# The development of children's ability to track and predict turn structure in conversation

Marisa Casillas[a,*], Michael C. Frank[b]

[a] *Max Planck Institute for Psycholinguistics, Nijmegen*
[b] *Department of Psychology, Stanford University*

**Abstract**

We investigate language acquisition through the lens of a fundamental conversational skill: turn taking. Children begin developing turn-taking skills in infancy, but take several years to assimilate their growing knowledge of language with this ability. In two eye-tracking experiments with children across a wide developmental range, we measured spontaneous predictions about upcoming speaker change while controlling the amount of linguistic information available. We found that children already predicted upcoming turns at age one, but they integrated linguistic cues differently at different ages. Children under three showed a general advantage for prosody over lexicosyntax, contrary to prior findings for adults. Children two and older showed more predictive switches for questions than non-questions, but only when lexical information was available. We found no evidence that lexicosyntax alone guides turn prediction—instead, participants' performance was best overall with access to lexicosyntax and prosody together. Our results point to the importance of studying children's linguistic processing in the context of conversation.

*Keywords:* Turn taking, Conversation, Development, Prosody, Lexical, Questions, Eye-tracking, Anticipation

## 1. Introduction

Spontaneous conversation is a universal context for using and learning language. Like other types of human interaction, it is organized at its core

*Corresponding author

by the roles and goals of its participants. But what sets conversation apart is its structure: sequences of interconnected, communicative actions that take place across alternating turns at talk. Sequential, turn-based structures in conversation are strikingly uniform across language communities and linguistic modalities. Turn-taking behaviors are also cross-culturally consistent in their basic features and the details of their implementation (De Vos et al., in preparation; Dingemanse et al., 2013; Stivers et al., 2009). How does this ability develop?

Children participate in sequential coordination with their caregivers starting at three months of age—before they can rely on any linguistic cues in taking turns (see, among others, Bateson, 1975; Hilbrink et al., in preparation; Jaffe et al., 2001; Snow, 1977). Of course, infant turn taking is different from adult turn taking in several ways. Infant turn taking is heavily scaffolded by caregivers, has distinct timing in comparison to adult turn taking, and lacks semantic content (Hilbrink et al., in preparation; Jaffe et al., 2001). However, children's early, turn-structured social interactions are presumably a critical precursor to their conversational turn taking. Along these lines, early non-verbal interactions establish the protocol by which children come to use language with others. And as they acquire language, they also come to integrate it into the preverbal turn-taking systems.

In this study, we investigate when children begin to make predictions about upcoming turn structure in conversation, and how they integrate language into their predictions as they grow older. In what follows, we first give a basic review of turn-taking research and the state of current knowledge about adult turn prediction. We then discuss recent work on the development of turn-taking skills before turning to the details of our own study.

### *1.1. Turn taking*

Turn taking itself is not unique to conversation. Many other human activities are organized around sequential turns at action. Traffic intersections and computer network communication both use turn-taking systems. Children's early games (e.g., give-and-take, peek-a-boo) have built-in, predictable turn structure (Ratner and Bruner, 1978; Ross and Lollis, 1987). Even monkeys take turns: Non-human primates such as marmosets and Campbell's monkeys vocalize contingently with each other in both natural and lab-controlled environments (Lemasson et al., 2011; Takahashi et al., 2013). In all these cases, turn taking serves as a protocol for interaction, allowing the participants to coordinate with one another through sequences of contingent action.

Conversation distinguishes itself from non-conversational turn-taking behaviors by the complexity of the turn sequencing involved. In the examples above (traffic, games, and monkeys) the set of sequence and action types is far more limited and predictable than what we find in everyday talk. For example, conversational turns come grouped into semantically-contingent sequences of action. The groups can span turn-by-turn exchanges (e.g., simple question–response, "How are you?"–"Fine.") or sequence-by-sequence exchanges (e.g., reciprocals, "How are you?"–"Fine, and you?"–"Great!"). Sequences of action drive the conversation forward into the next, relevant sequences of talk (e.g., "And you?"–"Great!"–"Why's that?"; Schegloff, 2007). To take a turn, participants need to make predictions about what conversational content will be relevant next. In some cases, relevant next turns are somewhat obvious (e.g., question–response) while, in other cases, there are multiple relevant next actions to choose from, or no obvious next action at all (e.g., after a closed sequence).

Despite this complexity, conversational turn taking is often precise in its timing. Across a diverse sample of conversations in 10 languages, one study found a consistent average transition time of 0–200 msec at points of speaker switch (Stivers et al., 2009). Experimental results and current models of speech production suggest that it takes approximately 600 msec to produce a content word, and even longer to produce a simple utterance (Griffin and Bock, 2000; Levelt, 1989). So in order to achieve 200 msec turn transitions, speakers must begin formulating their response before the prior turn has ended (Levinson, 2013). Moreover, to formulate their response early on, speakers must track and anticipate what types of response might become relevant next. They also need to predict the content and form of upcoming speech so that they can launch their articulation at exactly the right moment. Prediction thus plays a key role in timely turn taking.

### *1.2. Adults' turn prediction*

Adults have a lot of information at their disposal to help make accurate predictions about upcoming turn content. Lexical, syntactic, and prosodic information (e.g., *wh-* words, subject-auxiliary inversion, and list intonation) can all inform addressees about upcoming linguistic structure (De Ruiter et al., 2006; Duncan, 1972; Ford and Thompson, 1996; Torreira et al., under review). Non-verbal cues (e.g., gaze, posture, and pointing) often appear at turn-boundaries and can sometimes act as late indicators of an upcoming

speaker switch (Rossano et al., 2009; Stivers and Rossano, 2010). Additionally, the sequential context of a turn can make it clear what will come next: answers after questions; thanks or denial after compliments, et cetera (Schegloff, 2007).

Prior work suggests that adult listeners primarily use lexicosyntactic information to accurately predict upcoming turn structure (De Ruiter et al., 2006). De Ruiter and colleagues (2006) asked participants to listen to snippets of spontaneous conversation and to press a button whenever they anticipated that the current speaker was about to finish her turn. The speech snippets were controlled for the amount of linguistic information present; some were normal, but others had flattened pitch, low-pass filtered speech, or further manipulations. De Ruiter and colleagues found that, with pitch-flattened speech, the timing of participants' button responses was comparable to their timing with the full linguistic signal. But, when no lexical information was available, participants' responses were significantly earlier than the full linguistic signal. The authors concluded that lexicosyntactic information[1] was necessary and possibly sufficient for turn-end projection, while intonation was neither necessary nor sufficient. Congruent evidence comes from studies varying the predictability of lexicosyntactic and pragmatic content: Adults anticipate turn ends better when they can more accurately predict the exact words that will come next (Magyari and De Ruiter, 2012; see also Magyari et al., In press). They can also identify speech acts within the first word of an utterance (Gísladóttir et al., under review), allowing them to start planning their response at the first moment possible (Bögels et al., in preparation).

The role of prosody for adult turn-prediction is still a matter of debate. De Ruiter and colleagues' (2006) experiment only tested the role of intonation, which is only a partial index of prosody. Prosodic structure is also tied closely to the syntax of an utterance, and so the two linguistic signals are difficult to control independently (Ford and Thompson, 1996). Torreira, Bögels and Levinson (under review) used a combination of button-press and verbal responses to investigate the relationship between lexicosyntactic and prosodic cues in turn-end prediction. Critically, their stimuli were

[1]The "lexicosyntactic" condition only included flattened pitch and so was not exclusively lexicosyntactic—the speech would still have residual prosodic structure, including syllable duration and intensity.

cross-spliced so that each item had full prosodic cues to accompany the lexicosyntax. Because of the splicing, they were able to create items that had syntactically-complete units with no intonational phrase boundary at the end. Participants never verbally responded or pressed the "turn-end" button when hearing a syntactically-complete phrase without an intonational phrase boundary. And when intonational phrase boundaries were embedded in multi-utterance turns, participants were tricked into pressing the "turn-end" button 29% of the time. Their results suggest that listeners actually do rely on prosodic cues to execute a response (see also de De Ruiter et al. (2006):525). These experimental findings corroborate other corpus and experimental work promoting a combination of cues (lexicosyntactic, prosodic, and pragmatic) as key for accurate turn-end prediction (Duncan, 1972; Ford and Thompson, 1996; Hirvenkari et al., 2013).

In sum, adults accurately and spontaneously make predictions about upcoming turn structure. Their predictions rely on a sophisticated body of knowledge about linguistic structure, non-verbal signals, and social actions. Knowing this, we could expect that children's acquisition of turn-taking skills is closely tied to their knowledge about language, gaze, gesture, and social cues. But children's turn taking starts early in infancy, long before first words or gestures emerge. So a primary role for lexicosyntactic cues doesn't fit well with children's pre-verbal turn taking.

### *1.3. Children's turn prediction*

#### *1.3.1. Observational studies*

The majority of work on children's early turn taking has focused on observations of spontaneous interaction. Children's first turn-like structures appear as early as two to three months in proto-conversation with their caregivers (Bruner, 1975, 1985). During proto-conversations, caregivers interact with their infants as if they were capable of making meaningful contributions; they take every look, vocalization, arm flail and burp as "utterances" in the joint discourse (Bateson, 1975; Jaffe et al., 2001; Snow, 1977). Infants catch onto the structure of proto-conversations quickly. By three to four months they notice disturbances to the contingency of their caregivers' response and, in reaction, change the rate and quality of their vocalizations (Bloom, 1988; Masataka, 1993). Infants at this age also notice changes to social contingency outside of turn structure. In the Still Face paradigm, caregivers interact with their infants and then suddenly halt, taking on a neutral expression with a

sustained gaze. When faced with this sudden disappearance of social contingency, infants three months and older try a range of methods to reinitiate the interaction, such as vocalization, reaching, and smiling before looking away or getting upset (Rochat et al., 1998; Toda and Fogel, 1993).

The timing of children's responses to their caregivers' speech shows a non-linear pattern of fall-rise-fall from early infancy to middle childhood. A recent study by Hilbrink et al. (in preparation) finds that infants' turn timing at three months is often too early or too late: they start vocalizing in overlap on nearly 40% of their caregivers' turns, and their non-overlapped vocalizations come after an average inter-turn silent gap of 675 msec (adult average: 200 msec). Between four and nine months, children begin to reduce the number of turns happening in overlap while also improving on their average response latency. But then, later on, children's response latencies slow down again, peaking at average gaps of 1100 msec at nine months, with only very gradual improvement after that (Hilbrink et al., in preparation). While children's avoidance of overlap is nearly adult-like by nine months, the timing of their non-overlapped responses stays much longer than the 200 msec standard for the next several years (Garvey, 1984; Ervin-Tripp, 1979).

The protracted development of children's timing may be attributable to their linguistic development: Taking turns on time is easier when the response is a simple vocalization rather than a linguistic utterance. Integrating language into the turn-taking system may be one major factor in children's delayed responses (Casillas et al., under review). If response planning (i.e., language production) is the primary hurdle in children's spontaneous turn taking, we should find evidence that children understand turn-taking behaviors before they are able to produce the behaviors themselves; this hypothesis has been recently explored in experimental settings.

*1.3.2. Experimental studies*

Children begin to develop specific expectations about conversational behavior before they begin to speak. Sometime between four and six months, children begin to attend differently to face-to-face and back-to-back conversation; six-month-olds follow conversational speakers more with their gaze when at least one speaker is looking at the other (Augusti et al., 2010). At ten months, infants expect people to look and talk at other people, and not to objects (Beier and Spelke, 2012). At twelve months infants expect to see responses to verbal (but not non-speech) utterances in face-to-face contexts (Thorgrímsson et al., under review).

There are mixed results regarding when children begin to anticipate turn structure in conversation. One study found that 12-month-olds make more predictive gaze shifts to a responder while watching human verbal conversation compared to conversation-like interactions with objects (Bakker et al., 2011), but another only found a similar effect at 36 months (von Hofsten et al., 2009). Neither of these two studies had baselines to which the turn-relevant looking behavior could be compared, however. A baseline measurement is critical because there may be developmental differences in gaze shifting between conversational participants, even if the shifting is not related to turn structure. Such developmental differences could produce artifactual changes in measures of turn-contingent shifting.

