The development of predictive processes in children's discourse understanding

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

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

Abstract

We investigate children's online predictive processing as it occurs naturally, in conversation. We showed 1–7 year-olds short videos of improvised conversation between puppets, controlling for available linguistic information through phonetic manipulation. Even one- and two-year-old children made accurate and spontaneous predictions about when a turn-switch would occur: they gazed at the upcoming speaker before they heard a response begin. This predictive skill relies on both lexical and prosodic information together, and is not tied to either type of information alone. We suggest that children integrate prosodic, lexical, and visual information to effectively predict upcoming linguistic material in conversation.

Keywords: Prediction; online comprehension; turn-taking; timing; child language; prosody; eye-tracking

Introduction

Conversation is the primary way we use language. It is most often spoken face-to-face with two or more speakers, and is deeply embedded in our current interactional context. Participants in conversation don't just listen; given that inter-turn gaps are so brief, speakers must be simultaneously planning at least part of their response while the current speaker is still talking (Sacks et al., 1974; Stivers et al., 2009). So under normal conditions, listeners deal with critically different processing pressures during conversation than they do in rigidly controlled experiments. For children especially, experience and skill in processing and conversation can be critical to later language development (e.g., Weisleder, 2012). The current study seeks to draw a link between language processing in the lab and language processing in broader contexts by tracing predictive processing during conversation across a broad developmental sample.

Timing is critical in conversation because, in addition to parsing the linguistic signal for its parts and meanings, conversational participants are interested in the upkeep of the ongoing interaction. For example, if someone asks you, "What are your plans for dinner?" you are obligated to do more than just parse the linguistic signal; you must respond. You can't just respond at your convenience either—especially for an implied invitation like this one, a slight hesitation might communicate that you will turn the offer down. This is a substantial cognitive load to bear since speakers must quickly figure out what was said and how to respond. Predictive processes can help maintain the flow of conversation by allowing us to plan for what is likely to happen next in the interaction.

When listening to a single utterance, we make predictions about what the speaker will say next. Many studies have shown that we can use a wide variety of linguistic and non-linguistic cues to incrementally update our expectations about what linguistic material will follow (e.g., Altmann & Kamide, 1999; Ito & Speer, 2008; Snedeker & Yuan, 2008).

In multi-utterance contexts, like conversation, speakers must also coordinate their ongoing actions, and so our predictive prowess is even more to our advantage than it is in the lab. There is both naturally-occurring and experimental evidence that adults effortlessly anticipate when a speaker switch will occur during conversation. In order to respond with brief gaps, they need to accurately predict when to begin speaking (Sacks et al., 1974). There is also experimental evidence that adults can anticipate upcoming turn-structure—when asked to press a button when they think a speaker will finish her turn, adult listeners demonstrate incredible timing accuracy (M = 168 ms from the offset of speech; de Ruiter et al., 2006). They also spontaneously track turn-timing and anticipate upcoming speakers with their gaze when watching videos of conversation (Tice & Henetz, 2011).

Here we ask: do children also make online predictions about conversation? We focus on how children process the multi-utterance speech around them because, as mentioned, children's conversational skill and experience can influence their later language learning.

Children learn language in the context of conversation, and their conversational skills allow them to practice comprehending and using language with others. Children begin to take turns in early infancy, but their coordination of turn-timing with others takes several years to develop. By four months, infants regularly engage in coordinated back-and-forth interactions with their caregivers (Masataka, 1993). Twelve-month-old infants who are watching two-person conversations can (1) track who is speaking, and (2) expect speech to be responded to verbally (Thorgrímsson et al., 2011). Despite this, even at 5;0, children's timing is significantly delayed compared to adults'—their response delay at 3;0 is up to 10 times slower (Casillas et al., in prep)—leading some to believe that children can't or don't perform the same predictive processing that adults do (Garvey & Berninger, 1981).[1]

We propose that, on the contrary, children develop their predictive turn-taking skill early in life, and that their apparent delay is due to the time needed to plan and execute a response. Thus, when children simply observe an ongoing interaction, they show predictive timing similar to adult norms. Casillas and Frank (2012) found that when children and adults watched videos of conversation in a language they didn't speak, they were able to use the available information

[1] Cf. Snedeker & Yuan, 2008 for more on children's sentence processing.

