Journal of Memory and Language 92 (2017) 234–253

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

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, The Netherlands
[b] Department of Psychology, Stanford University, United States

ARTICLE INFO

Article history:
Received 3 October 2014
revision received 27 June 2016

Keywords:
Turn taking
Conversation
Development
Questions
Eye-tracking
Anticipation

ABSTRACT

Children begin developing turn-taking skills in infancy but take several years to fluidly integrate their growing knowledge of language into their turn-taking behavior. In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling linguistic cues to turn structure. In Experiment 1, we showed English and non-English conversations to English-speaking adults and children. In Experiment 2, we phonetically controlled lexicosyntactic and prosodic cues in English-only speech. Children spontaneously made anticipatory gaze switches by age two and continued improving through age six. In both experiments, children and adults made more anticipatory switches after hearing questions. Consistent with prior findings on adult turn prediction, prosodic information alone did not increase children's anticipatory gaze shifts. But, unlike prior work with adults, lexical information alone was not sufficient either—children's performance was best overall with lexicosyntax and prosody together. Our findings support an account in which turn tracking and turn prediction emerge in infancy and then gradually become integrated with children's online linguistic processing.

Introduction

Spontaneous conversation is a universal context for using and learning language. Like other types of human interaction, it is organized at its core 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, Torreira, & Levinson, 2015; Dingemanse, Torreira, & Enfield, 2013; Stivers et al., 2009).

Children participate in sequential coordination (proto-turn taking) with their caregivers starting at three months of age—before they can rely on any linguistic cues (see, among others, Bateson, 1975; Hilbrink, Gattis, & Levinson, 2015; Jaffe et al., 2001; Snow, 1977). However, infant turn taking is different from adult turn taking in several ways: it is heavily scaffolded by caregivers, has different inter-turn timing, and lacks semantic content (Hilbrink et al., 2015; Jaffe et al., 2001). But children's early, turn-structured social interactions are presumably a critical precursor to their later conversational turn taking, establishing the protocol by which children come to use language with others. How then do children integrate linguistic knowledge with these preverbal turn-taking abilities?

In this study, we investigate when children begin to make predictions about upcoming turn structure in

* Corresponding author at: Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.

E-mail addresses: marisa.casillas@mpi.nl (M. Casillas), mcfrank@stanford.edu (M.C. Frank).

http://dx.doi.org/10.1016/j.jml.2016.06.013
0749-596X/© 2016 Elsevier Inc. All rights reserved.

conversation and how online linguistic processing becomes integrated into their predictions as they grow older. 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 presenting the details of the present study.

Adult 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 & Bruner, 1978; Ross & 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, Narayanan, & Ghazanfar, 2013). In all these cases, turn taking serves as a protocol for interaction, allowing the participants to coordinate with each other through sequences of contingent action.

Conversational turn taking distinguishes itself from other turn-taking behaviors by the complexity of the sequencing involved. 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?"–"Fi ne, and you?"–"Great!"). Compared to other turn-taking behaviors, the possible sequence and action types in everyday talk are diverse and unpredictable.

Despite this complexity, conversational turn taking is precise in its timing. Across a diverse sample of conversations in 10 languages, one study found a consistent average inter-turn silence of 0–200 ms at points of speaker switch (Stivers et al., 2009). Experimental results and current models of speech production suggest that it takes approximately 600 ms to produce a content word, and even longer to produce a simple utterance (Griffin & Bock, 2000; Levelt, 1989). In order to achieve 200 ms turn transitions, speakers must begin formulating their response before the prior turn has ended (Levinson, 2013; Levinson, 2016). 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.

Adults have a lot of information at their disposal to help make accurate predictions. 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, Mitterer, & Enfield, 2006; Duncan, 1972; Ford & Thompson, 1996; Bögels & Torreira, 2015). 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, Brown, & Levinson, 2009; Stivers & Rossano, 2010). Additionally, the sequential context of a turn can make the next action obvious: answers after questions, thanks or denial after compliments, etc. (Schegloff, 2007).

Prior work suggests that adult listeners primarily use lexicosyntactic information to accurately predict upcoming turn structure. De Ruiter et al. (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 his or 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. 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 responded significantly earlier within the turn. 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 & De Ruiter, 2012; see also Magyari, Bastiaansen, De Ruiter, & Levinson, 2014). They can also identify speech acts within the first word of an utterance (Gísladóttir, Chwilla, & Levinson, 2015), allowing them to start planning their response at the first moment possible (Bögels, Magyari, & Levinson, 2015).

Despite this body of evidence, the role of prosody for adult turn prediction is still a matter of debate. De Ruiter et al.'s (2006) experiment focused on the role of intonation, which is only a partial index of prosody. Prosody is tied closely to the syntax of an utterance, so the two linguistic signals are difficult to control independently (Ford & Thompson, 1996). Bögels and Torreira (2015) 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 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 within multi-utterance turns, participants were tricked into pressing the "turn-end" button 29% of the time. These findings suggest that listeners actually do rely on prosodic cues to execute a response, and that their use of prosodic cues interacts with their predictions about the unfolding syntactic structure (see also De Ruiter et al., 2006, 525). These experimental findings corroborate other corpus

[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.

and experimental work promoting a combination of cues (lexicosyntactic, prosodic, and pragmatic) as key for accurate turn-end prediction (Duncan, 1972; Ford & Thompson, 1996; Hirvenkari et al., 2013).

Turn taking in development

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 after birth, in proto-conversation with their caregivers (Bruner, 1975; Bruner, 1985; Snow, 1977). During proto-conversations, caregivers treat their infants as 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; Toda & Fogel, 1993).

The timing of children's responses to their caregivers' speech shows a non-linear pattern. Infants' contingent vocalizations in the first few months of life show very fast timing (though with a lot of vocal overlap). But by nine months, their timing slows down considerably, only to gradually speed up again after 12 months (Hilbrink et al., 2015). For children, taking turns with brief transitions between speakers is more difficult than avoiding speaker overlap; children's incidence of overlap is nearly adult-like by nine months, but the timing of their non-overlapped (i.e., gapped) responses remains longer than the adult 200 ms standard for the next few years (Casillas, Bobb, & Clark, 2016; Garvey, 1984; Garvey & Berninger, 1981; Ervin-Tripp, 1979). This puzzling pattern is likely due to children's linguistic development: taking turns on time is easier when their response is a simple vocalization rather than a linguistic utterance. Integrating language into the turn-taking system may therefore be a major factor in children's delayed responses (Casillas et al., 2016).