Keitel and colleagues (2013) addressed the random baseline issue in their study of 6-, 12-, 24-, and 36-month-olds. They asked participants to watch short videos of conversation, and tracked their eye movements at points of speaker change. They found that children's anticipatory gaze frequency was only greater than chance for 36-month-olds and adults. Their study was the first to focus on the role of linguistic processing in children's turn predictions. They showed their participants two types of conversation videos: one normal and one with flattened pitch (i.e., with flattened intonation contours), finding that only 36-month-olds were affected by a lack of intonation contours. The adult control group made equal numbers of anticipatory looks in the videos, with and without intonation contours, consistent with prior adult findings (De Ruiter et al., 2006). Keitel and colleagues concluded that children's ability to predict upcoming turn structure relies on their ability to comprehend the stimuli (emerging around 36 months), especially with respect to semantic access. They also suggest that intonation takes a secondary role in turn prediction, but only *after* children acquire more sophisticated, adult-like language comprehension systems (sometime after 36 months).

Although the Keitel et al. (2013) study constitutes a substantial advance over previous work, it nevertheless has several limitations. Because these limitations directly inform our own study design, we review them in some detail. First, their estimates of baseline gaze frequency ("random" in their terminology) were computed from switches that occurred specifically while the speaker's turn was ongoing; in contrast, estimates of anticipatory shifting frequency were computed from switches that occurred within the time window immediately surrounding inter-turn gaps. Assuming that gaze switches happen most often during the inter-turn gap (Hirvenkari et al., 2013), the switches made during ongoing speech (their baseline) would be minimized.

This baseline thus maximized the possibility of finding a difference between "random" and anticipatory gaze frequencies. A more conservative baseline would be to compare participants' looking behavior at turn transitions to their looking behavior during randomly-selected windows of time throughout the stimulus. We follow this conservative approach in our work.

Second, the conversation stimuli they used were somewhat unusual. The average gap between turns was 900 msec, which is much longer than typical adult timing, where gaps average around 200 msec (Stivers et al., 2009). The speakers in the videos were also asked to minimize their movements while performing a scripted and adult-directed conversation, which would have created a somewhat unnatural stimulus. In addition, in order to produce more naturalistic conversation, it would have been ideal to localize sources for the two voices in the video (i.e., to have the voices come out of separate left and right speakers). But both voices were recorded and played back on the same audio channel, which may have made it more difficult to distinguish the two talkers. Again, we attempt to address these issues in our current study.

Despite these minor methodological issues, the Keitel et al. (2013) study still demonstrates intriguing age-based differences in children's ability to predict upcoming turn structure, and the results suggest that both semantic and intonational development *do* play a role in children's looking patterns. Our current work thus takes this paradigm as our starting point.[2]

*1.3.3. Prosodic development*

The roles of prosody and lexicosyntax in children's turn predictions are currently unknown, but children understand more about prosody early in life. Children begin to acquire prosody in the womb, and can distinguish their native language's rhythm type from others (e.g., syllable-timed vs. stress-timed) 2–5 days after birth (Mehler et al., 1988; Moon et al., 1993; Nazzi and Ramus, 2003). Beginning between four and five months, infants prefer pauses in speech to be inserted at prosodic boundaries, and by 6 months they can start using prosodic markers to pick out sub-clausal syntactic units (Jusczyk et al., 1995; Soderstrom et al., 2003). They show preference for the typical stress patterns of their native language over others by 6–9 months (e.g., iambic vs. trochaic), and can use prosodic information to segment the speech stream into smaller chunks from 8 months onward (Johnson and

[2]See also Casillas and Frank (2012, 2013).

Jusczyk, 2001; Jusczyk et al., 1993; Morgan and Saffran, 1995). In comparison, children show only a limited lexical inventory at six months, and begin to recognize function words just before their first birthdays, with syntactic categorization beginning around 14 months (Bergelson and Swingley, 2013; Shi and Melancon, 2010). Two-month-olds notice changes in word order, but this ability appears to rely on prosodic cueing (Mandel et al., 1996). Generally speaking then, our current knowledge about children's linguistic development points to a possible early advantage for prosody in children's turn-taking predictions.

### *1.4. The Current Study*

We report here on the role of linguistic processing in children's predictions about upcoming turn structure. We focus in particular on how children use prosodic and lexicosyntactic information to make their predictions. Prior work has focused mainly on lexicosyntax and intonation, and not on prosody proper (De Ruiter et al., 2006; Keitel et al., 2013, but see Torreira et al., under review), even though infants seem to acquire the basic rhythmic properties of the prosodic signal first (Mehler et al., 1988; Moon et al., 1993; Nazzi and Ramus, 2003).

In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling for the amount of lexicosyntactic and prosodic information available. Experiment 1 uses a paradigm in which English-speaking participants view video clips of naturalistic conversations from several different languages in order to create conditions under which lexicosyntactic information was differentially available, but prosodic and non-linguistic cues were intact. This paradigm showed minimal differences between the predictive looking behavior of preschoolers and adults. In Experiment 2, we created artificial (puppet) visual stimuli to better separate these linguistic cues from one another. In this more controlled paradigm, we found that children's predictive looking behavior improved from ages one to six, but that even one-year-olds made more anticipatory looks than would be expected by chance.

In both experiments children consistently looked faster to responders after hearing questions, compared to non-questions. Both prosodic and lexicosyntactic information played a role in children's predictions about turn structure, but the two information sources were used differently at different ages. Our findings overall support an account in which predictive processes

for turn taking in conversation are present early, but their integration with linguistic information takes substantial practice.

## 2. Experiment 1

We recorded participants' eye movements as they watched six short videos of two-person (dyadic) conversation. Each video featured an improvised discourse in one of five languages (English, German, Hebrew, Japanese, and Korean). The participants, all native English speakers, were only expected to understand the two videos in English. We showed participants non-English videos to limit their access to lexical information while maintaining their access to other cues to turn boundaries (e.g., (non-native) prosody, gaze, breath, phrase final lengthening). Using this method, we compared children and adult's anticipatory looks from the current speaker to the upcoming speaker at points of turn transition in English and non-English videos.

### *2.1. Methods*

#### *2.1.1. Participants*

We recruited 74 children between ages 3;0–5;11 and 11 undergraduate adults to participate in the experiment. Our child sample included 19 three-year-olds, 32 four-year-olds, and 23 five-year-olds, all enrolled in a local nursery school. All participants were native English speakers. Approximately one-third (N=25) of the children's parents and teachers reported that their child regularly heard a second (and sometimes third or further) language, but only one child frequently heard a language that was used in our non-English video stimuli, and we excluded his data from analyses. None of the adult participants reported fluency in a second language.

#### *2.1.2. Materials*

*Video recordings.* We recorded pairs of talkers while they conversed in a sound-attenuated booth (see sample frame in Figure 1). Each talker was a native speaker of the language being recorded, and each talker pair was male-female. Using a Marantz PMD 660 solid state field recorder, we captured audio from two lapel microphones, one attached to each participant, while simultaneously recording video from the built-in camera of a MacBook laptop computer. The talkers were volunteers, all members of the local post-graduate community, and were acquainted with their recording partner ahead of time.

Figure 1: Example frame from a conversation video used in Experiment 1.

Each recording session began with a 20-minute warm-up period of spontaneous conversation during which the pair talked for five minutes on four topics (favorite foods, entertainment, hometown layout, and pets). Then we asked talkers to choose a new topic—one relevant to young children (e.g., riding a bike, eating breakfast)—and to improvise a dialogue on that topic. We asked them to speak as if they were on a children's television show in order to elicit child-directed speech toward each other. We recorded until the talkers achieved at least 30 seconds of uninterrupted discourse with enthusiastic, child-directed speech. Most talker pairs took less than five minutes to complete the task, usually by agreeing on a rough script at the start. We encouraged talkers to ask at least a few questions to each other during the improvisation. The resulting conversations were therefore not entirely spontaneous, but were as close as possible while still remaining child-oriented in topic, prosodic pattern, and lexicosyntactic construction.[3]

After recording, we combined the audio and video files by hand, and cropped each recording to the 30-second interval with the most turn activity. Because we recorded the conversations in stereo, the male and female voices came out of separate speakers during video playback. This gave each voice in the videos a localized source (from the left or right loudspeaker). We coded each turn transition in the videos for language group (English vs. non-

[3]All of the non-English talkers were fluent in English as a second language, and some fluently spoke a third or more language. We chose male-female pairs as a natural way of creating contrast between the two talker voices. See an example run-though of the videos here: www.youtube.com/channel/UCGEZQcM9t8Zfjqqi_B1Q5Sw.

English), inter-turn gap duration (in milliseconds), and speech act (question vs. non-question). The non-English stimuli were coded for speech act from a monolingual English-speaker's perspective, i.e., which turns "sound like" questions, and which don't.

Because the conversational stimuli were recorded semi-spontaneously, the duration of turn transitions and the number of speaker transitions in each video was variable. We measured the duration of each turn transition from the audio recording associated with each video. We excluded turn transitions longer than 550 msec and shorter than 0 msec (those with transitional overlap) from analysis.[4] This left approximately equal numbers of turn transitions available for analysis in the English (N=22) and non-English (N=21) videos. On average, the inter-turn gaps for English videos (mean=273 msec, median=291 msec) were slightly longer than for non-English videos (mean=226 msec, median=205 msec). The longer gaps in the English videos could give them a slight advantage: Our definition of an "anticipatory gaze shift" includes shifts that are initiated during the gap between turns (Figure 2), so participants had slightly more time to make anticipatory shifts in the English videos.

Questions made up approximately half of the turn transitions in the English (54%; N=12) and non-English (48%; N=10) videos. In the English videos, inter-turn gaps were slightly shorter for questions (mean=271 msec, median=254 msec) than non-questions (mean=276 msec, median=300 msec). The opposite was true for the non-English videos: question transitions (mean=264 msec, median=257 msec) had longer inter-turn gaps than non-question transitions (mean=191 msec, median=168 msec). Thus non-question (i.e., declarative) transitions in non-English videos were the briefest type of turn transition in the stimuli on average.

*2.1.3. Procedure*

Participants sat in front of an SMI 120Hz corneal reflection eye-tracker mounted beneath a large flatscreen display. The display and eye-tracker were

[4]Overlap occurs when a responder begins a new turn before the current turn is finished. When overlap occurs, observers cannot switch their gaze in anticipation of the response because the response begins earlier than expected; participants expect conversations to proceed with "one speaker at a time" (Sacks et al., 1974). As such, they would still be fixated on the prior speaker when the overlap started, and then would have to switch their gaze *reactively* to the responder.

secured to a table with an ergonomic arm that allowed the experimenter to position the whole apparatus at a comfortable height, approximately 60 cm from the viewer. We placed stereo speakers on the table, to the left and right of the display.

Before the experiment started, we warned adult participants that they would see videos in several languages and that, though they weren't expected to understand the content of non-English videos, we *would* ask them to answer general, non-language-based questions about the conversations. Then after each video we asked participants one of the following randomly-assigned questions: "which speaker talked more?", "which speaker asked the most questions?", "which speaker seemed more friendly?", and "did the speakers' level of enthusiasm shift during the conversation?" We also asked if the participants could understand any of what was said after each video. The participants responded verbally while an experimenter noted their responses.

Children were less inclined to simply sit and watch videos of conversation in languages they didn't speak, so we used a different procedure to keep them engaged. The experimenter started each session by asking the child about what languages he or she could speak, and about what other languages he or she had heard of. Then the experimenter expressed her own enthusiasm for learning about new languages, and invited the child to watch a video about "new and different languages" together. If the child agreed to watch, the experimenter and the child sat together in front of the display, with the child centered in front of the tracker and the experimenter off to the side. If the child began to look bored during video playback, the experimenter would talk during the filler videos, either commenting on the previous conversation ("That was a neat language!") or giving the language name for the next conversation ("This next one is called Hebrew. Let's see what it's like."). The experimenter's comments reinforced the video-watching as a joint task.

All participants (child and adult) completed a five-point calibration routine before the first video started. We used a dancing Elmo for the children's calibration image. During the experiment, participants watched all six 30-second conversation videos with 15–30 second filler videos in-between them (e.g., running puppies, singing muppets, flying bugs). The first and last conversations were in American English and the intervening conversations were Hebrew, Japanese, German, and Korean. The presentation order of the non-English videos was shuffled into four lists, which participants were assigned to randomly. The entire experiment, including instructions, took 10–15 minutes.