(prosodic, temporal, and visual) to track and anticipate the ongoing turn structure with their gaze. Because some linguistic units are more informative than others in predicting turn-boundaries (e.g., words > intonation; de Ruiter et al., 2006), we also hypothesize that, like adults, children's online predictions about turn-taking are more heavily influenced by lexical information than they are by prosodic information. By testing these proposals we can (1) track children's development of predictive turn processing during discourse while (2) also beginning to tease apart which linguistic cues children attend to in making their predictions about turn-structure. We measured children's online anticipation about who will speak next in conversation and found that children use multiple linguistic cues to make accurate predictions about what will come next—and they do so even at 1– 2 years old.

Method

We tracked children's eye movements as they watched short videos of conversation to measure their predictive gaze to upcoming speakers at points of speaker-transfer. We controlled the audio signal to limit children's access to either prosodic information or lexical information, making comparisons to their gaze behavior in normal audio conditions and conditions without any linguistic information. We focus here on effects of linguistic information, so we eliminated visual cues to turn-taking by using videos of puppets to replace the original videos of our speakers.

Participants

We recruited 129 children ages 1;0–7;0 from the Children's Discovery Museum in San Jose, CA, to participate in the current study. We collected data from 20-23 children for each of the six 1-year age groups. All participants were native English speakers, though some parents reported that their child heard a second (and sometimes third) language at home.[2]

Materials: Puppet videos

Audio-recordings We recorded six 20-25 second two-person conversations for use in the puppet videos. Each of the six conversations featured a native English-speaking male and female talker. Talkers were directed to improvise a short conversation on a given topic (one of: 'riding bikes', 'pets', 'breakfast', 'birthday cake', 'rainy days', 'the library'). We asked talkers to talk "as if they were on a children's television show" to establish a child-friendly style. We gave talkers approximately five minutes to work out a basic conversation and then perform it with minimal practice. We edited each conversation to a 20-25 second clip for use in the final video stimuli.

[2]The 27 bilingual children heard and used English at least 50% of the time, as reported by parents. This proportion is representative of the area where we collected data, which has a large population of fully- or partially-bilingual speakers. We replicated all analyses below, excluding bilingual speakers and saw essentially no difference in the qualitative or quantitative pattern of results reported below.

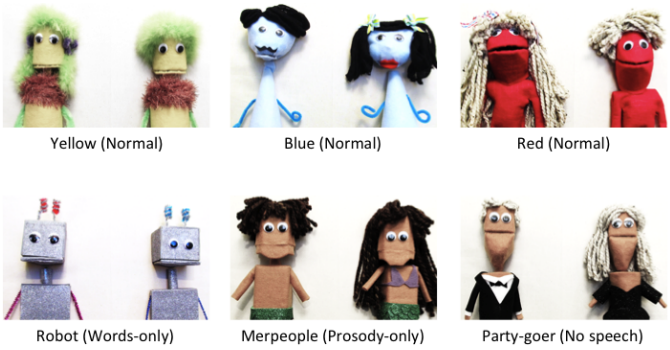

Figure 1: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

Audio Manipulation To control for the linguistic information available in the final puppet videos, we phonetically manipulated the recordings to fall into four conditions: Normal, Words-only, Prosody-only, and No Discernible Speech. *Normal* videos simply used the 20-25 second audio recording. *Words-only* videos featured manipulated speech in which intonation was flattened to each talker's average pitch (F0) and every syllable nucleus and coda duration were set to each talker's average nucleus and coda duration.[3] To do this we used PSOLA resynthesis in Praat (Boersma & Weenink, 2012). The resulting audio signal was devoid of pitch and durational cues to turn-boundary, so we referred to this audio as 'robot' speech when talking to children. *Prosody-only* videos also featured manipulated speech, in which the original audio recording was low-pass filtered at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006). Low-pass filtering removes the phonetic information used to distinguish between phonemes, and so the resulting audio has no identifiable words, but retains the original intonational and rhythmic qualities of the conversation. Low-pass filtered audio sounds muffled, like voices under water, so we referred to this audio as 'merperson' speech. To create *Non-discernible speech* audio, we overlaid eight different child-oriented conversations (not including the original one) to create multi-talker babble. This is sometimes referred to as 'cocktail party' speech, but we referred to it as 'birthday party' speech. Finally, the *Prosody-only* audio sounded much quieter than the other conditions because it lacked acoustic energy above 500 Hz, so all other audio conditions were adjusted to match its lower volume.

