical results suggest that adults were able to integrate non-native cues in their online turn processing more effectively than children were, providing some guidance for an account of the development of turn-end anticipation. For both age groups, there was a significant effect of turn type: question vs. non-question. There may have been many divergent cues in these cases which led participants to earlier and more successful anticipation in the presence of questions. However, since the determination of what counts as a "question" in the non-English videos mainly relied on prosodic similarity to English questions, we have reason to believe that it is precisely because speakers rely on intonational

[1] Judgments were made by the first author primarily based on auditory information, including but not limited to a rising intonation. This judgment is meant to represent which switches the participants were most likely to think were questions in the non-English videos.

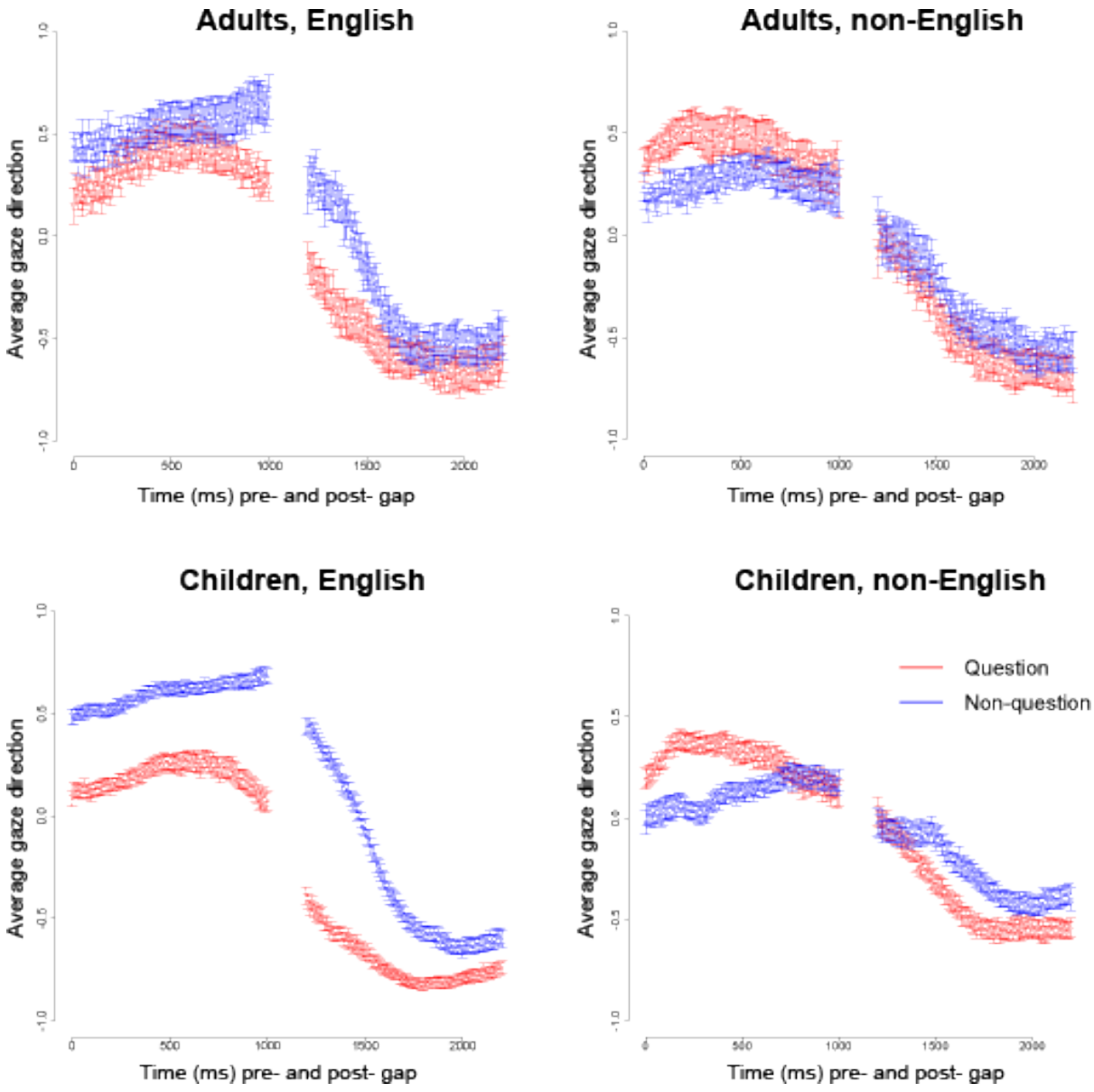

Figure 3: Children and adults' trajectory of gaze over the preceding and following 1-second window of inter-speaker gaps for questions and non-questions in English and non-English videos. Error bars show standard error of the mean across participants.