Before children manage to fully integrate linguistic processing into their turn-taking behaviors (for both turn prediction and production), they can rely on non-verbal interactional cues, including silence, eye gaze, body orientation, and gesture, to identify the boundaries of social actions. For example, with little to no linguistic knowledge, children are often able to infer desired responses to offers and requests by taking account of their interlocutor's non-verbal communicative behavior, the structure of routine events, and the affordances of the current interactional context (Nomikou & Rohlfing, 2011; Reddy, Markova, & Wallot, 2013; Shatz, 1978). With respect to turn taking in particular, children's spontaneous vocalizations during interaction demonstrate a sensitivity to short inter-speaker gaps from infancy (Hilbrink et al., 2015). Thus, before children can anticipate turn structure by integrating linguistic cues from unfolding speech, they might react to silence as a cue to upcoming speaker change. Interactional silence itself may then serve as one of children's first cues to turn structure, giving them information about when to respond before they can rely on language.

As children's language competence and speed of processing increases (Kail, 1991), they become better equipped to use linguistic cues in making predictions about upcoming turn structure. Studies of early linguistic development point to a possible early advantage for prosody over lexicosyntax in children's turn-taking predictions. Infants can distinguish their native language's rhythm type from others soon after birth (Mehler et al., 1988; Nazzi & Ramus, 2003). They also 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 & Jusczyk, 2001; Morgan & Saffran, 1995). Four- to five-month-olds also prefer pauses in speech to be inserted at prosodic boundaries, and by 6 months infants can use prosodic markers to pick out sub-clausal syntactic units, both of which are useful for extracting turn structure from ongoing speech (Jusczyk, Hohne, Mandel, & Strange, 1995; Soderstrom, Seidl, Kemler Nelson, & Jusczyk, 2003). In comparison, children show at best a very limited lexical inventory before their first birthday (Bergelson & Swingley, 2013; Shi & Melancon, 2010).

Keitel, Prinz, Friederici, Hofsten, and Daum (2013) were one of the first to explore how children use linguistic cues to predict upcoming turn structure. They asked 6-, 12-, 24-, and 36-month-old infants, and adult participants to watch short videos of conversation and tracked their eye movements at points of speaker change. They showed their participants two types of videos—one normal and one with flattened pitch—to test the role of intonation in participants' anticipatory predictions about upcoming speech. Comparing children's anticipatory gaze frequency to a random baseline, they found that only 36-month-olds and adults made anticipatory gaze switches more often than expected by chance, and that only 36-month-olds were affected by flattened intonation contours. This finding led Keitel and colleagues to conclude that children's ability to predict upcoming turn structure relies on their ability to comprehend the stimuli lexicosemantically. They also suggest that intonation might play a secondary role in turn prediction, but only after children acquire more sophisticated, adult-like language comprehension skills (also see Keitel & Daum, 2015).

Although the Keitel et al. (2013) study constitutes a substantial advance over previous work in this domain, it has some 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 not random. Instead, they used gaze switches during ongoing speech as a baseline. But ongoing speech is the period in which switching is least likely to occur (Hirvenkari et al., 2013)—their baseline thus maximizes the chance of finding a difference in gaze frequency at turn transitions compared to the baseline. A more conservative baseline would compare participants' looking behavior at turn transitions to their looking behavior during randomly selected windows of time throughout the stimulus, including turn transitions. We follow this conservative approach in the current study.

Second, the conversation stimuli Keitel et al. (2013) used were somewhat unusual. The average gap between turns was 900 ms, a duration much longer than typical adult timing, which averages around 200 ms (Stivers et al., 2009). The speakers in the videos were also asked to minimize their movements while performing scripted, adult-directed conversation, which would have created a somewhat unnatural interaction. Additionally, to produce more naturalistic conversation, it would have been ideal to localize the sound 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 difficult to distinguish the two talkers. Again, we attempt to address these issues in our current study. Despite these minor methodological drawbacks, the Keitel et al. (2013) study still demonstrates interesting age-based differences in children's predictions about upcoming turn structure. Our current work takes these findings as a starting point.[2]

The current study

Our goal in the current study is to find out when children begin to make predictions about upcoming turn structure and to understand how their predictions are affected by linguistic cues to turn taking across development. We present two experiments in which we measured children's anticipatory gaze to responders while they watched conversation videos with natural (people speaking English vs. non-English; Experiment 1) and non-natural (puppets with phonetically manipulated speech; Experiment 2) control over the presence of lexical and prosodic cues. We tested children across a wide range of ages (Experiment 1: 3–5 years; Experiment 2: 1–6 years), with adult control participants in each experiment. We additionally tested for the use of one non-verbal cue: inter-turn silence.

We highlight four primary findings: first, although children and adults use linguistic cues to make predictions about upcoming turn structure, they do so primarily to predict speaker transitions after questions (a "speech act" effect). This intriguing effect, which has not been reported previously, suggests that participants track unfolding speech for cues to upcoming speaker change, which may affect how they use linguistic cues more generally for anticipatory processing in conversation. Second, we find that children make more predictions than expected by chance starting at age two, but that this effect is small at first, and continues to improve through age six, along with children's use of linguistic cues to anticipate answers after question turns. Third, children and adults often used inter-turn silence (a non-verbal cue to turn structure) to make more predictive gaze switches to the responder, suggesting that non-verbal cues are useful for predicting turn structure early on and continue to be important in adulthood. Finally, we find no evidence for an early prosodic advantage in children's anticipations and, further, no evidence that lexical cues alone are comparable to the full linguistic signal in aiding children's predictions (as is proposed for adults; De Ruiter et al., 2006). Anticipation is strongest for stimuli with the full range of linguistic cues. Our findings support an account in which turn prediction emerges in infancy and becomes integrated with online linguistic processing gradually, possibly because of children's increased linguistic knowledge and speed of processing with development.

Experiment 1

We recorded participants' eye movements as they watched six short videos of two-person (dyadic) conversation that were interspersed with attention-getting filler videos. Each conversation video featured an improvised discourse in one of five languages (English, German, Hebrew, Japanese, and Korean). Participants saw two videos in English and one in every other language. 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-English prosody, gaze, in-breaths, phrase final lengthening). Using this method, we analyzed 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.

Methods

Participants

We recruited 74 children ages 3;0–5;11 and 11 undergraduate adults to participate in the experiment. We recruited adult participants through the Stanford University Psychology participant database. Adult participants were either paid or received course credit for their time. Our child sample included 19 three-year-olds, 32 four-year-olds, and 23 five-year-olds, all enrolled in a local nursery school and all of whom volunteered their time. 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 the analyses.[3] None of the adult participants reported fluency in a second language.

Materials

We recorded pairs of talkers while they conversed in a sound-attenuated booth (Fig. 1). Each talker was a native

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

[3] Multilingual children may make predictions about upcoming turn structure differently from their monolingual peers due to their more varied experiences with linguistic cues to turn taking. We are unable to test this hypothesis here due to the variability in multilingual language input and the diverse set of languages being learned in our sample. The same applies to Experiment 2.