Video-recordings We then created puppet video-recordings to match the final audio signals. The puppets were designed to be minimally expressive so that the experimenter could only control the opening and closing of their mouths. There were three *Normal* condition puppet pairs—'red',

[3]We excluded occasional emphatically lengthened monosyllabic words like [waʊː] 'woooow!' from the calculation of the average and the resulting length manipulation.

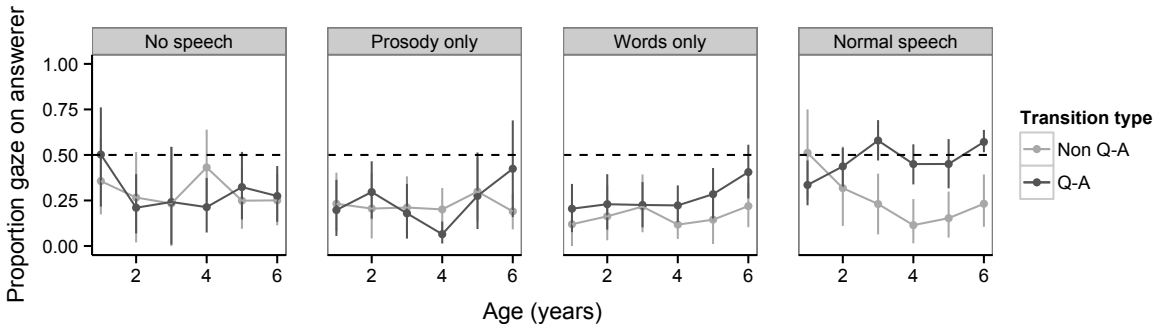

Figure 2: Proportion gaze to the answerer during the first 333 ms of the answer. Age in years is plotted on the x-axis for each of the four conditions (Question-Answer switches = dark gray; Non-Question-Answer switches = light gray). The vertical bars show the 95% confidence intervals around each point.

Condition	Current	Non-current	Elsewhere
No discernible speech	0.51	0.15	0.35
Prosody only	0.55	0.14	0.31
Words only	0.65	0.14	0.21
Normal speech	0.68	0.14	0.18

Table 1: Overall proportion gaze, averaged across all participants, to the current and non-current speakers (and elsewhere) during utterances.

'blue' and 'yellow' ones—and one puppet pair for each of the other conditions: 'robots', 'merpeople', and 'party-goers' (Figure 1). Three conversation topics ('birthday cake', 'pets', and 'breakfast') were used for the *Normal* conversations, and three ('riding bikes', 'rainy days', and 'the library') were used for the other three conditions. We created three versions of the experiment so that each of the six puppet pairs was associated with at least three different conversation topics. We then hand-aligned the final audio to the puppet video recordings and ensured that half of the videos in each version were female-left-male-right and vice-versa by flipping the video and audio channels as needed.[4]

Procedure

We seated children in front of a large screen with speakers placed below and at the sides of the screen. Mounted beneath the screen was an SMI 120 Hz remote infrared eye-tracker that continuously recorded their eye movements throughout the experiment. Children then watched a series of short videos comprising six brief puppet conversations and five engaging filler videos (e.g., running puppies and music). The filler videos were inserted between the puppet videos, which were ordered randomly for each participant. The six puppet videos fell into four audio conditions: *Normal* (3), *Words only* (1), *Prosody only* (1), and *No Discernible Speech* (1). The entire experiment took less than five minutes for most children.