*2.1.4. Data preparation and coding*

To determine whether participants predicted upcoming turn transitions, we needed to define a set of criteria for what counted as an anticipatory gaze shift. Prior work using similar experimental procedures has found that adults and children make anticipatory gaze shifts to upcoming talkers within a wide time frame; the earliest shifts occur before the end of the prior turn, and the latest occur after the onset of the response turn, with most shifts occurring in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, 2011). Following prior work, we measured how often our participants shifted their gaze from the prior to the upcoming speaker *before* the shift in gaze could have been initiated in reaction to the onset of the speaker's response. In doing so, we assumed that it takes children 333 msec to plan an eye movement, following Fernald and colleagues' (2008) similar 'looking while listening' method. We assumed that it takes adults 200 msec to plan an eye movement, following standards from adult anticipatory processing studies (e.g., Kamide et al., 2003).

We checked each participant's gaze at each turn transition for three characteristics (Figure 2): (1) that the participant fixated on the prior speaker for at least 100 msec at the end of the prior turn, (2) that sometime thereafter the participant switched to fixate on the upcoming speaker for at least 100 ms, and (3) that the switch in gaze was initiated within the first 333 msec of the response turn, or earlier. These criteria guarantee that we only counted gaze shifts when: (1) participants were tracking the previous speaker, (2) switched their gaze to track the upcoming speaker, and (3) did so before they could have simply reacted to the onset of speech in the response. Under this assumption, a gaze shift that was initiated within the first 333 msec of the response (or earlier) was planned *before* the child could react to the onset of speech itself.

As mentioned, most anticipatory switches happen in the inter-turn gap, but we also allowed anticipatory gaze switches that occurred in the final syllables of the prior turn. Early switches are consistent with the distribution of responses in explicit turn-boundary prediction tasks. For example, in a button press task, adult participants anticipate turn ends approximately 200 msec in advance of the turn's end, and anticipatory responses to pitch-flattened stimuli come even earlier (De Ruiter et al., 2006). We therefore allowed switches to occur as early as 333 msec before the end of the prior turn for children, and 200 msec before the end of the prior turn for adults.

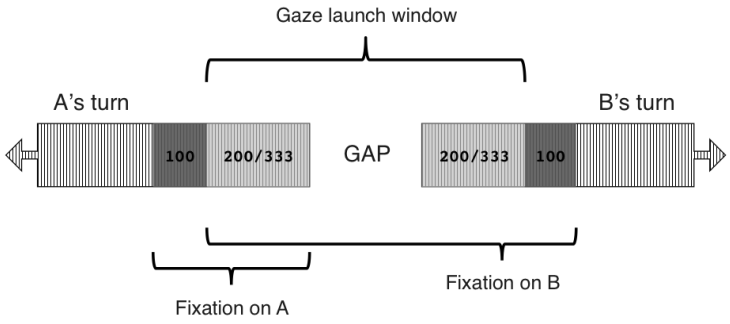


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition. The criteria were the same for adults and children, except children were assumed to have a 333 msec window for planning eye movements, while adults were only given 200 msec.

For very early and very late switches, our requirement for 100 msec of fixation on each speaker would sometimes extend outside of the transition window boundaries (333 msec before and after the inter-turn gap). The maximally available fixation window was 100 msec before and after the earliest and latest possible switch point (433 msec before and after the inter-turn gap). We did not count switches made during the fixation window as anticipatory. We *did* count switches made during the inter-turn gap. The period of time from the beginning of the possible fixation window on the prior speaker to the end of the possible fixation window on the responder was our total analysis window (433 msec + the inter-turn gap + 433 msec), which we use to create the random baselines described below. For adult participants we used 200 msec (instead of 333 msec) as the assumed time to plan an eye movement in defining all analysis windows.

*Random baseline analysis.* To ensure that participants' looking behavior was turn-related and not simply the result of random switching, we corrected for the probability of making an anticipatory switch at random. We estimated the baseline chance of making an anticipatory switch at each turn transition for each participant. We ran each participants' eye-tracking signal through our switch identification procedure 2.1.4 with 100 randomly-shuffled versions of the analysis windows (Figure 3). For each of the 100 randomly-shuffled versions, we recorded whether the participant made a "switch" or not. We then averaged participants' switches from all 100 randomly-shuffled versions of the analysis windows, yielding a value between 0 and 1 for each turn tran-

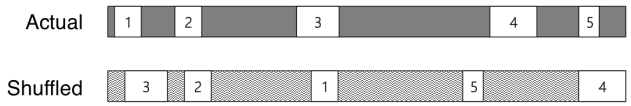


Figure 3: Example of shuffling for five turn transition analysis windows. The windows were ±433 msec around the inter-turn gap for children, and from ±300 msec for adults.

sition for each participant. We then subtracted the random baseline value from each participant's actual switch result (1 or 0) for each turn transition to arrive at the final, baseline-corrected values. Subtracting the random baseline this way gives us an estimate of how often participants switched above chance across ages, conditions, and transition types. If participants made the same number of anticipatory switches, regardless of turn transition location, there would be no evidence that their switching behavior was linked to the conversation stimulus.

### *2.2. Results and discussion*

#### *2.2.1. Descriptive analysis*

Participants looked at the screen most of the time during video playback (81% and 91% on average for children and adults, respectively). They primarily kept their eyes on the person who was currently speaking in both English and non-English videos: They gazed at the current speaker between 38% and 63% of the time, looking back at the addressee between 15% and 20% of the time (Table 1). Even three-year-olds looked more at the current speaker than anything else, whether the videos were in a language they could understand or not. Children looked at the current speaker less than adults did during the non-English videos. Despite this, their looks to the addressee did not increase substantially in the non-English videos, indicating that their looks away were probably related to boredom rather than confusion about ongoing turn structure. Overall, participants' pattern of gaze to current speakers indicated that they performed basic turn tracking during the videos, regardless of language.

We identified anticipatory gaze switches for all 43 usable turn transitions, based on the criteria outlined in Section 2.1.4. Because participants had greater access to linguistic information in English, we expected to see greater anticipation in the English videos than in the non-English videos. We also predicted that anticipation would be greater following questions compared to

| Age group | Condition | Speaker | Addressee | Other onscreen | Offscreen |
|---|---|---|---|---|---|
| 3 | English | 0.61 | 0.16 | 0.14 | 0.08 |
| 4 | English | 0.60 | 0.15 | 0.11 | 0.13 |
| 5 | English | 0.57 | 0.15 | 0.16 | 0.12 |
| Adult | English | 0.63 | 0.16 | 0.16 | 0.05 |
| 3 | Non-English | 0.38 | 0.17 | 0.20 | 0.25 |
| 4 | Non-English | 0.43 | 0.19 | 0.21 | 0.18 |
| 5 | Non-English | 0.40 | 0.16 | 0.26 | 0.18 |
| Adult | Non-English | 0.58 | 0.20 | 0.16 | 0.07 |

Table 1: Average proportion of gaze to the current speaker and addressee during periods of talk.

non-questions; questions have early cues to upcoming turn transition (e.g., *wh-* words, subject-auxiliary inversion), and also make a next response immediately relevant. We also predicted that anticipatory looks would increase with development, along with children's increased linguistic competence.

To estimate the effects of linguistic condition and age on switching behavior, we estimated each participants' random baseline switching behavior for each turn transition. Participants at all ages made anticipatory gaze switches more often than would be expected by chance, whether they could understand the language in the video or not (Figure 4). Anticipatory switch behavior in the non-English videos could not have been based on lexicosyntactic information, so both adults and children must have relied on other sources of information to make their predictions (e.g., non-native prosodic cues and non-verbal signals). For all further analyses we then corrected each participant's data by subtracting their random baseline estimations from their actual switching data for each turn transition.

Overall, a greater proportion of participants made anticipatory switches in the English videos (0.17) compared to the non-English videos (0.09). But the tendency to make more anticipatory switches for English than non-English videos disappeared with age (Figure 5). While anticipatory switches increased with age for non-English videos, they decreased with age for English videos, resulting in nearly equal looking behavior for the two language conditions for adults. Otherwise, age-based differences in predictive gaze shifts were minor.

Participants also showed an advantage for questions (Figure 6). Children

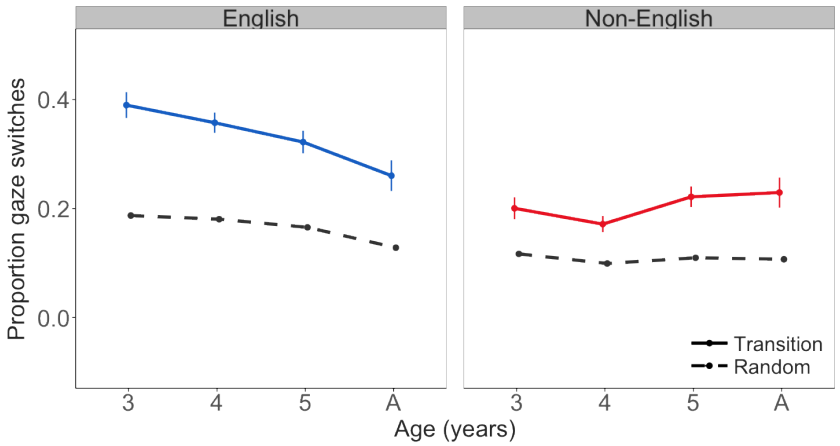


Figure 4: Proportion of anticipatory gaze switches made for actual (solid) and randomly-shuffled (dashed) turn transitions in each condition, and across ages. Error bars represent the standard error of the mean.

made anticipatory switches twice as often when they heard a question, compared to a non-question (18.7% and 7.3%). But the advantage for questions was not universal—it depended on language condition and age. The advantage for questions didn't emerge until age five in the non-English videos.

*2.2.2. Statistical analysis*

We quantified these findings using two linear mixed-effects models of participants' baseline-corrected anticipatory switches (Bates et al., 2014; R Core Team, 2014). We built one model each for children and adults. The child model included condition (English vs. non-English)[5], transition type (question vs. non-question), age, and duration of the inter-turn gap as predictors, with full interactions between condition, transition type, and age. We included the duration of the inter-turn gap as a predictor since longer gaps also provide more opportunities to make anticipatory switches (Figure 2). We ad-

[5]Because each non-English language was represented by a single stimulus, we cannot compute reliable cross-language differences. Gaze behavior might be best for languages that have the most structural overlap with participants' native language: 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). We would need multiple items from each of the language to check for similarity effects of specific linguistic features.

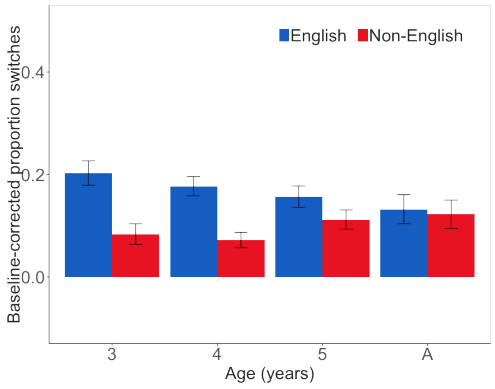


Figure 5: Baseline-corrected proportion of anticipatory gaze switches made for turn transitions in each language condition, and across ages. Error bars represent the standard error of the mean.

ditionally included random effects of item (turn transition) and participant, with random slopes of condition, transition type, and their interaction for participants (Barr et al., 2013). The adult model was exactly the same as the children's, excluding age-related effects.

Children's anticipatory gaze switches showed a main effect of language condition ($\beta$=-0.448, $SE$=0.126, $t$=-3.56, $p$<.001).[6] As predicted, children made more anticipatory looks while they were watching videos that they could understand. In the English videos, children had access to lexicosyntactic information, which could have helped them make better predictions about upcoming turn structure. Although children performed better on the English videos, lexicosyntactic information was not exclusively responsible for their anticipatory switches; children still performed remarkably well when no lexical information was present at all, and even three-year-olds made anticipatory switches at an above-chance rate in the non-English videos.