Fig. 1. Example frame from a conversation video used in Experiment 1.

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 and were acquainted with their recording partner ahead of time.

Each recording session began with a 20-min 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-friendly speech toward each other. We recorded until the talkers achieved at least 30 s of uninterrupted discourse with enthusiastic, child-friendly 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.[4]

After recording, we combined the audio and video recordings by hand, and cropped each one to the (approximate) 30-s 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 condition (English vs. non-English), inter-turn gap duration (in milliseconds), and transition type (question vs. non-question). Each non-English turn was coded as a question or non-question from a monolingual English-speaker's perspective, i.e., turns that "sound like" questions and turns that do not. We asked five native American English speakers to listen to the audio recording for each non-English turn and judge whether it sounded like a question. We marked non-English turns as questions when at least 4 of the 5 listeners (80%) said that the turn "sounded like a question". Thus, "question" cues in the non-English condition only *resembled* native English question cues, and were therefore likely harder to identify than cues to questionhood in the English condition. However, since participants did not speak the non-English languages and would only ever treat "question-sounding" turns as questions, we proceeded with these analyses to see how pervasive question effects were—could they show up even without lexical access? If participants primarily rely on prosodic cues to question turns, it's possible that even non-English prosody can elicit anticipatory gaze switches for question-like turns.

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 ms and shorter than 90 ms from analysis, additionally excluding overlapped transitions.[5] This left approximately equal numbers of turn transitions available for analysis in the English ($N = 20$) and non-English ($N = 16$) videos. On average, the inter-turn gaps for English videos (mean = 318, median = 302, stdev = 112 ms) were slightly longer than for non-English videos (mean = 286, median = 251, stdev = 122 ms).

Questions made up exactly half of the turn transitions in the English ($N = 10$) and non-English ($N = 8$) videos. In the English videos, inter-turn gaps were slightly shorter for questions (mean = 310, median = 293, stdev = 112 ms) than non-questions (mean = 325, median = 315, stdev = 118 ms). Non-English videos did not show a large difference in transition time for questions (mean = 270, median = 257, stdev = 116 ms) and non-questions (mean = 302, median = 252, stdev = 134 ms).

Procedure

Participants sat in front of an SMI 120 Hz corneal reflection eye-tracker mounted beneath a large flatscreen display. The display and eye-tracker were secured to a table with an ergonomic arm that allowed the experimenter to position the whole apparatus at a comfortable height and 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

[4] All of the non-English talkers were fluent in English as a second language, and some fluently spoke three or more languages. We chose male–female pairs as a natural way of creating contrast between the two talker voices.

[5] 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 began earlier than expected. Participants expect conversations to proceed with "one speaker at a time" (Sacks, Schegloff, & Jefferson, 1974). They would therefore still be fixated on the prior speaker when the overlap started, and would have to switch their gaze *reactively* to the responder.

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. Each conversation video was preceded and followed by a 15–30 s attention-getting filler video (e.g., running puppies, singing muppets, flying bugs). If the child began to look bored, the experimenter would talk during the fillers, 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-s conversation videos. 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 min.

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 (Hirvenkari et al., 2013; Keitel et al., 2013; Tice (Casillas) 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 participants 200 ms to plan an eye movement, following standards from adult anticipatory processing studies (e.g., Kamide, Altmann, & Haywood, 2003).

We checked each participant's gaze at each turn transition for three characteristics (Fig. 2): (1) that the participant fixated on the prior speaker for at least 100 ms at the end of the prior turn, (2) that immediately 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 200 ms 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 the assumption that it takes at least 200 ms to plan an eye movement, gaze shifts initiated within the first 200 ms of the response (or earlier) were planned *before* participants 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 anticipated turn ends approximately 200 ms in advance of the turn's end, and anticipatory responses to pitch-flattened stimuli came even earlier (De Ruiter et al., 2006). We therefore allowed switches to occur as early as 200 ms before the end of the prior turn. Again, because it

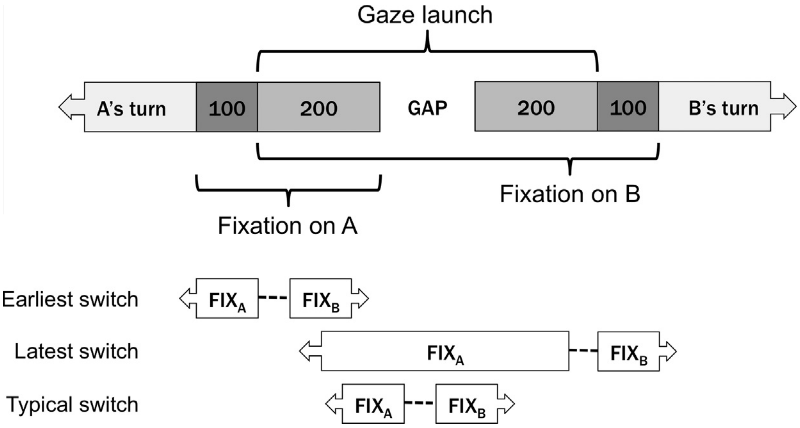

Fig. 2. Schematic summary of the criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition. FIX = hypothetical fixation on speaker A or speaker B; dashed lines = hypothetical saccadic time.

takes 200 ms to plan an eye movement, we counted anticipatory switches, at the latest, 200 ms after the onset of speech. Therefore, for very early and very late switches, our requirement of 100 ms of fixation on each speaker would sometimes extend outside of the gaze launch window boundaries (200 ms before and after the inter-turn gap; dark gray boxes Fig. 2). The maximally available fixation window was therefore 100 ms before and after the earliest and latest possible switch point (300 ms 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 (300 ms + the inter-turn gap + 300 ms).

Predictions

We expected participants to show greater anticipation in the English videos than in the non-English videos because of their increased access to linguistic information in English. We also predicted that anticipation would be greater following questions compared to 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. Our third prediction was that anticipatory looks would increase with development, along with children's increased linguistic competence and speed of processing. Finally, we predicted that transitions with longer inter-turn gaps would show greater anticipation because longer gaps provide (a) more time to make a gaze switch and (b) are themselves a cue to possible upcoming speaker switch.

Results

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 or not the videos were in a language they could understand. 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 demonstrated that they performed basic turn tracking during the videos, regardless of language. Fig. 3 shows participants' anticipatory gaze rates across age, language condition, and transition type.

Table 1
Average proportion of gaze to the current speaker and addressee during periods of talk across ages in Experiment 1.

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

Statistical models

We identified anticipatory gaze switches for all 36 usable turn transitions, based on the criteria outlined above, and analyzed them for effects of language, transition type, and age with two mixed-effects logistic regressions (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). We built one model each for children and adults. We modeled children and adults separately because effects of age are only pertinent to the children's data.