[4]See a sample of the final videos and data from all conditions in one version at: http://langcog.stanford.edu/materials/anticip.html

Data analysis

For each participant in the study, we only included data from those video segments in which the participant gazed at the video for more than 75% of its duration. In prior work (Casillas & Frank, 2012; Tice & Henetz, 2011) adults and children 3;0 and older made anticipatory gaze shifts to upcoming talkers while watching videos of conversation. The shifts sometimes began before the prior turn ended, within the final 300 ms of speech. To determine whether children 1;0-7;0 in our data made similar anticipatory shifts, we conducted our analyses contingent on looks to the prior listener. Specifically, we only included children who were looking at the prior speaker 333 ms before the prior turn ended. This follows contingent looking analyses in other child language work (Fernald et al., 2008) and guarantees that the children in our analyses were prepared make a gaze switch to the upcoming speaker. We then averaged gaze to the upcoming speaker during the first 333 ms of the answer.[5] Since each child in our analysis started by looking at the prior speaker, looks to the upcoming speaker at the answer onset will represent the magnitude of children's anticipatory gaze shifts. Because prior work has found that children shift their gaze more quickly after hearing a question than non-question (Casillas & Frank, 2012), we separated these in our analysis. When gaps are too long they can signal a troubled speaker transition or a disfluency that might need conversational repair (Jefferson, 1974). For this reason we excluded all turn-transitions longer than 550 ms in our stimuli.

Results

In all conditions, participants were nearly three times more likely to keep their eyes on a talker when that person was speaking, rather than when they were silent (Table 1). Par-

[5]We assume here that it takes children ~333 ms to plan an eye movement, following Fernald and colleagues (2008). A significant shift in gaze to the next speaker before 333 ms of speech indicates that the eye movement was planned before the response began. We saw anticipation in all conditions, so below we compare anticipation *across* conditions by analyzing looks to the upcoming speaker at the onset of the response turn.