Contrary to our expectation, there was no main effect of age ($\beta$=-0.03, $SE$=0.027, $t$=-1.145, $p$=.252). Children's anticipatory switches *did* increase with age for the non-English videos, but it also decreased with age for the English videos, resulting in a significant interaction of age and language

[6]We derive $p$-values here by treating $t$ as $z$ (with a two-tailed assumption), as is justified for large data sets (Barr et al., 2013).

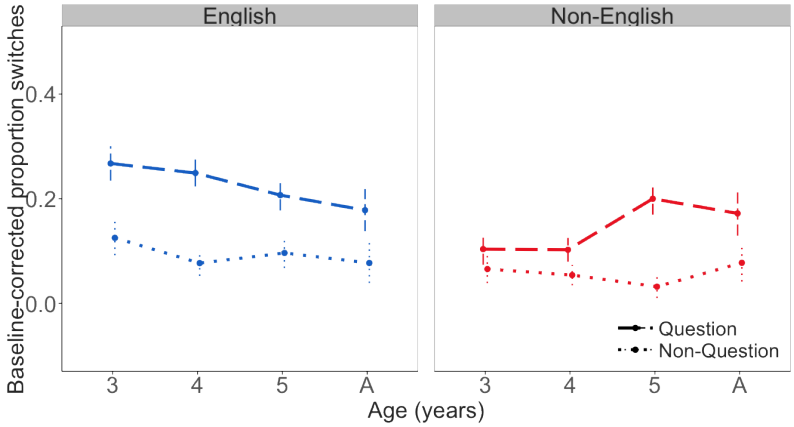


Figure 6: Baseline-corrected proportion of anticipatory gaze switches made for question- (dashed) and non-question- (dotted) turn transitions in each condition, and across ages. Error bars represent the standard error of the mean.

condition ($\beta$=0.083, $SE$=0.027, $t$=3.063, $p$<.001). We had expected that, as children got older, they would also show more anticipatory switches. But this is contradicted by decreased looks with age for English videos, which even extended to the adult data. One possible explanation is that older children and adults found the videos in English easy to comprehend, and so tracked the conversation less closely. Compared to three-year-olds, adults and older children spent more time looking away from the current speaker and addressee in the English conversations. In contrast, compared to three-year-olds, they spent more time looking at the current speaker and less time looking away in the non-English conversations (Table 1).

It is also possible that the timing of anticipatory gaze switches changes as children get older. That is, the moment at which observers chose to launch their gaze away from the current speaker might change with age. For example, children could have reacted more immediately to early cues in the turn (e.g., "Which one...") as they got older. If a change in timing were sufficiently large, we might not capture it in our analysis window, which was designed to capture switches close to the turn transition. To test for a change in switch timing, we would have to run new analyses on a larger set of turn transitions—ones with controlled early and late cues (see, e.g., Bögels et al., in preparation). Within the current analysis windows there

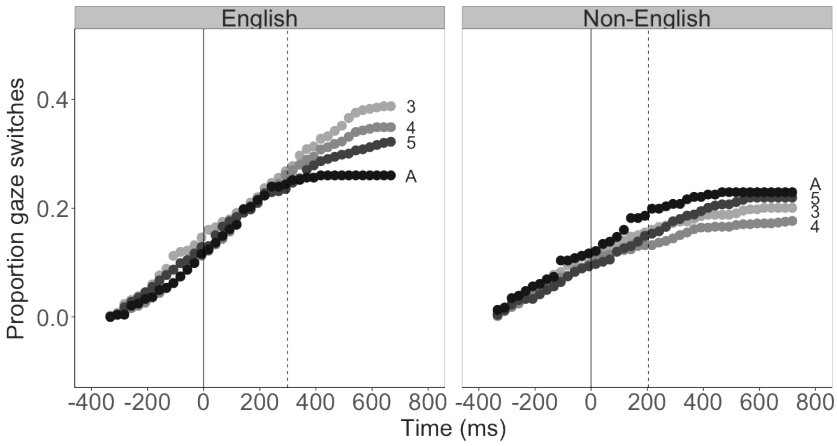


Figure 7: Cumulative proportion of anticipatory gaze switches planned during the transition window, across conditions and ages. Age is indicated by hue (darker = older). The solid line is the offset of the pre-gap turn, and the dashed line is the median onset of the post-gap turn in each condition. "Early switches" occur before 0.

were some "early switches" (i.e., those launched before the end of the prior turn). But participants made approximately equal numbers of early switches, regardless of age and condition: 11–15% (Figure 7). The lack of increased early switching with age might suggest that changes in timing are either larger than the analysis window, or are not at play.

Children made more anticipatory switches after hearing questions than non-questions, but the advantage for questions was driven by age and linguistic access. There was no main effect of transition type alone ($\beta$=-0.209, $SE$=0.151, $t$=-1.384, $p$=.166). Rather, the model confirmed a two-way interaction of linguistic condition and transition type ($\beta$=0.445, $SE$=0.19, $t$=2.344, $p$<.02). Presumably, question effects were larger in the English videos than the non-English videos because there were more cues to questionhood in English (i.e., non-verbal, prosodic, *and* lexicosyntactic). Children younger than five did not distinguish questions from non-questions in the non-English videos (Figure 6). The children's model confirmed a three-way interaction of age, linguistic condition, and transition type ($\beta$=-0.091, $SE$=0.042, $t$=-2.183, $p$<.05). Lexicosyntactic information might then be critical for young children to identify questions in conversation. Cues like subject-auxiliary inversion, and *wh*-words occur early in the turn and are clear indicators that an answer will come next. Importantly, the turns

marked as "questions" in the non-English data were not real questions, but turns that *sounded like* questions to native English speakers. Question-like prosodic cues in the non-English videos could have been markedly different from question prosody in English, making it more difficult for three- and four-year-olds to identify "questions" in those conversations.

The increased switching after questions indicates that, so long as they have native lexical and prosodic cues, children as young as three spontaneously distinguish questions from other types of speech acts in third-party conversation. Children may treat questions differently from non-questions in conversation because they are predictably followed by an immediate response (i.e., the answer). Because the form of an answer is partially determined by its question (e.g., locative expressions for "where" questions), children could even be predicting what type of response they will hear, and look at the responder for confirmation. In contrast, non-questions do not usually require an immediate response, so it is not as easy to predict what will come next.

Inter-turn gap duration did not affect children's anticipatory switches ($\beta$=0.202, $SE$=0.151, $t$=1.340, $p$=.180), but it was the main factor driving adults' switches ($\beta$=0.648, $SE$=0.047, $t$=13.855, $p$<.001). The adult model also showed a marginal effect of transition type, with adults looking more after questions compared to non-questions ($\beta$=-0.117, $SE$=0.069, $t$=-1.704, $p$=.088). Language condition and its interaction with transition type were not significant. Language did not affect adults' anticipatory looking behavior, possibly because the animated, child-directed prosody and non-verbal cues made it possible for them to follow the interaction and track the current speaker easily.

*2.2.3. Summary*

Several results from Experiment 1 were unexpected. First, we found no significant main effect of age in the children's data, and we saw very little difference between adults and children (besides the age-related interactions). It appears that, especially in English, children were as good at tracking the current speaker and anticipating upcoming turns as adults were. To find out when children first begin making spontaneous anticipatory looks to responders, it is necessary to recruit younger participants. Experiment 2 thus includes a sample of one- and two-year-olds.

Second, children, but not adults, showed an advantage for lexicosyntactic information in their anticipatory switches. An advantage for lexicosyntactic information is consistent with prior work on adult turn taking (De Ruiter

et al., 2006), so it's curious that the adults did not show a difference in anticipatory switching between videos they did and did not understand. As mentioned above, it is likely that they used non-verbal cues to track the interactions and consequently found the videos so easy to follow that they did not track the turn taking closely. To identify which linguistic cues adults and children rely on, it is necessary to control the presence of non-verbal cues in the stimulus—Experiment 2 implements this control.

Finally, we had predicted that children would show early advantages for prosodic cues over syntactic ones because children begin developing language-specific knowledge about prosody long before lexicosyntax. We weren't able to directly test this prediction in Experiment 1 because we used a limited range of ages (three- to five-year-olds) and because we only controlled for lexicosyntactic information. Children had prosodic cues in both English and non-English videos, although the prosodic signal was non-native in the latter case. So, although using non-English videos was a natural way to control for lexical information, it did not allow us to assess the role of prosody. Relatedly, the advantage for questions was greater in the English videos than the non-English videos, suggesting that lexicosyntactic cues are important for children in identifying questions during conversation. Again, to test this idea, we would need to directly compare lexicosyntactic and prosodic cues in the participants' native language. In sum, Experiment 1 lays the analytic groundwork for a method that allows for greater experimental control, which we introduce in Experiment 2.

## 3. Experiment 2

We improved our design by using native-language stimuli, controlling for lexical *and* prosodic information, eliminating non-verbal cues, and testing children from a wider age range. All of the videos in Experiment 2 were in the participants' native language (American English). To tease apart the role of lexical and prosodic information, we phonetically manipulated the speech signal for pitch, syllable duration, and lexical access. By testing one- to six-year-olds we hoped to find the developmental onset of turn-predictive gaze. We also hoped to measure changes in the relative roles of prosody and lexicosyntax across development.

Non-verbal cues in Experiment 1 (e.g., gaze and gesture) could have helped participants make predictions about upcoming turn structure (Rossano et al., 2009; Stivers and Rossano, 2010). Since our focus is on linguistic cues,

we eliminated all gaze and gestural signals in Experiment 2 by replacing the videos of human actors with videos of puppets. Puppets are less realistic and expressive than human actors, but they create a natural context for having somewhat motionless talkers in the videos (thereby allowing us to eliminate gestural and gaze cues). Additionally, the prosody-controlled condition included small but global changes to syllable duration that would have required complex video manipulation or precise re-enactment with human talkers, neither of which was feasible. For these reasons, we decided to substitute puppet videos for human videos in the final stimuli.

As in the first experiment, we recorded participants' eye movements as they watched six short videos of dyadic conversation, and then analyzed their anticipatory glances from the current speaker to the upcoming speaker at points of turn transition.

### *3.1. Methods*

#### *3.1.1. Participants*

We recruited 27 undergraduate adults and 129 children between ages 1;0–6;11 to participate in our experiment. We recruited our child participants from the Children's Discovery Museum in San Jose, California, targeting approximately 20 children for each of the six 1-year age groups (range=20–23). All participants were native English speakers, though some parents (N=27) reported that their child heard a second (and sometimes third) language at home. None of the adult participants reported fluency in a second language. We ran Experiment 2 at a local children's museum because it gave us access to children with a more diverse range of ages.

#### *3.1.2. Materials*

We created 18 short videos of improvised, child-friendly conversation (Figure 8). To eliminate non-verbal cues to turn transition and to control the types of linguistic information available in the stimuli we (1) recorded improvised conversations, (2) phonetically manipulated those recordings to limit the availability of prosodic and lexical information, and (3) recorded a new set of videos that featured puppets as talkers, using the manipulated audio as the puppets' speech.

*Audio recordings.* The recording session was set up in the same way as the first experiment, but with a shorter warm up period (5–10 minutes) and a pre-determined topic for the child-friendly improvisation ('riding bikes', 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the

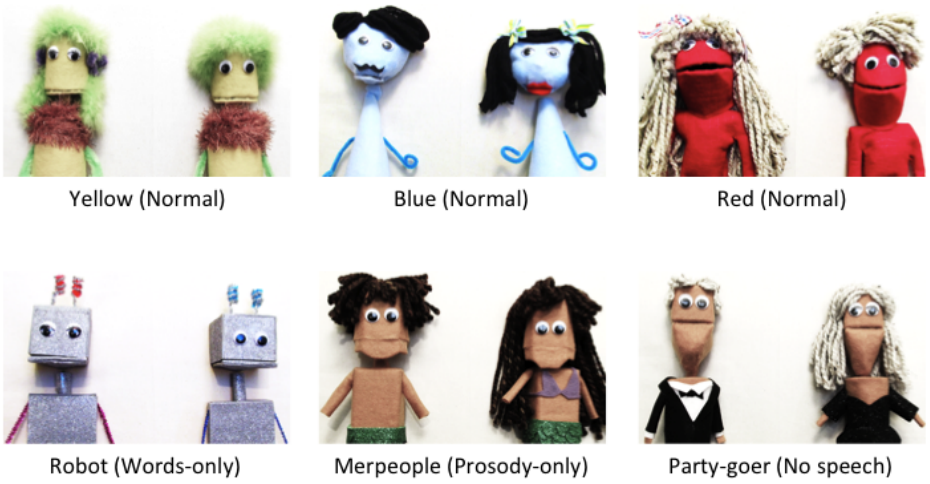


Figure 8: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

talkers were native English speakers, and were recorded in male-female pairs. As before, we asked talkers to speak "as if they were on a children's television show" and to ask at least a few questions during the improvisation. We cut each audio recording down to the 20-second interval with the most turn activity. The 20-second clips were then phonetically manipulated and used in the final video stimuli.