The child model included condition (English vs. non-English),[6] transition type (question vs. non-question), age (3, 4, 5; numeric; intercept as age = 0), and duration of the inter-turn gap (seconds, e.g., 0.441) as predictors, with two-way interactions between gap duration and the other simple fixed effects (language condition, transition type, and age) and a three-way interaction between language condition, transition type, and age. We included the two-way interactions with gap duration in case the effect of inter-turn silence changes with age or linguistic cueing (e.g., if children older children rely less on silence as a cue).[7] We also included random effects of item (turn transition) and participant, with maximal random slopes of condition, transition type, and their interaction for participants (Barr, Levy, Scheepers, & Tily, 2013).[8]

The adult model included fixed effects of condition, transition type, and their interaction, plus two-way interactions between gap duration and the other simple fixed effects (language condition and transition type, as in the child model). The adult model also included random effects of item and participant with maximal random slopes of condition, transition type, and their interaction for participant.

[6] Because each non-English language was represented by a single stimulus, we cannot treat individual languages as factors. Gaze behavior might be best for non-native 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, Hirschberg, & Swerts, 2005). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

[7] We test these two-way interactions with gap duration in all of the models reported in this paper. Higher-order interactions with gap duration usually resulted in model non-convergence due to distributional sparsity when three or more predictor values were considered, so we did not include them.

[8] The models we report in this paper are all qualitatively unchanged by the exclusion of their random slopes. We have left the random slopes in because of minor participant-level variation in the predictors modeled.

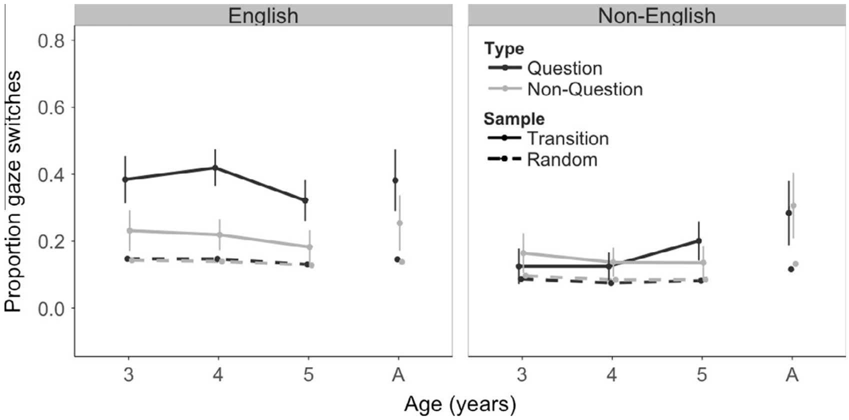

Fig. 3. Anticipatory gaze rates across language condition and transition type for the real and randomly permuted datasets. Vertical bars represent 95% confidence intervals.

Children's anticipatory gaze switches showed effects of language condition ($\beta = -3.65$, $SE = 1.16$, $z = -3.15$, $p < .01$) and transition type ($\beta = -2.95$, $SE = 1.13$, $z = -2.61$, $p < .01$) with additional effects of an age-by-language condition interaction ($\beta = 0.5$, $SE = 0.212$, $z = 2.35$, $p < .05$), a language condition-by-transition type interaction ($\beta = 2.69$, $SE = 1.35$, $z = 1.99$, $p < .05$), and a transition type-gap duration interaction ($\beta = 5.52$, $SE = 2.28$, $z = 2.42$, $p < .05$). There were no significant effects of age or gap duration alone ($\beta = -0.002$, $SE = 0.26$, $z = -0.009$, $p = .99$ and $\beta = 2.25$, $SE = 3.19$, $z = 0.7$, $p = .48$, respectively).

Adults' anticipatory gaze switches showed an effect of transition type ($\beta = -3.3$, $SE = 0.93$, $z = -3.54$, $p < .001$) and significant interactions between language condition and transition type ($\beta = 1.23$, $SE = 0.63$, $z = 1.96$, $p < .05$) and transition type and gap duration ($\beta = 7.12$, $SE = 2.2$, $z = 3.24$, $p < .01$). There were no significant effects of language condition or gap duration alone ($\beta = -0.06$, $SE = 0.75$, $z = -0.08$, $p = .94$ and $\beta = 0.13$, $SE = 1.77$, $z = 0.08$, $p = .94$, respectively).

Random baseline comparison

Our primary analysis (above) makes the assumption that participants' eye movements generally follow the turn structure of the stimulus, i.e., that participants track the current speaker and switch their gaze to the upcoming speaker near turn transitions. As just described, based on this assumption, we used linear mixed effects regressions to see how anticipatory looking is affected by aspects of participant group (e.g., age) and stimulus (e.g., transition type, language condition). But what if the assumption that participants generally track turn structure were wrong? Could these results have emerged if participants' eye movements were *not* linked to turn structure? For example, if participants were randomly looking back and forth between the two speakers, we might still find some anticipatory switching by chance. To test whether our primary results (the regression output above) could have arisen from random switching we conducted a secondary analysis comparing participants' anticipatory gaze at real and randomly shuffled points of turn transition.

We conducted this analysis by running the same regression models on participants' eye-tracking data, only this time calculating their anticipatory gaze switches with respect to randomly permuted turn transition windows. This process involved: (1) randomizing the order and temporal placement of the analysis windows within each stimulus (Fig. 4; "analysis window" is as shown in Fig. 2) to randomly redistribute the analysis windows across the eye-tracking signal, (2) re-running each participant's eye tracking data through switch identification (described above) on each of the randomly permuted analysis windows, and (3) modeling the anticipatory switches from the randomly permuted data (our random baseline dataset) with the same statistical models we used for the original dataset (Table 2). Importantly, although the onset time of each transition was shuffled within the eye-tracking signal, the other intrinsic properties of each turn transition (e.g., prior speaker identity, transition type, gap duration, language condition, etc.) stayed constant across each permutation.