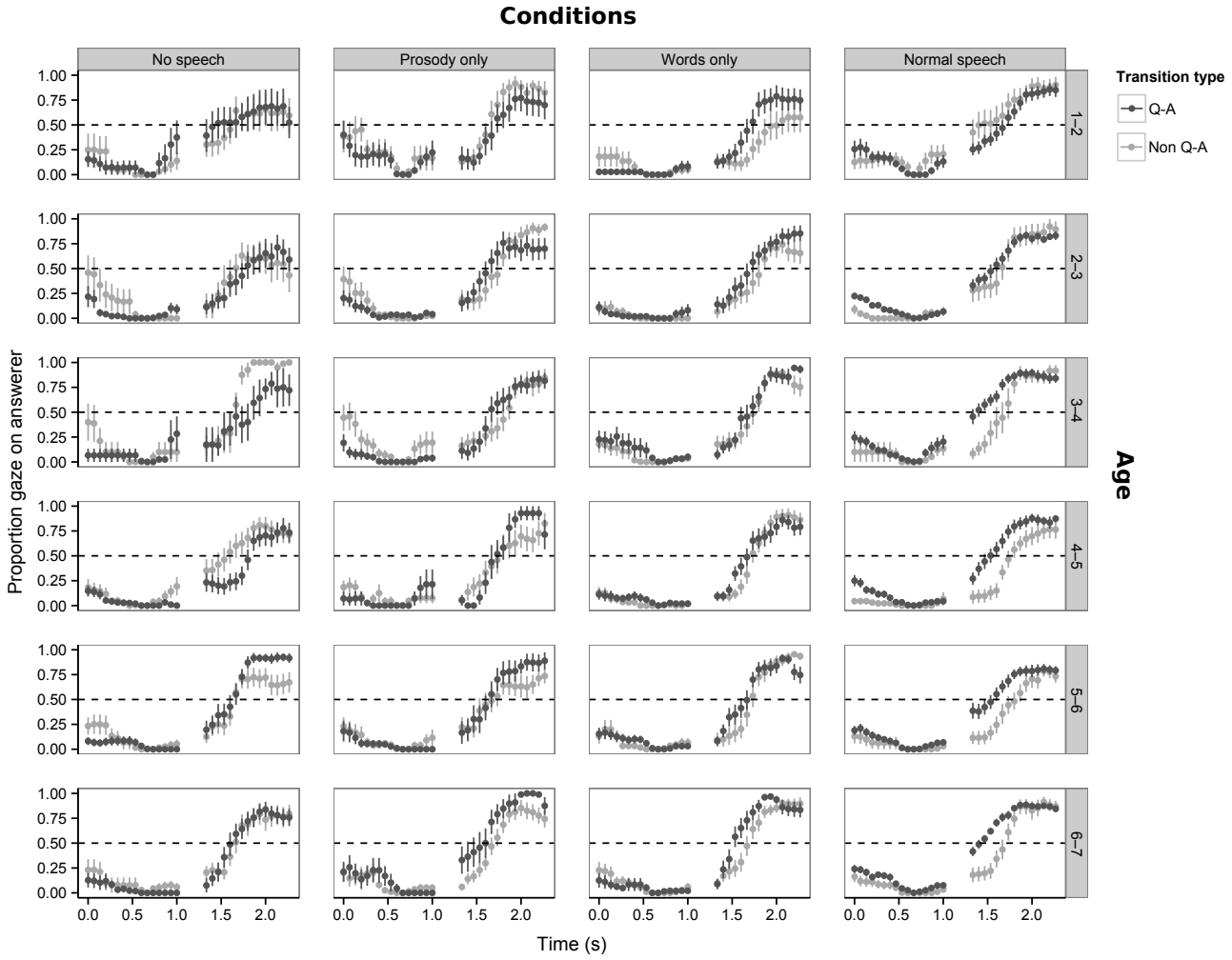

Figure 3: Proportion gaze to the answerer during the last 1 second of the prior turn and the first 1 second of the upcoming turn, broken down by participant age and linguistic condition (Question-Answer switches = dark gray; Non-Question-Answer switches = light gray). Error bars indicate the standard error of the mean. The inter-turn gap is represented by the blank area along the trajectory. Included speaker switches had gaps ranging from 3–497 ms (M=308 ms). The gap shown above is 300 ms. Gaps in the stimuli varied in length, so looking data during this period isn't plotted.

ticipants looked away from the talkers 18–35% of the time. This closely matches our prior results (Casillas & Frank, 2012), though children looked elsewhere more often in the *No Discernible Speech* and *Prosody Only* conditions than in the other two. Children's consistent looks to the current, rather than the non-current, talker suggest that the participants were tracking basic turn-taking with their gaze by using information from the audio, video, or (most likely) both. Participants most consistently looked at the current speaker (and looked away least) in the *Normal Speech* condition.

Children of all ages and in all conditions made anticipatory shifts to upcoming speakers (Figures 2 and 3). Even in the *No Discernible Speech* condition—in which the children saw puppets mouthing words to unrelated multi-talker babble—children shifted their gaze toward upcoming speakers by the time the response began. This anticipatory shift was much smaller in magnitude than what we found for *Normal Speech* with the same children (~25% vs. ~40%).

Perhaps surprisingly, when children only had access to prosodic or lexical information, they performed similarly to when they had no linguistic information at all with slight, if any, improvement with age (Figures 2 and 3). In the *Words Only* condition, looks to the answerer showed a small, but consistently greater magnitude for Question-Answer turn switches than for non-Question-Answer switches. Switch in gaze to the upcoming speaker was strongest in the *Normal Speech* condition, in which older children also clearly distinguished their gaze to Question-Answer and non-Question-Answer switches.