*Audio Manipulation.* We created four versions of each audio clip: *normal*, *words only*, *prosody only*, and *no speech*. That is, one version with a full linguistic signal (*normal*), and three with incomplete linguistic information (hereafter "limited cue" conditions). The *normal* clips were the unmanipulated, original audio clips.

The *words only* clips were manipulated to have robot-like speech: we flattened the intonation contours to each talker's average pitch (F0) and we reset the duration of every nucleus and coda to each talker's average nucleus and coda duration.[7] We made duration and pitch manipulations using PSOLA resynthesis in Praat (Boersma and Weenink, 2012). Thus, the *words only* versions of the audio clips had no pitch or durational cues

[7]We excluded hyper-lengthened words like [waʊː] 'woooow!'. These were rare in the clips.

to upcoming turn boundaries, but did have intact lexicosyntactic cues (and residual phonetic correlates of prosody, like intensity).

We created the *prosody only* clips by low-pass filtering the original recording at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006). This manipulation creates a "muffled speech" sound because low-pass filtering removes most of the phonetic information used to distinguish between phonemes. The *prosody only* versions of the audio clips lacked lexical information, but retained their intonational and rhythmic cues to upcoming turn boundaries.

The *no speech* condition served as a non-linguistic baseline. For this condition, we replaced the original clip with multi-talker babble: Eight different child-oriented conversations (not including the original one), overlaid and cropped to the duration of the video. Thus, the *no speech* audio clips lacked any linguistic information to upcoming turn boundaries—the only cue to turn taking was the opening and closing of the puppets' mouths.

Finally, because low-pass filtering removes significant acoustic energy, the *prosody only* clips were much quieter than the other three conditions. Our last step was to downscale the intensity of videos from the three other conditions to match the volume of the *prosody only* clips. We referred to the conditions as "normal", "robot", "mermaid", and "birthday party" speech when interacting with participants.

*Video recordings.* We created puppet video recordings to match the manipulated 20-second audio clips. The puppets were minimally expressive; the experimenter could only control the opening and closing of their mouths; their head, eyes, arms, and body stayed still. Puppets were positioned looking forward to eliminate shared gaze as a cue to turn structure (Thorgrímsson et al., under review). We took care to match the puppets' mouth movements to the syllable onsets as closely as possible, and avoided any mouth movement before the onset of a turn. We then added the manipulated audio clips to the puppet video recordings by hand.

We used three pairs of puppets used for the *normal* condition—'red', 'blue' and 'yellow'—and one pair of puppets for each limited cue condition: "robots", "merpeople", and "party-goers" (Figure 8). We randomly assigned half of the conversation topics ('birthday cake', 'pets', and 'breakfast') to the *normal* condition, and half to the limited cue conditions ('riding bikes', 'rainy days', and 'the library'). We then created three versions of the experiment, so that each of the six puppet pairs was associated with three different conversation topics across the different versions of the experiment (18 videos

in total). We ensured that the position of the talkers (left and right) was counterbalanced in each version by flipping the video and audio channels as needed.[8]

The duration of turn transitions and the number of speaker changes across videos was variable because the conversations were recorded semi-spontaneously. We measured turn transitions from the audio recording of the *normal*, *words only*, and *prosody only* conditions. There was no audio from the original conversation in the *no speech* condition videos, so we measured turn transitions from the video recording, using ELAN video editing software (Wittenburg et al., 2006).

There were 79 turn transitions for analysis, after excluding transitions longer than 550 msec and shorter than 0 msec. The remaining turn transitions were distributed evenly across transition types (questions N=47 and non-questions N=32) and conditions, keeping in mind that there were three *normal* videos and only one limited cue video for each experiment version: *normal* (N=36), *words only* (N=13), *prosody only* (N=12), and no speech (N=18). Inter-turn gaps for questions (mean=358, median=405) were longer than those for non-questions (mean=296, median=288) on average, but gap duration was overall comparable across conditions: *normal* (mean=333, median=337), *words only* (mean=345, median=383), *prosody only* (mean=320, median=336), and *no words* (mean=324, median=328). The longer gaps for question transitions could give them an advantage because our anticipatory measure includes shifts initiated during the gap between turns (Figure 2).

### *3.2. Procedure*

We used the same experimental apparatus and procedure as in the first experiment. Each participant watched six puppet videos in random order, with five 15–30 second filler videos placed in-between (e.g., running puppies, moving balls, flying bugs). Three of the puppet videos had *normal* audio while the other three had *words only*, *prosody only*, and *no speech* audio. This experiment required no special instructions so the experimenter immediately began each session with calibration (same as before) and then stimulus presentation. The entire experiment took less than five minutes.

*Data preparation, coding, and random baseline analysis.* We coded each

[8]See the videos here: www.youtube.com/channel/UCGEZQcM9t8Zfjqqi_B1Q5Sw.

| Age group | Speaker | Addressee | Other onscreen | Offscreen |
|---|---|---|---|---|
| 1 | 0.44 | 0.14 | 0.23 | 0.19 |
| 2 | 0.50 | 0.13 | 0.24 | 0.14 |
| 3 | 0.47 | 0.12 | 0.25 | 0.16 |
| 4 | 0.48 | 0.11 | 0.29 | 0.12 |
| 5 | 0.54 | 0.11 | 0.20 | 0.14 |
| 6 | 0.60 | 0.12 | 0.18 | 0.10 |
| Adult | 0.69 | 0.12 | 0.09 | 0.10 |

Table 2: Average proportion of gaze to the current speaker and addressee during periods of talk across ages.

turn transition for its linguistic condition (*normal*, *words only*, *prosody only*, and *no speech*) and transition type (question/non-question)[9], identified anticipatory gaze switches to the upcoming speaker, and estimated a random baseline of anticipatory switches using the methods from Experiment 1.

### *3.3. Results and discussion*

#### *3.3.1. Descriptive analysis*

Participants' pattern of gaze indicated that they performed basic turn tracking across all ages and in all conditions. Participants looked at the screen most of the time during video playback (82% and 86% average for children and adults, respectively). Children and adults primarily kept their eyes on the person who was currently speaking: they gazed at the current speaker between 44% and 69% of the time, looking back at the addressee between 11% and 14% of the time (Table 2). They tracked the current speaker in every condition—even one-year-olds looked more at the current speaker than at anything else in the three limited cue conditions (40% for *words only*, 43% for *prosody only*, and 39% for *no speech*). There was a steady overall increase in looks to the current speaker with age and added linguistic information (Tables 2 and 3). Looks to the addressee also decreased with age, but the change was minimal.

We identified anticipatory gaze switches for all 79 usable turn transitions using the same criteria as in Experiment 1. We expected to see the

[9]We coded *wh*-questions as "non-questions" for the *prosody only* videos. Polar questions had a final rising prosodic contour, but *wh*-questions did not (Hedberg et al., 2010).

| Condition | Speaker | Addressee | Other onscreen | Offscreen |
|---|---|---|---|---|
| Normal | 0.58 | 0.12 | 0.17 | 0.13 |
| Words only | 0.54 | 0.11 | 0.24 | 0.10 |
| Prosody only | 0.48 | 0.12 | 0.26 | 0.15 |
| No speech | 0.44 | 0.13 | 0.26 | 0.18 |

Table 3: Average proportion of gaze to the current speaker and addressee during periods of talk across conditions.

most anticipatory switches in the *normal* condition and the least anticipatory switches in the *no speech* condition because they contained the most and least linguistic information, respectively. We expected to replicate our finding that questions result in more anticipatory switches than non-questions, with the added hypothesis that the question effect is driven by lexicosyntactic cues. We also anticipated an overall increase in anticipatory switches with age. Finally, since the development of prosodic skills partially precedes the development of lexicosyntax, we expected to see more switches in the *prosody only* condition compared to the *words only* condition in the youngest age groups.

We found that participants at all ages made anticipatory gaze switches more often than would be expected by chance in almost all conditions, but sometimes by a small margin (Figure 9). For example, anticipatory switches in the *no speech* videos were often modestly above chance (with the exception of 5- and 6-year-olds). One interpretation of this finding is that children do not use linguistic information to predict upcoming turns—the only cue to turn taking in the *no speech* condition was the alternating mouth movements of the conversing puppets. With sufficiently long inter-turn gaps, participants could simply switch their gaze to the responder when the prior talker finishes talking—a strategy that doesn't rely on linguistic information. But this interpretation is not compatible with our the effects of linguistic information that do emerge in the data: an early advantage for prosody and a later advantage for lexical question cues.

We saw a generalize early advantage for prosody over lexicosyntax in children's spontaneous turn predictions. Notably, while children performed well with *prosody only* speech from the start, they seemed to only reliably exceed chance in the *words only* condition at ages three and older. By age three, the general advantage for prosody over lexicosyntax seems to disappear with

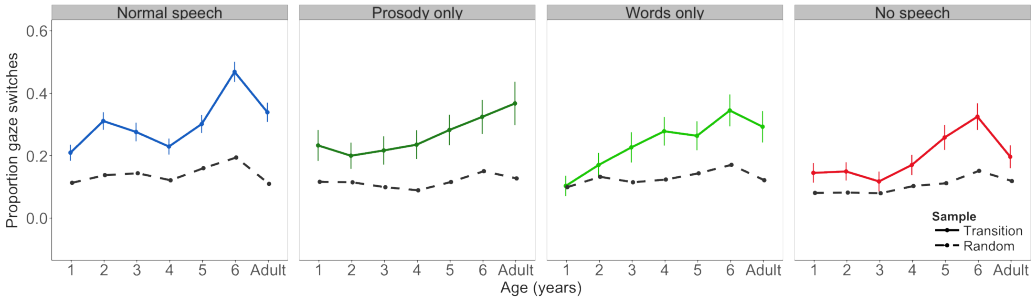


Figure 9: Baseline-corrected proportion of anticipatory gaze switches made for actual (solid) and randomly-shuffled (dashed) turn transitions in each condition, and across ages. Error bars represent the standard error of the mean.

more equal rates of switching overall for the two partial linguistic conditions.

We observed that, overall, children and adults made the fewest anticipatory switches in the *no speech* videos (9%) and the most in the *normal* videos (16%), with the *prosody only* and *words only* videos falling in-between (14% and 11%, respectively; Figure 10). However, these condition-based differences were primarily driven by transition type effects (question vs. non-question).

As in Experiment 1, we made condition, age and transition type comparisons based on baselin-corrected switching values; we corrected each participant's gaze data by subtracting their random baseline from their actual switching for each turn transition.

We found that participants made more anticipatory gaze switches following questions than non-questions, but only in the two lexical conditions (*normal* and *words only* speech). The advantage for questions was strongest in the *normal* speech condition, where it emerged early in development. The question effect became larger with age in both lexical conditions. While children made small, sometimes non-linear, gains in their anticipatory switches for non-questions as they grew older, they made large gains in their anticipatory switches with age (especially between ages one and two; Figure 11). The increased switching for questions between ages one to two indicates that, around this age, children begin to leverage lexicosyntactic cues in the speech signal to identify questions in ongoing conversation. In the lexical conditions, children had access to early cues to questionhood, such as *wh*-words and subject-auxiliary inversion, to predict an upcoming speaker change. In

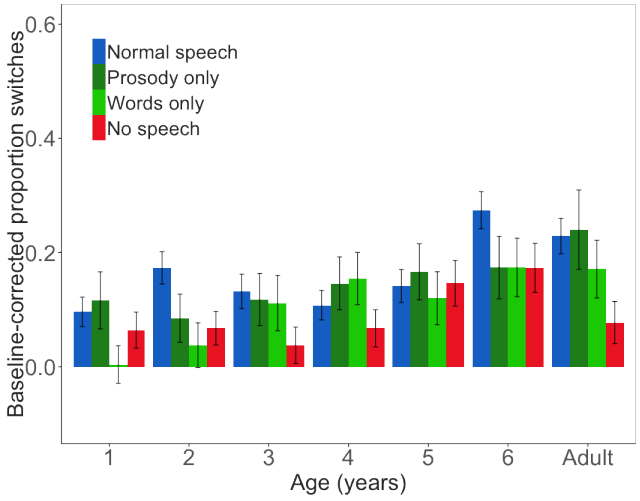


Figure 10: Baseline-corrected proportion of anticipatory gaze switches made for turn transitions in each speech condition, and across ages. Error bars represent the standard error of the mean.

the *normal* condition, where the question effect was strongest, lexical and syntactic cues were reinforced by question prosody for polar questions.