The random shuffling procedure de-links participants' gaze data from the turn structure in the original stimulus, thereby allowing us to compare turn-related (original) and non-turn-related (randomly permuted) looking behavior using the same eye movement data. We created 5000 permutations of the original turn transitions, thereby creating 5000 anticipatory gaze datasets with randomly de-linked gaze data. Because the randomly shuffled turn transitions could occur anywhere in the stimulus (so long as they didn't overlap each other within a single iteration), the resulting turn-transition windows collectively covered the entire stimulus—during speech and silence, during speaker change and speaker continuation, and during all turn transitions in the stimulus, even those excluded in the original analyses (e.g., because they were overlapped). This technique crucially differs from that used by Keitel et al. (2013) and Keitel and Daum (2015), which tests anticipatory gaze at turn transitions against anticipatory gaze during speech. Pooled together, our 5000 anticipatory gaze datasets yielded an average anticipatory switch rate for each participant over all possible starting points in

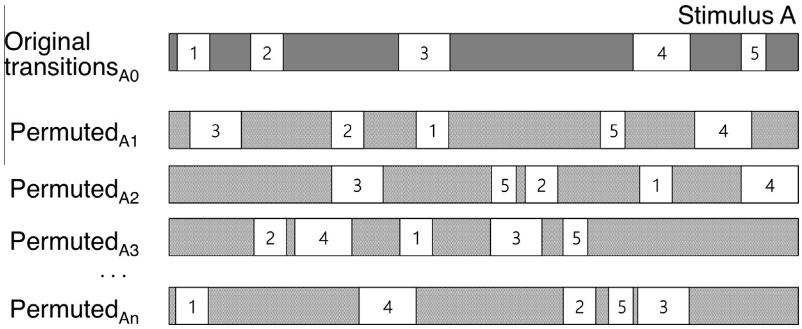

Fig. 4. Example of analysis window permutations for a stimulus with five turn transitions. The windows included ±300 ms around the inter-turn gap.

Table 2
Model output for participants' anticipatory gaze switches in Experiment 1.

	Estimate	Std. Error	z value	Pr(>|z|)
Children				
(Intercept)	−0.604	1.242	−0.486	0.627
Age	−0.002	0.261	−0.009	0.993
LgCond = *non-English*	−3.65	1.16	−3.146	0.002**
TType = *non-Question*	−2.95	1.13	−2.61	0.009**
GapDuration	2.247	3.194	0.704	0.482
Age ∗ LgCond = *non-English*	0.5	0.212	2.353	0.019*
Age ∗ TType = *non-Question*	0.009	0.196	0.044	0.965
LgCond = *non-English* ∗ TType = *non-Question*	2.692	1.347	1.999	0.046*
Age ∗ GapDuration	−0.577	0.627	−0.921	0.357
LgCond = *non-English* ∗ GapDuration	1.143	2.287	0.5	0.617
TType = *non-Question* ∗ GapDuration	5.519	2.282	2.418	0.016*
Age ∗ LgCond = *non-English* ∗ TType = *non-Question*	−0.433	0.304	−1.426	0.154
Adults				
(Intercept)	−0.584	0.64	−0.913	0.361
LgCond = *non-English*	−0.059	0.751	−0.079	0.937
TType = *non-Question*	−3.298	0.933	−3.536	0.0004***
GapDuration	0.132	1.766	0.075	0.941
LgCond = *non-English* ∗ TType = *non-Question*	1.234	0.629	1.961	0.0498*
LgCond = *non-English* ∗ GapDuration	−1.519	2.192	−0.693	0.488
TType = *non-Question* ∗ GapDuration	7.116	2.195	3.241	0.001**

the stimuli: a random baseline. Using this technique we compared participants' anticipatory switches at turn transition windows to their anticipatory switches over the stimulus as a whole. If participants looked randomly back and forth between the speakers, we would have seen similar patterns in both cases.

Rather than simply comparing participants' overall anticipatory switch rates with real and random transition windows, we estimated the likelihood that each of the predictor effects in the original data (e.g., the effect of language condition; Table 2) could have arisen with random gaze switching: we ran identical statistical models on the real and randomly permuted data sets. This tells us not only whether participants' switches were above chance, but whether the specific underlying effects of their anticipatory gaze patterns (e.g., the effect of language condition) were above that expected by chance. Because these analyses are complex and secondary to the main results, we report their full details in Supplementary material A.

Our baseline analyses revealed that none of the significant predictors from models of the original, turn-related data can be explained by random looking. For the children's data, the original z-values for language condition, transition type, the age-language condition interaction, the transition type-gap duration interaction, and the language condition-transition type interaction were all greater than 95% of z-values from models of the randomly permuted data (99.3%, 99.1%, 98.9%, 97%, and 96%, respectively, all $p < .05$). Similarly, the adults' data showed significant differentiation from the randomly permuted data for all three significant predictors from the real transition dataset. Transition type, the interaction between transition type and gap duration, and the interaction between language condition and transition type showed z-values that exceeded 100%, 99.8%, and 95% of random z-values, respectively (all $p \leqslant .05$). See Supplementary material A for more information on each predictor's random permutation distribution.[9]

[9] This baseline analysis tests "random looking" against "turn-driven looking", but it does not test subtypes of turn-driven looking. For example, children might switch their gaze from the current speaker to the addressee out of boredom with the ongoing speech rather than from active anticipation of an upcoming response. We address this hypothesis about "boredom" gaze switches vs. "turn-transition" gaze switches in Supplementary material C.

Developmental effects

The models reported above revealed a significant interaction of age and language condition (Table 2) that was unlikely be due to random gaze switching (Fig. 3). To further explore this effect, we compared the effect of language condition across age groups: using the permuted datasets described above, we extracted the average difference score for the two language conditions (English minus non-English) for each participant, computing an overall average for each random permutation of the data. Then, within each permutation, we made pairwise comparisons of the average difference scores across participant age groups. This process yielded a distribution of random permutation-based difference scores that we could then compare to the difference score in the actual data. Details are given in Supplementary material B.

These analyses revealed that, while 3- and 4-year olds showed similarly-sized effects of language condition, 5-year-olds had a significantly smaller effect of language condition, compared to both younger age groups. The difference in the language condition effect between 5-year-olds and 3-year-olds was greater than would be expected by chance (99.52% of the randomly permuted data sets; $p < .01$). Similarly, the difference in the language condition effect between 5-year-olds and 4-year-olds was greater than would be expected by chance (99.96% of the data sets; $p < .001$). See Supplementary material Fig. B.1 for each difference score distribution.

When does spontaneous turn prediction emerge developmentally? We tested whether the youngest age group (3-year-olds) already exceeded chance in their anticipatory gaze switches by comparing children's real gaze rates to the random baseline in the English condition with two-tailed *t*-tests. We used the English condition because we are most interested in finding out when children begin to make spontaneous turn predictions for natural speech. We found that three-year-olds made anticipatory gaze switches significantly above chance, when all transitions were considered ($t(22.824) = -4.147$, $p < .001$) as well as for question transitions alone ($t(21.677) = -5.268$, $p < .001$).

Discussion

Children and adults spontaneously tracked the turn structure of the conversations, making anticipatory gaze switches at an above-chance rate across all ages and conditions. Children's anticipatory gaze rates were affected by language condition, transition type, age, and gap duration (Table 2), none of which could be explained by a baseline of random gaze switching (see Supplementary material Fig. A.1). These data show a number of important features that bear on our questions of interest.