To test the reliability of the differences in anticipation between conditions and switch types, we fit a linear mixed-effects model (Gelman & Hill, 2007) to participants' aver-

age gaze at the upcoming speaker during the first 333 ms of the response. We included turn-switches and subjects as random effects, using maximal random effects structure (Barr et al., 2013) to control for variability between participants on the switch type (Question-Answer vs. non-Question-Answer) and linguistic condition. We also included a three-way interaction term for age, switch type, and condition with two-way interaction terms for age and gap duration and condition and gap duration.[6]

Model coefficients suggest that there were two significant effects in the gaze data. First, duration is a significant predictor of anticipation; longer gap durations result in more anticipation ($\beta = 0.73$, s.e. $= 0.28$, $t = 2.61$). Second, there is a highly significant three-way interaction between age, switch type, and condition ($\beta = 0.11$, s.e. $= 0.03$, $t = 3.16$) in predicting anticipation. This derives from the *Normal Speech* condition, in which children's differential looking behavior for Question-Answer vs. non-Question-Answer switches clearly diverges with age. No other coefficients reached significance.

Are children simply reacting to turn-ends and then looking to the other puppet, or are they instead anticipating the end of the prior speaker's turn and looking early on? To test this we fit a second model on turn-transitions that lasted less than 200 ms. Anticipatory looks in this subset of the data must have been planned before the prior turn ended. Model coefficients suggest anticipation still occurs with 2 two-way interactions between age and condition for *Words only* and *Normal* speech ($\beta = 0.07$, s.e. $= 0.03$, $t = 2.18$ and $\beta = 0.08$, s.e. $= 0.06$, $t = 2.31$, respectively).[7]

General Discussion

Children's looking patterns suggest that they reach at least two developmental benchmarks for predictive processing for discourse. First, children recognize that turn-taking requires immediate responses, and they quickly integrate linguistic and non-linguistic cues to shift their gaze in anticipation of a response. We saw this behavior from all children in our data set. As children grow older, they become more sensitive to linguistic cues, using them to distinguish between different conversational actions (questions vs. non-questions) and make earlier and swifter predictive shifts.

Since children in this age range still appear quite delayed in their own turn-taking, these results are strong evidence that children's apparent delays in everyday conversation are not due to the ability to predict when a turn-switch will occur. We propose that these delays are instead due to the cost of planning a response. Children's turn-timing during conversation is most delayed when they must make a complex response, and so a three-year-old's timing during conversation may appear to be slower than a one-year-old's (Casillas et al., in prep). But in our task, when the cognitive load was lightened so children were only required to perform comprehension, we saw that children's skill in predicting turn-structure develops early on and becomes more sensitive to discourse distinctions with age, using linguistic information to distinguish between different conversational acts (e.g., questions).

Children made their earliest and most consistent predictive looks in the condition where they had all linguistic information available to them. These results strengthen claims from previous work that young children spontaneously anticipate what is coming next in conversation (Casillas & Frank, 2012). By testing a broad age range, we found that children show greater anticipation, with a greater advantage for question- over non-question switches as they get older. Children in our study could effectively make predictions about normal speech by age 1;0, but that they begin distinguishing between different types of conversational actions (questions vs. non-questions) by the time they are 3;0 (Figure 3). Question effects are strongest when *both* prosodic and lexical cues are present, contrary to prior findings with adult listeners that found lexical information sufficient to predict upcoming turn-end boundaries (de Ruiter et al., 2006).

Children's performance was significantly downgraded by phonetically controlled stimuli such that their predictive eye movements were comparable to conditions in which they had no linguistic information at all. We suggest that children were able to make anticipatory shifts without linguistic information because they simply waited for one puppet to stop talking before looking to the next. Rather than anticipating the end of the ongoing turn, these children are likely anticipating the start of the next speaker's turn, which explains the significant effect of longer inter-turn gaps. In contrast, anticipation in the *Normal Speech* and *Words only* conditions still occurs when gaps are shorter than 200 ms, in which case children do not have time to simply react to the end of the prior turn and make significant shifts by the start of the response—in these cases they must have instead anticipated the end of the prior turn.