Children showed increased anticipatory switching with age across all conditions, making almost as many transitions as adults by ages five and six. Adults reacted to the *no speech* condition differently than the children did, tracking the interaction less closely than in the three linguistic conditions. They may have minimized switching in the *no speech* condition because they realized that there was no conversation to follow.

### *3.3.2. Statistical analysis*

We tested these findings in two linear mixed-effects models of participants' baseline-corrected anticipatory switches—one model each for children and adults. The child model included condition (*normal*, *words only*, *prosody only*, and *no speech*; *normal* in the intercept), transition type (question vs. non-question), age, and gap duration as fixed effects, with full interactions between condition, transition type, and age. The model also included random effects of item (turn transition) and participant, with random slopes of condition and transition type for participants. The adult model matched the child model, excluding age-related effects.

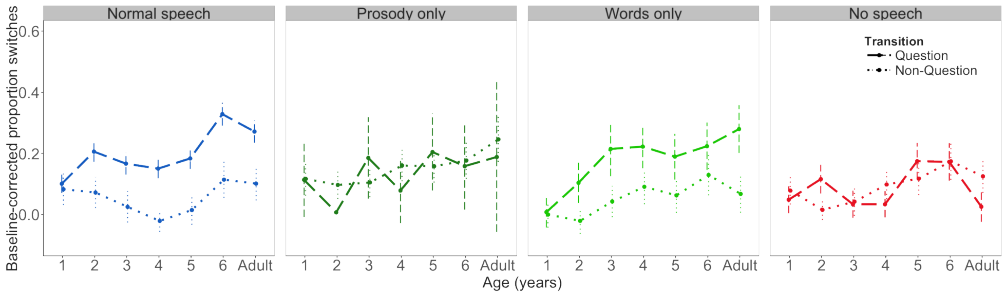


Figure 11: Baseline-corrected proportion of anticipatory gaze switches made for question- (dashed) and non-question- (dotted) turn transitions in each condition, and across ages. Error bars represent the standard error of the mean.

There was a significant effect of gap duration. Children had time to make more anticipatory switches in analysis windows with longer gaps ($\beta$=0.294, $SE$=0.091, $t$=3.231, $p$<.01). An effect of gap duration is expected since prior work using spontaneous anticipatory switching has found that most switches occur in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, 2011).

The only effects of linguistic condition derived from interactions with age and transition type, despite large differences in the linguistic information available across conditions. (Figure 11). The model of children's data confirmed two three-way effects of age, condition, and transition type. Children made significantly more anticipatory switches for questions with age in the *normal* ($\beta$=0.04, $SE$=0.008, $t$=5.168, $p$<.001) and *words only* conditions ($\beta$=0.031, $SE$=0.01, $t$=3.170, $p$<.01). Overall, they made a switch twice as often after hearing a question (16%) than a non-question (8%) in the *words only* condition, and nearly four times as often in the *normal* condition (19% vs. 5%). Since the question effect was limited to the lexical conditions, there was no main effect of transition type. There were also no further interactions of age and condition, or of transition type and condition. There was, however, a significant two-way interaction of transition type and age ($\beta$=-0.031, $SE$=0.01, $t$=-3.18, $p$<.01) in addition to the three-way interactions for the *normal* and *words only* conditions already mentioned.

Last but not least, we confirmed that children generally made more anticipatory switches as they got older: The model showed a significant main effect of age ($\beta$=0.028, $SE$=0.009, $t$=3.018, $p$<.01). Though the increase in

anticipatory looks with age came, in large part, from questions effects in the lexical conditions, children also showed increased switches for non-questions and non-lexical conditions. The increases related to non-questions and non-lexical conditions were overall smaller and sometimes followed a non-linear trajectory (Figure 11). As in Experiment 1, early switches (planned before the end of the prior turn) did not show large increases with age, with adults making the fewest early switches in the three linguistic conditions (Figure 12).

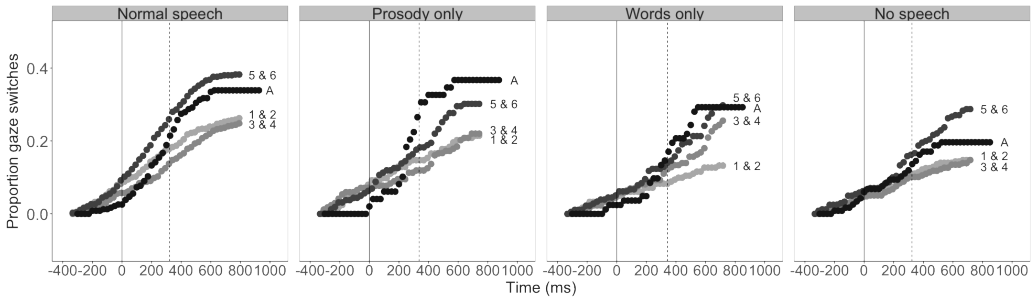


Figure 12: Cumulative proportion of anticipatory gaze switches planned during the transition window, across conditions and ages. Age is indicated by hue (darker = older). The solid line is the offset of the pre-gap turn, and the dashed line is the median onset of the post-gap turn in each condition. "Early switches" occur before 0.

The advantage for questions over non-questions in these results reinforces our findings from Experiment 1. They suggest that even two-year-olds can reliably distinguish questions from other transition types in third-party conversation, especially when full linguistic information is available. Inspecting the data more closely, we found that 43% of one-year-olds made more anticipatory switches after questions than non-questions in the *normal* condition, and 24% of them did so in the *words only* condition. Meanwhile, 61% of two-year-olds show the effect in the *normal* condition, and 52% of them do in the *words only* condition, which is already close to the average for older children and adults (Table 4).

Our results suggest that the question effect emerges between ages one and two, but further testing is needed to establish whether this is the case. For example, the size and emergence of the question effect for various question types and question cues could differ (*wh-* vs. *yes-no*; early vs. late in the utterance), but we do not have enough data to make those distinctions in the

| Age group | Normal speech | No speech |
|---|---|---|
| 1 | 0.43 | 0.24 |
| 2 | 0.61 | 0.52 |
| 3 | 0.63 | 0.42 |
| 4 | 0.78 | 0.52 |
| 5 | 0.91 | 0.52 |
| 6 | 0.75 | 0.50 |
| Adult | 0.67 | 0.57 |

Table 4: Proportion of participants showing numerically more anticipatory switches for questions than non-questions in the *normal* and *words only* conditions.

current analysis. Also, 43% of one-year-olds already showed more switching for questions than non-questions in the *normal* speech condition—the only condition with the full range of linguistic cues children hear in everyday life. Future work with more fine-grained age samples and better controlled *normal* speech might then find that children can identify questions in conversation even earlier.

Finally, we confirmed that young children showed a benefit for prosody over lexicosyntax. We conducted two *post-hoc* $t$-tests for children under three, comparing their anticipatory switching rate against their estimated random switching rate in the *prosody only* and *words only* conditions. We found that anticipatory gazes significantly exceeded chance for the *prosody only* condition ($t(187.677)$=-3.004, $p$=.003), but not for the *words only* condition ($t(229.828)$=-0.8307, $p$=.407). This finding suggests that there is indeed an early advantage for prosody over lexicosyntax in children under three. The advantage for prosody was not driven by a question effect; children and adults did not consistently show more switches after questions than non-questions in the *prosody only* condition (Figure 11). So, whereas the early advantage for prosody appeared for both transition types (question and non-question), the later-emerging advantage for lexicosyntax and prosody together was primarily limited to question transitions. This pattern of results may suggest a qualitative shift in the kind of predictions children make while observing a conversation, with linguistic cues (like prosody) playing different roles at different ages.

The model of adults' anticipatory gaze switches confirmed main effects of gap duration and linguistic condition. Longer gaps led to more anticipatory

switches ($\beta$=0.656, $SE$=0.176, $t$=3.726, $p$<.001). Adults also made significantly more anticipatory switches in the *normal* condition than in the *no speech* baseline, indicating that linguistic access did affect their overall performance ($\beta$=-0.211, $SE$=0.085, $t$=-2.486, $p$<.05). There were no other main effects of linguistic condition. The model also showed a marginal two-way interaction of condition and transition type, with adults making more anticipatory gazes for questions in the *normal* condition ($\beta$=0.194, $SE$=0.11, $t$=1.773, $p$=.076). Though the question effect was numerically present in both lexical conditions, it was only significant for adults in the *normal* condition, in which they had both lexicosyntactic and prosodic cues to questionhood. This result suggests that prosodic cues still have a role to play in adult question recognition during third-party conversation. Since question effects were primarily limited to the *normal* condition, there was no main effect of transition type in the adults' data.

*3.3.3. Summary*

The core aims of Experiment 2 were to gain better traction on the individual roles of prosody and lexicosyntax in children's turn predictions, and to expand our age range to capture more developmental change. We found that effects of linguistic processing and age were present in this dataset, but both were primarily found within the question effect. Children and adults made more anticipatory switches after questions than non-questions, but did so more often when they had greater access to linguistic information, whether increased access came from age or with speech condition. Our results hint that lexical information is critical in identifying questions during conversation, especially for children. Even so, performance was consistently best in the *normal* condition, in which participants could integrate lexical and prosodic information together.

By extending the age range of participants, we were able to see more of the developmental trajectory in children's predictions about turn structure. We saw that one-year-olds already made more anticipatory switches than would be expected by chance in two of the four conditions. As they got older, children demonstrated above-chance anticipatory switches in more conditions and made more anticipatory switches overall. Like in Experiment 1, children between ages three and five behaved similarly to each other, and comparably to adults, showing small but steady gains in their anticipatory switches with age. Though the advantage for questions generally increased with age, the biggest developmental shift in our data occurred between one

and two years, when the question effect began to reliably emerge in the two lexical conditions. It will be up to future work to disentangle which lexical and prosodic cues children begin to pick out in distinguishing questions from other speech acts, and when.

The results from Experiment 2 complement those of Experiment 1, but still leave some open questions. For example, we were surprised to find no independent effects of linguistic condition outside of age and transition type interactions. In fact, the primary evidence that participants actually *used* linguistic information came from the question effects, which were limited to the two lexical conditions. Given prior work, it is surprising that taking away lexical and prosodic information didn't have a larger impact on participants' anticipatory looking. We take this issue up further in the general discussion.

We found that the importance of lexicosyntactic and prosodic information for turn prediction differed depending on age and transition type. While even one-year-olds made anticipatory switches at an above-chance rate in the *normal* and *prosody only* conditions, children didn't consistently surpass chance in the *words only* condition until age three. This pattern of development suggests an early importance for prosodic information in children's turn anticipation. In contrast, the question effect—a primary driver of anticipatory switches in our data—seems to rely on lexical information, with some room for further support from prosody. So, although prosody may help children early on, it does not, on its own, help children to distinguish questions from non-questions. Rather, for both adults and children, prosody's role in the question effect was an additive one: The question effect was strongest when children had access to lexical *and* prosodic information. In short, prosody has different roles to play in turn prediction before and after the lexically-driven question effect emerges. Importantly, these results do not support a straightforward view of lexicosyntax as the primary linguistic cue for turn predictions (even for adults; De Ruiter et al., 2006; Magyari and De Ruiter, 2012), but do support accounts of multiple cue integration (Duncan, 1972; Ford and Thompson, 1996; Hirvenkari et al., 2013; Torreira et al., under review).