First, both adults' and children's anticipations were strongly affected by transition type. Both groups made more anticipatory switches after hearing questions, compared to non-questions, especially for the English stimuli compared to the non-English stimuli. Overall, participants made few anticipatory switches after non-questions, even in the English videos when they had full linguistic access. Prior work using online, metalinguistic tasks has shown that participants can use linguistic cues to accurately predict upcoming turn ends (Bögels & Torreira, 2015; Magyari & De Ruiter, 2012; De Ruiter et al., 2006). The current results add a new dimension to our understanding of how listeners make predictions about turn ends: both children and adults spontaneously monitor the linguistic structure of unfolding turns for cues to imminent responses.

Second, children made more anticipatory switches overall in English videos, compared to non-English videos. This effect suggests that linguistic access is important for children's ability to anticipate upcoming turn structure, consistent with prior work on turn-end prediction in adults (De Ruiter et al., 2006; Magyari & De Ruiter, 2012) and children (Keitel et al., 2013).

Third, we saw that older children made anticipatory switches more reliably than younger children, but only in the non-English videos. In the English videos, children anticipated well at all ages, especially after hearing questions. This interaction between age and language condition suggests that the 5-year-olds were able to leverage anticipatory cues in the non-English videos in a way that 3- and 4-year-olds could not, possibly by shifting more attention to the non-English prosodic or non-verbal cues. Prior work on children's turn-structure anticipation has proposed that children's turn-end predictions rely primarily on lexicosemantic structure (and not, e.g., prosody) as they get older (Keitel et al., 2013). The current results suggest more flexibility in children's predictions; when they do not have access to lexical information, older children and adults find alternative cues to turn taking behavior.

Finally, children and adults made more anticipatory switches in transitions with longer inter-turn gaps, though this effect was limited to non-question turns (Table 2). This finding suggests that gap duration indeed serves as a cue to upcoming turn structure; while short gaps may be perceived as within-turn pauses (Männel & Friederici, 2009), long gaps could instead be indicative of between-turn pauses (where speaker transition occurs). Participants might use long silences to retroactively assign turn boundaries and anticipate speaker switches that were otherwise not anticipated (in this case, because the preceding turn was not a question). An alternative explanation for effects of gap duration is that longer inter-turn gaps result in longer analysis windows, which gives participants more time to make an anticipatory gaze. However, if participants are generally more likely to make a switch at question transitions (as our results suggest), and if question-driven switches aren't already at ceiling when gaps are short, we would expect that longer gaps would benefit questions more than non-questions—the opposite pattern from what the data show here. We take this as evidence that inter-turn silence may be most useful when participants have limited ability to make predictions about upcoming speaker transitions.

In Experiment 2, we followed up on these findings, improving on two aspects of the design: first, our language manipulation in this first experiment was too coarse to provide data regarding specific linguistic information channels (e.g., the effect of prosodic information alone). In Experiment 2, we compared lexicosyntactic and prosodic cues with phonetically altered speech and used puppets to eliminate non-verbal cues to turn taking. Second, we were not able to pinpoint the emergence of anticipatory switching because the youngest age group in our sample was already able to make anticipatory switches at above-chance rates. In Experiment 2, we explored a wider developmental range.

Experiment 2

Experiment 2 used English-only stimuli, controlled for lexical and prosodic information, eliminated non-verbal cues, and tested children from a wider age range. 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 1- to 6-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 gestural cues in Experiment 1 could have helped participants make predictions about upcoming turn structure (Rossano et al., 2009; Stivers & Rossano, 2010). Since our focus here 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. Additionally, the prosody-controlled condition (described below) 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 use puppet videos rather than 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.

Methods

Participants

We recruited 27 undergraduate adults and 129 children ages 1;0–6;11 to participate in our experiment. Adult participants were recruited again via the Stanford University Psychology participant database and were either paid or received course credit for their time. We recruited our child participants from the Children's Discovery Museum in San Jose, California,[10] targeting approximately 20 children for each of the six one-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.

Materials

We created 18 short videos of improvised, child-friendly conversation (Fig. 5). To eliminate non-verbal cues to turn transition and to control the types of linguistic information available in the stimuli we first audio-recorded improvised conversations, then phonetically manipulated those recordings to limit the availability of prosodic and lexical information, and finally recorded video to accompany the manipulated audio, featuring puppets as talkers.

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 min) and a pre-determined topic for the child-friendly improvisation ('riding bikes', 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the 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 (approximate) 20-s interval with the most turn activity. The 20-s clips were then phonetically manipulated and used in the final video stimuli.

Audio manipulation. We created four versions of each audio conversation: *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 "partial cue" conditions). The *normal* conversations were the unmanipulated, original audio clips.

The *words only* conversations were manipulated to have robot-like speech: we flattened the intonation contours to each talker's average pitch (F_0) and we reset the duration of every nucleus and coda to each talker's average nucleus and coda duration.[11] We made duration and pitch manipulations using PSOLA resynthesis in Praat (Boersma & Weenink, 2012). Thus, the *words only* versions of the conversations had no pitch or durational cues to upcoming turn boundaries, but did have intact lexicosyntactic cues (and some residual phonetic correlates of prosody, e.g., intensity).

We created the *prosody only* conversations 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" effect because low-pass filtering removes most of the phonetic information used to distinguish between phonemes. The *prosody only* versions of the conversations 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 audio clip for the conversation with multi-talker babble: we overlaid multiple child-oriented conversations (excluding the original one), and then cropped the result to the duration of the original conversation clip. Thus, the *no speech* conversation lacked any linguistic information to upcoming turn boundaries—the only cue to turn taking was the opening and closing of the puppets' mouths.

[10] We ran Experiment 2 at a local children's museum because it gave us access to children with a wider range of ages. Participants were volunteers.

[11] We excluded hyper-lengthened words like [wa] 'wooooow!'.