One limitation of the current study is that, by using puppets for the visual signal, we removed all visual cues to turn-taking except mouth movement. We did this to focus our analysis on linguistic cues, but visual cues are culturally variable and important indicators of conversational timing and coordination (Kendon, 1967; Stivers et al., 2009). In related work (Casillas & Frank, 2012), we asked 3-5 year-old children to watch short clips of conversation in languages they didn't speak. We saw larger and earlier-initiated anticipatory shifts in that experiment even though children in that study had no access to lexical information, only non-native prosodic and visual cues. Since children in the current study have smaller shifts, even in the *Normal Speech* condition, we suspect that visual cues play a large role in helping children guess what will come next, and that children integrate these cues with linguistic information when given the chance. Further work will be required to test this hypothesis. Also, children rarely hear phonetically-controlled speech, and may not have been able to process it as efficiently as normal speech, though they still were able to make small anticipatory shifts.

[6] Longer gaps give more time for gaze shift.

[7] There were also marginal effects of Age and Condition overall (t=-1.85). There were not enough non-question switches under 200 ms to test for effects of switch type in this model.

Conclusion

Just as children must learn to break into the linguistic stream and segment it into words, they must also learn to break into the interactional stream of conversation and segment it into turns. Using children's spontaneous gaze behavior while watching improvised conversations, this study has attempted to link online predictive processes with naturalistic, conversation-based stimuli. We have focused here on children's predictive skill in conversation because children's conversational skills can impact the form of their linguistic input and may be critical to understanding what children hear and how they practice language. Children's turn-taking skills help them become active interactants who have control over the linguistic input and practice they receive.

The implications of conversation-specific skills for language development are likely important (e.g., Weisleder, 2012), but are still largely unknown. Within single- and multi-utterance sequences, children's ability to predict what's coming next can aid in their uptake of new information (Fernald et al., 2008). By being able to predict what upcoming turn-structure will look like and anticipating the type of response needed for different types of actions (e.g., questions vs. non-questions), children are developing conversational skills that affect their input more globally: they can become more successful participants in multi-party interaction. Their skill in prediction within and across utterances then affects the type and quality of linguistic practice that children get during development.

Our findings indicate that rapid turn-timing is one of the earliest properties of organized interaction that children acquire, and that over the first seven years of life, children come to rely on their linguistic knowledge to refine and build on their predictions about what to expect next in conversation. So while children learn about language, they can use their linguistic knowledge online to take turns more effectively, and as children learn to take turns, they can use language more effectively in conversation with others.

Acknowledgments

We thank C. Kurumada, M. Lewis, and E. V. Clark for their helpful feedback, also the John Merck Scholars Program & the Hellman Faculty Fellows Program. This research is supported by an NSF dissertation grant to M. Casillas.

References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer software manual]. Available from `http://www.praat.org` ([Computer program] Version 5.3.16)

Casillas, M., Bobb, S. C., & Clark, E. V. (in prep). Turn-taking, timing, and access in early language acquisition.

Casillas, M., & Frank, M. C. (2012). Cues to turn boundary prediction in adults and preschoolers. In *Proceedings of SemDial 2012 (SeineDial): The 16th workshop on the semantics and pragmatics of dialogue.*

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., Zangl, R., Portillo, A., & Marchman, V. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. Fernandez, & H. Clahsen (Eds.), *Language acquisition and language disorders* (Vol. 44, pp. 97–135). Amsterdam/Philadelphia: John Benjamins.

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

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

Ito, K., & Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573.

Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society*, *2*, 181–199.

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

Masataka, N. (1993). Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three- to four-month-old Japanese infants. *Journal of Child Language*, *20*(02), 303–312.

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

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language*, *58*(2), 574–608.

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.

Thorgrímsson, G., Fawcett, C., & Liszkowski, U. (2011). *12-month-olds expect speech to provoke a response in others' interactions.* Poster presented at the Biennial Meeting of the Society for Research in Child Development, Montréal.

Tice, M., & Henetz, T. (2011). Turn-boundary projection: Looking ahead. In *Proceedings of the 33rd annual meeting of the Cognitive Science Society.*

Weisleder, A. (2012). *Richer language experience leads to faster understanding: Links between language input, processing efficiency, and vocabulary growth.* Unpublished doctoral dissertation, Stanford University.