We controlled linguistic information in Experiment 2 by using phonetically-manipulated speech, extending speech manipulations used in prior work (De Ruiter et al., 2006; Keitel et al., 2013). While phonetic manipulation gave us control over the linguistic signal, it also created speech sounds that children don't usually hear in their natural environment. Many prior studies have used phonetically-altered speech with infants and young children (cf.

Jusczyk, 2000), but almost none of them have done so in a conversational context. We found that the *normal* speech condition showed the strongest anticipation effects over all, and we have attributed this to the fact that it gave participants full (lexicosyntactic and prosodic) linguistic information. However, it was also the only natural speech condition. Children could have had trouble processing the *words only* and *prosody only* conditions because they were unfamiliar, and not just because they had less linguistic information available.

It would take a different type of stimulus to compare lexical and prosodic information without phonetic manipulation. For example, carefully scripted or cross-spliced conversations could control for the presence or absence of prosodic and lexicosyntactic cues to turn transition with natural speech. But since the non-natural (limited cue) linguistic conditions did still elicit some differences in anticipation, children must have processed at least some linguistic information in the *prosody only* and *words only* conditions. Perhaps the advantage of the *normal* condition was then not entirely due to its naturalness. We expect that future work with spontaneous conversational measures will still find an advantage for combined lexical and prosodic cues over lexical cues alone (see also Duncan, 1972; Ford and Thompson, 1996; Torreira et al., under review).

## 4. General Discussion

Children begin to develop conversational turn-taking skills long before their first words (Bateson, 1975; Hilbrink et al., in preparation; Jaffe et al., 2001; Snow, 1977). As they acquire language, they also acquire the information needed to make accurate predictions about upcoming turn structure. Until recently, we have had very little data on how children weave language into their already-existing turn-taking behaviors.

In two experiments investigating children's anticipatory gaze to upcoming speakers, we found evidence that turn prediction develops early in childhood, and that linguistic cues are integrated differently at different ages. Children under three showed a general advantage for prosody over lexicosyntax, contrary to findings for adults in prior work (De Ruiter et al., 2006). Children two and older also showed a slight advantage for lexicosyntax, but only in the form of a transition type effect: they made more anticipatory switches following questions than non-questions in conditions where lexical information was present. Two-year-olds thus appeared to be at a developmental tran-

sition point, still showing an overall advantage for prosody, but beginning to show an advantage for lexicosyntax within the domain of question-answer sequences. We found no evidence that, for children or adults, lexicosyntax alone was "sufficient" (equal to full linguistic information) for spontaneous turn prediction (De Ruiter et al., 2006, pg. 531): Participants' performance was best in conditions when they had access to the full linguistic signal.

### *4.1. Developmental implications*

The primary aim of our study was to find out how linguistic information affects children's predictions about upcoming turn structure across development. Two primary developmental findings emerged from our data. First, participants' spontaneous predictions largely occurred during question–answer transitions. Second, children's use of linguistic signals in making predictions depended on their age, with an early advantage for prosody over lexicosyntax and a later (partial) advantage for lexicosyntax over prosody.

#### *4.1.1. The advantage for questions*

In both of our experiments we found a robust advantage for questions in participants' anticipatory switching because participants made more switches after questions than non-questions. Children and adults both treated questions as if they had a special interactive status, and with good reason: Questions make a speaker switch immediately relevant, helping the listener to predict what will happen next (i.e., an answer from the addressee). Linguistic cues to questionhood often occur early in the utterance, making questions easy to identify early on and thereby giving observers time to plan a reaction (i.e., a gaze switch). Also, the form of a question can constrain the type of response that will come next (e.g., a location after a *where* question), further increasing the predictability of the interactional sequence.

Children get frequent and early experience with hearing and answering questions. Questions make up approximately one third of the utterances children hear, before and after the onset of speech, and even into their preschool years (Fitneva, 2012; Henning et al., 2005; Shatz, 1979).[10] Many of the questions directed to young children are "test' questions—questions that the caregiver already has the answer to (e.g., "What does a cat say?"). Caregivers use questions to get their young children's attention and to ensure

[10]There is substantial variation in this proportion by individual and socioeconomic class (Hart and Risley, 1992).

that information is in common ground, even if the responses are non-verbal or infelicitous (Bruner, 1985; Fitneva, 2012; Snow, 1977). So, in addition to having a special interactive status, questions are a core characteristic of many caregiver-child interactions.

We cannot definitively say with our current data what property of questions caused children to switch their gaze more often. It is true that questions make a speaker transition more likely and that questions have distinctive linguistic cues, both early and late, to identify themselves. But other speech acts and request formats have these same properties. Imperatives, compliments, and complaints all make a response from the addressee likely (Schegloff, 2007). Rhetorical and tag questions, on the other hand, take a similar form to prototypical polar questions, but do not always require an answer. So, though it is clear that children anticipate responses more often for questions than non-questions, their actual understanding of questionhood as a linguistically- or interactionally-marked type of speech act still needs further exploration.

*4.1.2. Linguistic cues for turn prediction*

Prior work with adults has found a consistent and critical role for lexicosyntax in predicting upcoming turn structure (De Ruiter et al., 2006; Magyari and De Ruiter, 2012), with more debate over the role of prosody (Duncan, 1972; Ford and Thompson, 1996; Torreira et al., under review). Knowing that children comprehend more about prosody than lexicosyntax early on (see introduction; also see Speer and Ito, 2009 for a review), we thought it possible that young children would instead show an advantage for prosody over lexicosyntax. Our results suggest that there are roles for both prosody and lexicosyntax in children's predictions about upcoming turn structure, but that there are different roles for these information sources at different ages.

Early on, there was an advantage for prosody over lexicosyntax in children's anticipatory switches: Anticipatory switching didn't reliably exceed chance for lexicosyntactic information until age three, but was already significantly greater than chance for prosodic information at age one. This early advantage for prosody was not driven by the advantage for questions: While some one-year-olds in our study already made more switches after questions than non-questions, most did not, plus prosody alone did not elicit a robust question effect at any age (Figure 11).

Although children showed an initial advantage for prosody, their pre-

dictions before age three were probably limited to basic prosodic contours and boundaries, and simple cues such as final lengthening and turn-final silence (Pannekamp et al., 2006). More complex structural predictions would have required them to access lexicosyntactic information because higher-level structure in the prosodic signal is linked to the syntactic structure being used.

In both experiments, even young children showed an advantage for questions in conditions where they had lexicosyntax, suggesting that lexicosyntax is a critical cue for question identification. The question effect emerged by age two, when children's syntactic comprehension and lexical inventories are still relatively limited (Clark, 2009). That being said, syntactic and lexical markers to questionhood in English are frequent, categorical, and early-occurring in the utterance (e.g., "do", *wh*-words, and subject-auxiliary inversion). Lexicosyntactic question cues were available on every instance of *wh-* and *yes/no* questions in our stimuli, whereas prosodic question cues were only salient on *yes/no* questions. Thus, the unambiguous presence of lexicosyntactic cues on *wh-* and *yes/no* questions may have been one reason why we saw an advantage for questions in the lexical conditions, but not in the *prosody only* condition.

Importantly, the question effect was not driven purely by lexicosyntactic information. In both experiments, prosody still made a contribution to the advantage for questions, even though prosody alone was not enough to elicit a question effect. For example, older children and adults in Experiment 1 showed a question effect in the languages they did not understand—presumably they used prosodic and non-verbal information to distinguish questions from non-questions. In Experiment 2, the advantage for questions was biggest for speech with lexicosyntactic *and* prosodic information together. All in all, children appear to improve at identifying questions in ongoing conversation by using lexicosyntactic cues as they get older. But, prosodic cues still aid in picking out questions from ongoing conversation.

The close link between prosodic and syntactic structure makes it difficult to tease apart how predictive processing in one domain (e.g., prosody) is distinct from the other (e.g., syntax) in turn prediction. We leave this challenge to future work.

### *4.2. Contributions of the current work*

Our findings are consistent with observational work on children's turn taking. Young children are slow turn-takers in spontaneous conversation

(Casillas et al., under review) even though they begin taking turns in infancy (Hilbrink et al., in preparation; Jaffe et al., 2001). We found that children begin to spontaneously predict upcoming turn structure at age one. Children's gaze switches were also most common in the inter-turn gap, suggesting that they expected an immediate response from the addressee. Our results therefore support the idea of an early-developed competence for the timing of turn taking. But because children do not themselves use adult-like conversational timing until much later (age six; Ervin-Tripp, 1979), our results also indirectly point toward response planning as a key factor contributing to children's slow responses in real-time interaction (Casillas et al., under review).

Our findings are only partially consistent with prior experimental work on turn taking. De Ruiter and colleagues (2006) found that adults' accuracy in identifying upcoming turn-ends was not significantly affected when intonation contours were flattened. Keitel and colleagues (2013) found the same for adults' anticipatory gaze switches, only finding that intonation flattening affected children's switches at 36 months (with switches by younger children never exceeding chance). Our results, in contrast, suggest that children's anticipatory switches emerge much earlier, and rely on prosody before they do lexicosyntax. Furthermore, our adult control participants patterned similarly to the children: They showed question effects for languages they didn't speak (Experiment 1) and showed the strongest question effects for lexicosyntax and prosody together (Experiment 2). Thus, neither the child nor the adult data support the idea that lexicosyntactic information is equal to full linguistic information in making predictions about upcoming turn structure (De Ruiter et al., 2006), at least in participants' spontaneous looking behavior. This may partially be due to the fact that we controlled syllable duration in addition to intonation, thereby eliminating two of the most salient cues to prosodic structure.

Despite extreme differences in linguistic information across conditions, independent effects of linguistic information (outside of interactions) were minimal in our experiments. But, using a button-press measure with similar stimuli, de Ruiter and colleagues (2006) found robust effects of linguistic condition. This disparity in findings could be partially explained by task differences. A button-press method explicitly encourages participants to make an early response, while observer gaze measures are purely passive. Participants' spontaneous anticipatory gaze switches may not elicit enough early responses to see large linguistically-driven differences in prediction. Indeed,

most switches happened during the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, 2011). Observers' gaze patterns suggest that they aren't targeting potential turn-ends with their predictions. Rather, they are targeting switches in speakership. So although button press measures may be better equipped to tell us what participants *can* do in predicting turn ends, they may also emphasize early responding (and consequently, lexicosyntactic cues) to a greater extent than we experience in everyday conversation. Within observer gaze experiments, questions may be the key to naturally eliciting early responses. We have seen in the current study that questions are interactionally sufficient to motivate early predictions.

Moving away from button press measures, Keitel and colleagues (2013) used observer gaze to investigate children's turn predictions, finding: (a) no evidence that children under three make switches above chance, (b) that three-year-olds' switches are significantly affected by flattened intonation, and (c) that adults' switches are not. Our current results are consistent with the general finding from their study, that prosody matters for young children's turn predictions. But our results also differ from theirs in a few important ways.

First, we found that children begin making above-chance anticipatory switches much earlier than three—our one-year-olds were already looking above chance in two of our four linguistic conditions in Experiment 2. This earlier emergence of anticipatory switching could be due to our more child-friendly stimuli, which we designed to be engaging and to mimic the kind of speech children often experience interactionally. Second, because we included more linguistic conditions to isolate the effects of prosody and lexicosyntax, we gained a broader perspective on how linguistic information changed children's predictions. Ultimately, the added conditions allowed us to find an early advantage for prosody and a later advantage for lexicosyntax.

Third, we found that anticipatory looking was primarily driven by question transitions, a pattern that has so far not been discussed in other observer gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, 2011). The question effects in the current study were critical to understanding the role of lexicosyntactic information in participants' turn predictions. Finally, we found that our adult controls were affected by the absence of prosody (within the question effect) while Keitel et al. (2013) found that adults' were unaffected by flattened pitch. But again, because we flattened pitch for individual speakers and controlled syllable duration, we are not able to directly compare our findings to theirs.