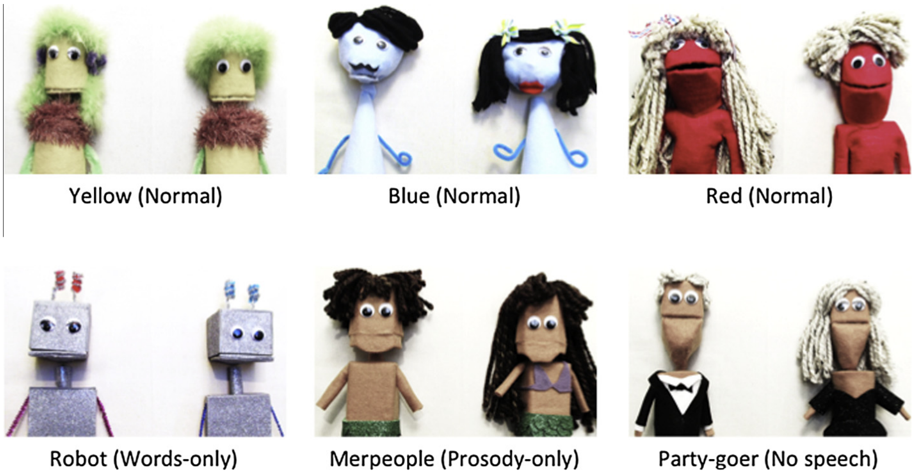

Fig. 5. The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

Finally, because low-pass filtering removes significant acoustic energy, the *prosody only* conversations were much quieter than the other three conditions. Our last step was to downscale the intensity of the audio tracks in 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-s audio clips. The puppets were minimally expressive; the puppeteer could only control the opening and closing of their mouths, and the puppets' heads, eyes, arms, and bodies stayed still. Puppets were positioned side-by-side, looking in the same direction to eliminate shared gaze as a cue to turn structure (Thorgrímsson, Fawcett, & Liszkowski, 2015). We took care to match the puppets' mouth movements to the syllable onsets as closely as possible, specifically avoiding mouth movement before the onset of a turn. We then added the manipulated audio clips to the puppet video recordings by hand with video editing software.

We used three pairs of puppets for the *normal* condition—'red', 'blue' and 'yellow'—and one pair of puppets for each partial cue condition: 'robots', 'merpeople', and 'party-goers' (Fig. 5). We randomly assigned half of the conversation topics ('birthday cake', 'pets', and 'breakfast') to the *normal* condition, and half to the partial 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; 6 videos per experiment version). 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.

As before, 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 signal 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 puppets' mouth movements in the video signal, using ELAN video annotation software (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006).

There were 85 turn transitions for analysis after excluding transitions longer than 550 ms and shorter than 90 ms. The remaining turn transitions had more questions than non-questions ($N = 47$ and $N = 38$, respectively), with transitions distributed somewhat evenly across conditions, keeping in mind that there were three *normal* videos and only one video for each partial cue condition in each experiment version: *normal* ($N = 36$), *words only* ($N = 13$), *prosody only* ($N = 17$), and *no speech* ($N = 19$). Inter-turn gaps for questions (mean = 366, median = 438, stdev = 138 ms) were longer than those for non-questions (mean = 305, median = 325, stdev = 94 ms) on average, but gap duration was overall comparable across conditions: *normal* (mean = 334, median = 321, stdev = 130 ms), *words only* (mean = 347, median = 369, stdev = 115 ms), *prosody only* (mean = 365, median = 369, stdev = 104 ms), and *no words* (mean = 319, median = 329, stdev = 136 ms).

Procedure

We used the same experimental apparatus and procedure as in the first experiment. Each participant watched six puppet videos in random order, with 15–30 s 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. As before, the experimenter immediately began each session with calibration and then stimulus presentation. Participants were given no instruction about how to watch the videos or what their purpose was, they were simply encouraged to watch the "(fun/nice) puppet videos". The entire experiment took less than five minutes.

Data preparation and coding

We coded each turn transition for its linguistic condition (*normal*, *words only*, *prosody only*, and *no speech*) and transition type (question/non-question),[12] and identified anticipatory gaze switches to the upcoming speaker using the methods from Experiment 1.

Results

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), primarily looking at the person who was currently speaking (Tables 3 and 4). 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 partial 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 3 and 4). Looks to the addressee also decreased with age, but the change was minimal. Fig. 6 shows participants' anticipatory gaze rates across age, the four language conditions, and transition type.

Table 3
Average proportion of gaze to the current speaker and addressee during periods of talk across ages in Experiment 2.

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 4
Average proportion of gaze to the current speaker and addressee during periods of talk across conditions in Experiment 2.

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

Statistical models

We identified anticipatory gaze switches for all 85 usable turn transitions, and analyzed them for effects of language condition, transition type, and age with two mixed-effects logistic regressions. We again built separate models for children and adults because effects of age were only pertinent to the children's data. The child model included condition (*normal*/*prosody only*/*words only*/*no speech*; with *no speech* as the reference level), transition type (question vs. non-question), age (1, 2, 3, 4, 5, 6; numeric, intercept as age = 0), and duration of the inter-turn gap (in seconds) as predictors, with full interactions between language condition, transition type, and age and two-way interactions between gap duration and the other basic fixed effects (age, linguistic condition, and transition type). We also included random effects of participant and item (turn transition), with maximal random slopes of transition type for participant. The adult model included condition, transition type, their interactions, gap duration, and two-way interactions between gap duration and condition and transition type, with participant and item as random effects and maximal random slopes of condition and transition type for participant.

Children's anticipatory gaze switches showed an effect of gap duration ($\beta = 3.85$, $SE = 1.73$, $z = 2.22$, $p < .05$), a two-way interaction of age and language condition (for *prosody only* speech compared to the *no speech* reference level; $\beta = 0.38$, $SE = 0.19$, $z = 1.97$, $p < .05$), a marginal two-way interaction of language condition and gap duration (for *prosody only* speech compared to the *no speech* reference level; $\beta = -4.77$, $SE = 2.63$, $z = -1.82$, $p = .07$), and a three-way interaction of age, transition type, and language condition (for *normal* speech compared to the *no speech* reference level; $\beta = -0.35$, $SE = 0.17$, $z = -2.05$, $p < .05$). There were no significant effects of age or transition type alone (Table 5; $\beta = -0.05$, $SE = 0.14$, $z = -0.38$, $p = .7$ and $\beta = -1.22$, $SE = 0.96$, $z = -1.27$, $p = .2$, respectively).

Adults' anticipatory gaze switches showed a significant effect of language condition (for *words only* speech compared to the *no speech* reference level; $\beta = 3.79$, $SE = 1.62$, $z = 2.34$, $p < .05$) and a marginal two-way interaction between language condition and transition type (for *words only* speech compared to the *no speech* reference level; $\beta = -1.68$, $SE = 0.89$, $z = -1.89$, $p = .06$). There was no significant effect of transition type alone (Table 6; $\beta = -0.02$, $SE = 1.44$, $z = -0.02$, $p = .99$).

Random baseline comparison

Using the same technique described in Experiment 1, we created and modeled random permutations of participants' anticipatory gaze switches. These analyses revealed that the significant predictors from models of the original, turn-related data were unlikely to be explained by random looking. In the children's data, the original model's z-values for gap duration, the two-way interaction of age and language condition (*prosody only*) and the three-way interaction of age, transition type, and language condition (*normal* speech) were all greater than 93% of the randomly permuted z-values (95.6%, 94%, and 93.3%, respectively, $p = .04$, .06, and .07). Similarly, the adults' data showed significant differentiation from the randomly permuted data for the effect of language condition (*words only* speech; greater than 98.3% of random z-values, $p < .02$). See Supplementary material A for more information on each predictor's random permutation distribution.