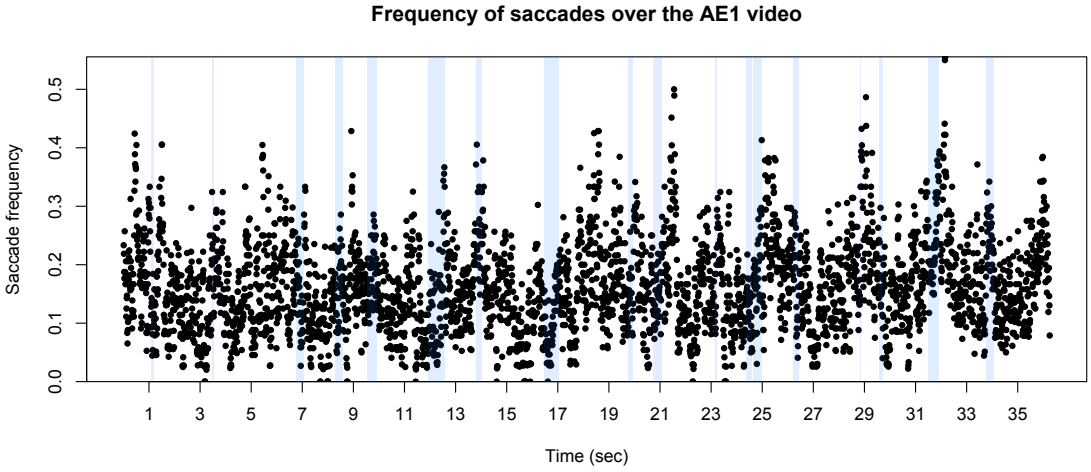

Figure 1: Frequency of saccades over the time course of one video in English. Vertical bars in blue indicate inter-speaker gaps.

Predictor	β	SE	t
Children			
Switches (Non-Question)	-0.31	0.04	-7.03
Lg group (Non-English)	-0.43	0.05	-9.29
Switches x Lg group	0.33	0.05	6.27
Adults			
Switches (Non-Question)	-0.17	0.1	-1.76
Lg group (Non-English)	-0.16	0.09	-1.77
Switches x Lg group	0.005	0.07	0.07

Table 2: Average direction of gaze to the each speaker while he or she is speaking and silent. 1 = looking exclusively at the current speaker and -1 = exclusively at the non-current speaker.

information that they show this advantage. Thus it would be inaccurate to characterize online turn-processing as solely dependent on lexical information. Rather, participants perform remarkably well when no lexical information is present at all.

Consistent with our previous work, the current results provide us with further empirical evidence for spontaneous anticipation of turn-end boundaries. Our results were calculated for without distinction between fixations, long movements, and saccades because of the frequent sampling of the tracker and our decision to analyze anticipation by averaging over pre- and post- gap windows. The anticipatory looking behavior we observed is unlikely to be due to continuous gaze shifting during the video, since saccades show spiked increases only near potential turn boundaries, not between. For example, a time-course rendering of eye-tracking data from one representative video of English conversation shows a considerable spike in saccades prior to turn gaps (Figure 3). Thus, we do not believe that random shifting accounts for our results.

Because each non-English language in this experiment is represented by a single stimulus, we cannot compute reliable across-language differences for each language. Since some of the languages have more overlap in linguistic structure with English, gaze behavior may be significantly better on these items. For example, English speakers can make predictions about the strength of upcoming Swedish prosodic boundaries nearly as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). A follow-up study of our work using eye tracking with multiple items from each language would enable us to check for effects of linguistic similarity in languages that the participants do not actually speak.