*4.3. Limitations*

In Experiment 2, we tried to improve our method for controlling linguistic information by phonetically manipulating improvised dialogic speech. Although this decision allowed us to follow prior work more closely (De Ruiter et al., 2006; Keitel et al., 2013), it also limits our interpretation of children's behavior. Low-pass filtered and robotic-like speech are not typical within children's speech environment, and so it may have been unfair to compare the limited cue conditions in Experiment 2 to the normal speech condition, which not only had full linguistic cues, but is also more typically present in children's speech environments. One alternative to explore for the future would be scripted speech, cross-spliced speech, or speech that is carefully selected from a large corpus, given a set of specific linguistic constraints.

Relatedly, since we used improvised conversation recordings, we did not control how often the speakers switched back and forth, or how often a switch occurred after a pause in speech. It is possible that, to some extent, children could have formed expectations about the turn structure on the basis of the frequent switching alone (see, e.g., Thorgrímsson et al., under review). If this statistical expectation were the only thing leading children's anticipatory gaze switches, we would have predicted more uniformity in looking behavior across the conditions and no major effects of linguistic signal, age, or transition type. But that is not what we saw. Nevertheless, another advantage of using scripted conversation would be the ability to control switch frequency and eliminate this issue.

Finally, we carefully designed a scheme for identifying anticipatory turn switches in the data, building off of Keitel and colleagues' (2013) analyses. In both our study and theirs, anticipatory switches that happen very early in the turn are not counted. Instead, they would actually count as "random" switches in the baseline analyses. These same early switches are counted as non-random in other studies of children's gaze to conversational stimuli (Bakker et al., 2011; von Hofsten et al., 2009). Since most studies, even those using a wider switching window (e.g., Hirvenkari et al., 2013) find that switches primarily happen close to or immediately after the inter-turn gap begins, the approach used here and in Keitel et al. (2013) probably captures most of participants' turn-relevant switching behavior. But if early looks do indeed increase as children get older and have more access to early syntactic cues, it would be helpful to develop a second measure of very early anticipatory looks.

*4.4. Conclusions*

Conversation plays a central role in children's language learning. It is the driving force behind what children say and what they hear. And yet many conversational skills, like turn taking, rely heavily on language. Adults use language to accurately predict turn structure in conversation, which facilitates their online comprehension and allows them to respond relevantly and on time. But young children do not have enough linguistic knowledge to take turns the way that adults do. Instead, they first acquire a preverbal turn-taking system for interaction, and then somehow assimilate language into that system as they acquire it. The simultaneous acquisition and integration of interactive and linguistic skills is no easy task, as evidenced by children's protracted development of turn-taking skills. In the current study we have investigated children's language acquisition through the lens of turn taking, thereby bringing into focus children's application of linguistic knowledge in real-time interaction. We found that children acquire basic predictive skills early on, but that the integration of linguistic cues takes time and is primarily driven by sequences of action—in our case, by questions and their answers.

## Acknowledgements

We gratefully acknowledge the parents and children at Bing Nursery School and the Children's Discovery Museum of San Jose. This work was supported by an ERC Advanced Grant to Stephen C. Levinson (269484-INTERACT), NSF graduate research and dissertation improvement fellowships to the first author, and a Merck Foundation fellowship to the second author. Earlier versions of these data and analyses were presented to conference audiences (Casillas and Frank, 2012, 2013). We also thank Tania Henetz, Francisco Torreira, Stephen C. Levinson, Eve V. Clark, and the First Language Acquisition group at Radboud University for their feedback on earlier versions of this work.

## References

Augusti, E.M., Melinder, A., Gredebäck, G., 2010. Look who's talking: Pre-verbal infants' perception of face-to-face and back- to-back social interactions. Developmental Psychology 1, 161.

Bakker, M., Kochukhova, O., von Hofsten, C., 2011. Development of social perception: A conversation study of 6-, 12-and 36-month-old children. Infant Behavior and Development 34, 363–370.

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, 255–278.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4: Linear mixed-effects models using Eigen and S4. URL: `https://github.com/lme4/lme4/http://lme4.r-forge.r-project.org/`. [Computer program] R package version 1.1-7.

Bateson, M.C., 1975. Mother-infant exchanges: The epigenesis of conversational interaction. Annals of the New York Academy of Sciences 263, 101–113.

Beier, J.S., Spelke, E.S., 2012. Infants' developing understanding of social gaze. Child Development 83, 486–496.

Bergelson, E., Swingley, D., 2013. The acquisition of abstract words by young infants. Cognition 127, 391–397.

Bloom, K., 1988. Quality of adult vocalizations affects the quality of infant vocalizations. Journal of Child Language 15, 469–480.

Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer. URL: `http://www.praat.org`. [Computer program] Version 5.3.16.

Bögels, S., Magyari, L., Levinson, S.C., in preparation. Language production planning during turn-taking starts as soon as possible .

Bruner, J., 1985. Child's talk: Learning to use language. Child Language Teaching and Therapy 1, 111–114.

Bruner, J.S., 1975. The ontogenesis of speech acts. Journal of Child Language 2, 1–19.

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

Casillas, M., Bobb, S.C., Clark, E.V., under review. Turn taking, timing, and planning in early language acquisition .

Casillas, M., Frank, M.C., 2012. Cues to turn boundary prediction in adults and preschoolers. Proceedings of SemDial .

Casillas, M., Frank, M.C., 2013. The development of predictive processes in children's discourse understanding, in: Proceedings of the 35th Annual Meeting of the Cognitive Science Society.

Clark, E.V., 2009. First language acquisition. Cambridge University Press.

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

De Vos, C., Torreira, F., Levinson, S.C., in preparation. Stroke-to-stroke turn boundaries in signed conversations .

Dingemanse, M., Torreira, F., Enfield, N., 2013. Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. PloS one 8, e78273.

Duncan, S., 1972. Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23, 283.

Ervin-Tripp, S., 1979. Children's verbal turn-taking, in: Ochs, E., Schieffelin, B.B. (Eds.), Developmental Pragmatics. Academic Press, New York, pp. 391–414.

Fernald, A., Zangl, R., Portillo, A.L., Marchman, V.A., 2008. Looking while listening: Using eye movements to monitor spoken language. Developmental Psycholinguistics: On-line methods in children's language processing , 113–132.

Fitneva, S., 2012. Beyond answers: questions and children's learning, in: de Ruiter, J.P. (Ed.), Questions: Formal, Functional, and Interactional Perspectives. Cambridge University Press, Cambridge, UK, pp. 165–178.

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

Garvey, C., 1984. Children's Talk. volume 21. Harvard University Press.

Gísladóttir, R., Chwilla, D., Levinson, S.C., under review. Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. .

Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. Psychological science 11, 274–279.

Hart, B., Risley, T.R., 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. Developmental Psychology 28, 1096.

Hedberg, N., Sosa, J.M., Görgülü, E., Mameni, M., 2010. The prosody and meaning of Wh-questions in American English, in: Speech Prosody 2010–Fifth International Conference.

Henning, A., Striano, T., Lieven, E.V., 2005. Maternal speech to infants at 1 and 3 months of age. Infant Behavior and Development 28, 519–536.

Hilbrink, E., Gattis, M., Levinson, S.C., in preparation .

Hirvenkari, L., Ruusuvuori, J., Saarinen, V.M., Kivioja, M., Peräkylä, A., Hari, R., 2013. Influence of turn-taking in a two-person conversation on the gaze of a viewer. PloS one 8, e71569.

von Hofsten, C., Uhlig, H., Adell, M., Kochukhova, O., 2009. How children with autism look at events. Research in Autism Spectrum Disorders 3, 556–569.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C.L., Jasnow, M.D., Rochat, P., Stern, D.N., 2001. Rhythms of dialogue in infancy: Coordinated timing in development. Monographs of the Society for Research in Child Development. JSTOR.

Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. Journal of Memory and Language 44, 548–567.

Jusczyk, P.W., 2000. The Discovery of Spoken Language. MIT press.

Jusczyk, P.W., Cutler, A., Redanz, N.J., 1993. Infants' preference for the predominant stress patterns of English words. Child Development 64, 675–687.

Jusczyk, P.W., Hohne, E., Mandel, D., Strange, W., 1995. Picking up regularities in the sound structure of the native language. Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research , 91–119.

Kamide, Y., Altmann, G., Haywood, S.L., 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language 49, 133–156.

Keitel, A., Prinz, W., Friederici, A.D., Hofsten, C.v., Daum, M.M., 2013. Perception of conversations: The importance of semantics and intonation in childrens development. Journal of Experimental Child Psychology 116, 264–277.

Lemasson, A., Glas, L., Barbu, S., Lacroix, A., Guilloux, M., Remeuf, K., Koda, H., 2011. Youngsters do not pay attention to conversational rules: is this so for nonhuman primates? Nature Scientific Reports 1.

Levelt, W.J., 1989. Speaking: From intention to articulation. MIT press.

Levinson, S.C., 2013. Action formation and ascriptions, in: Stivers, T., Sidnell, J. (Eds.), The Handbook of Conversation Analysis. Wiley-Blackwell, Malden, MA, pp. 103–130.

Magyari, L., Bastiaansen, M.C.M., De Ruiter, J.P., Levinson, S.C., In press. Early anticipation lies behind the speed of response in conversation. Journal of Cognitive Neuroscience .

Magyari, L., De Ruiter, J.P., 2012. Prediction of turn-ends based on anticipation of upcoming words. Frontiers in Psychology 3:376, 1–9.

Mandel, D.R., Kemler Nelson, D.G., Jusczyk, P.W., 1996. Infants remember the order of words in a spoken sentence. Cognitive Development 11, 181–196.

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, 303–312.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178.

Moon, C., Cooper, R.P., Fifer, W.P., 1993. Two-day-olds prefer their native language. Infant Behavior and Development 16, 495–500.

Morgan, J.L., Saffran, J.R., 1995. Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. Child Development 66, 911–936.

Nazzi, T., Ramus, F., 2003. Perception and acquisition of linguistic rhythm by infants. Speech Communication 41, 233–243.

Pannekamp, A., Weber, C., Friederici, A.D., 2006. Prosodic processing at the sentence level in infants. NeuroReport 17, 675–678.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`. [Computer program] Version 3.1.1.

Ratner, N., Bruner, J., 1978. Games, social exchange and the acquisition of language. Journal of Child Language 5, 391–401.

Rochat, P., Neisser, U., Marian, V., 1998. Are young infants sensitive to interpersonal contingency? Infant Behavior and Development 21, 355–366.

Ross, H.S., Lollis, S.P., 1987. Communication within infant social games. Developmental Psychology 23, 241.

Rossano, F., Brown, P., Levinson, S.C., 2009. Gaze, questioning and culture, in: Sidnell, J. (Ed.), Conversation Analysis: Comparative Perspectives. Cambridge University Press, Cambridge, pp. 187–249.

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

Schegloff, E.A., 2007. Sequence organization in interaction: Volume 1: A primer in conversation analysis. Cambridge University Press.

Shatz, M., 1979. How to do things by asking: Form-function pairings in mothers' questions and their relation to children's responses. Child Development 50, 1093–1099.

Shi, R., Melancon, A., 2010. Syntactic categorization in French-learning infants. Infancy 15, 517–533.

Snow, C.E., 1977. The development of conversation between mothers and babies. Journal of Child Language 4, 1–22.

Soderstrom, M., Seidl, A., Kemler Nelson, D.G., Jusczyk, P.W., 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. Journal of Memory and Language 49, 249–267.

Speer, S.R., Ito, K., 2009. Prosody in first language acquisition–Acquiring intonation as a tool to organize information in conversation. Language and Linguistics Compass 3, 90–110.

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

Stivers, T., Rossano, F., 2010. Mobilizing response. Research on Language and Social Interaction 43, 3–31.

Takahashi, D.Y., Narayanan, D.Z., Ghazanfar, A.A., 2013. Coupled oscillator dynamics of vocal turn-taking in monkeys. Current Biology 23, 2162–2168.

Thorgrímsson, G., Fawcett, C., Liszkowski, U., under review .

Tice (Casillas), M., Henetz, T., 2011. Turn-boundary projection: Looking ahead, in: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.

Toda, S., Fogel, A., 1993. Infant response to the still-face situation at 3 and 6 months. Developmental Psychology 29, 532.

Torreira, F., Bögels, S., Levinson, S.C., under review. Intonational phrasing is necessary for turn-taking in spoken interaction .

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006. Elan: a professional framework for multimodality research, in: Proceedings of LREC.