Finally, in the current study we did not include a baseline condition with no linguistic information at all. Tice & Henetz (2011) found that successful gaze anticipation relies on the presence of linguistic information for English. But, we have no direct comparison of gaze behavior in conditions without any linguistic information and with linguistic information in a language the participants don't speak. This must be added in future work.

4 Conclusion

Children and adults track the current speaker with their gaze. They also spontaneously make anticipatory looks to upcoming speakers at speaker exchanges, indexing their online processing of turn-structure (their anticipation of an ongoing turn's end and the beginning of a responder's turn). Their anticipatory gaze is stronger when prosodic and other non-lexical cues suggest question status (e.g., ending in a high-rise terminal).

Even without lexical information, we track turns as they unfold. Participants not only continued to track current speakers during non-English videos, they showed an advantage for question-type turns over non-question-type turns. A model of how we manage to take turns on time must account for prosodic and other non-lexical information.

We found that adults and children performed almost equally well, with the exception that children had more difficulty maintaining speaker tracking and anticipation during the non-English videos. This may in part be due to their uninhibited lack of interest which resulted in more variable looking patterns than well-behaved adults. Investigation of this possibility will require more data from both age groups and denser developmental data. Children's success in predicting turn-end boundaries and tracking the current speaker suggests that they master this skill early on. It therefore seems likely that their delays in responding to questions (Garvey and Berninger, 1981; Casillas et al., in preparation) has more to do with formulating a response than anticipating when to come in.

In the present study we used recordings of non-English languages to test for turn-processing success when lexical information is not present. Though the non-lexical stimuli are highly naturalistic, they do not directly test which *English* cues English speakers use. There is a significant effect of language group for child participants and a similar, but non-significant effect for adults, suggesting that we can most accurately measure turn-processing performance in English by using English stimuli. To perform the appropriate experiment, we must create phonetically-manipulated stimuli to control for turn-end linguistic cues including prosody and lexical information. We plan to run this follow-up study to compare how performance changes with carefully controlled, but less naturalistic stimuli.

Until recently, we did not have any experimental evidence of turn-end anticipation. But, in the past few years at least two studies have demonstrated that turn-end prediction is a measurable behavior (de Ruiter et al., 2006; Tice and Henetz, 2011). The present study is the first to show evidence that we spontaneously predict turn-end boundaries when attending to languages that we do not speak. This result tells us that the ability to predict upcoming turn-end boundaries is not reliant on lexical information alone; rather, we spontaneously apply (even non-native) prosodic and non-verbal information to continue tracking upcoming turn junctures accurately. Taking all of the experimental work on turn-end anticipation together, our turn processing mechanism is best characterized as a flexible one which makes use of the information available to it in the current conversational environment. These findings indicate that further experimental work will be able to distinguish what cues are attended to as speech unfolds and prediction takes place under different conditions.

5 Acknowledgements

This work is supported by an NSF GRF to M. Casillas. We thank the children, teachers, and directors at the preschool where the child participants were tested. We also thank Tania Henetz, Eve V. Clark, and Herb Clark for their invaluable input.

References

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

R. Carlson, J. Hirschberg, and M. Swerts. 2005. Cues to upcoming swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, (46):326–333.

M. Casillas, S. Bobb, and E. V. Clark. in preparation. Turn-taking, timing, and access in early language acquisition.

J. Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31(2):251–276.

J.-P. de Ruiter, H. Mitterer, and N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515.

C.E. Ford and S.A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

C. Garvey and G. Berninger. 1981. Timing and turn taking in children's conversations 1. *Discourse Processes*, 4(1):27–57.

A. Gelman and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*, volume 1. Cambridge University Press New York.

M. Heldner and J. Edlund. 2010. Pauses, gaps, and overlaps in conversations. *Journal of Phonetics*, (38):555–568.

M. Heldner, J. Edlund, and R. Carlson. 2006. Interruption impossible. In G. Bruce and M. Horne, editors, *Nordic Prosody: Proceedings of the IXth onference*, pages 225–233. Frankfurt am Main: Peter Lang.

A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22–63.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

T. Stivers, N.J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J.P. De Ruiter, K.E. Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

M. Tice and T. Henetz. 2011. Turn-boundary projection: Looking ahead. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.