Developmental effects

Our main goal in extending the age range to 1- and 2-year-olds in Experiment 2 was to find the age of emergence for spontaneous predictions about upcoming turn struc-

[12] We coded *wh*-questions as "non-questions" for the *prosody only* videos. Polar questions often have a final rising intonational contour, but *wh*-questions do not (Hedberg, Sosa, Görgülü, & Mameni, 2010).

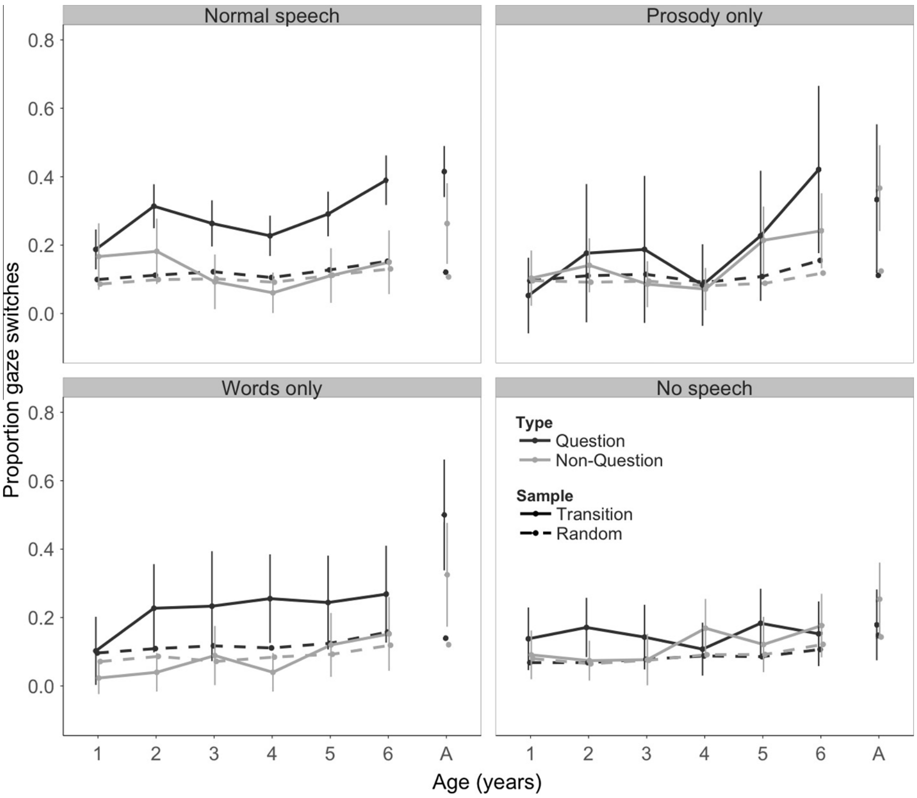

Fig. 6. Anticipatory gaze rates across language condition and transition type for the real and randomly permuted datasets. Vertical bars represent 95% confidence intervals.

Table 5
Model output for children's anticipatory gaze switches in Experiment 2.

	Estimate	Std. Error	*z* value	Pr(>|*z*|)
Children				
(Intercept)	−3.452	0.76	−4.543	5.55e−06***
Age	−0.054	0.143	−0.379	0.705
TType = *non-Question*	−1.217	0.958	−1.27	0.204
GapDuration	3.852	1.735	2.221	0.026*
Age ∗ TType = *non-Question*	0.152	0.141	1.081	0.28
Age ∗ GapDuration	0.214	0.266	0.805	0.421
TType = *non-Question* ∗ GapDuration	0.995	2.134	0.466	0.641
Condition = *normal*	0.54	0.742	0.728	0.467
Age ∗ Condition = *normal*	0.125	0.103	1.221	0.222
Condition = *normal* ∗ TType = *non-Question*	0.908	0.748	1.215	0.224
Age ∗ Condition = *normal* ∗ TType = *non-Question*	−0.355	0.173	−2.051	0.04*
Condition = *normal* ∗ GapDuration	−0.431	1.67	−0.258	0.797
Condition = *prosody*	0.549	1.452	0.378	0.705
Age ∗ Condition = *prosody*	0.375	0.191	1.967	0.049*
Condition = *prosody* ∗ TType = *non-Question*	1.076	1.105	0.974	0.33
Age ∗ Condition = *prosody* ∗ TType = *non-Question*	−0.296	0.235	−1.257	0.209
Condition = *prosody* ∗ GapDuration	−4.767	2.625	−1.816	0.069 (.)
Condition = *words*	0.684	1.06	0.645	0.519
Age ∗ Condition = *words*	0.127	0.136	0.934	0.35
Condition = *words* ∗ TType = *non-Question*	−1.244	1.031	−1.207	0.228
Age ∗ Condition = *words* ∗ TType = *non-Question*	0.111	0.225	0.495	0.621
Condition = *words* ∗ GapDuration	−2.285	2.232	−1.024	0.306

ture. As in Experiment 1, we used two-tailed *t*-tests to compare children's real gaze rates to the random baseline rates in the *normal* speech condition (in which the speech stimulus is most like what children hear every day). We tested real gaze rates against baseline rates for three age groups: one-, two-, and three-year-olds. Two- and three-year-old

Table 6
Model output for adults' anticipatory gaze switches in Experiment 2.

	Estimate	Std. Error	z value	Pr(>|z|)
Adults				
(Intercept)	−3.117	1.176	−2.649	0.008**
TType = *non-Question*	−0.022	1.44	−0.015	0.988
GapDuration	4.073	2.947	1.382	0.167
TType = *non-Question* ∗ GapDuration	1.304	3.859	0.338	0.735
Condition = *normal*	0.39	1.316	0.296	0.767
Condition = *normal* ∗ TType = *non-Question*	−0.709	0.754	−0.94	0.347
Condition = *normal* ∗ GapDuration	2.1	3.336	0.629	0.529
Condition = *prosody*	0.757	2.193	0.345	0.73
Condition = *prosody* ∗ TType = *non-Question*	0.386	1.065	0.362	0.717
Condition = *prosody* ∗ GapDuration	−1.118	4.543	−0.246	0.805
Condition = *words*	3.792	1.621	2.338	0.019*
Condition = *words* ∗ TType = *non-Question*	−1.678	0.889	−1.888	0.059 (.)
Condition = *words* ∗ GapDuration	−5.653	3.861	−1.464	0.143

children made anticipatory gaze switches significantly above chance both when all transitions were considered (2-year-olds: $t(26.193) = -4.137$, $p < .001$; 3-year-olds: $t(22.757) = -2.662$, $p < .05$) and for question transitions alone (2-year-olds: $t(25.345) = -4.269$, $p < .001$; 3-year-olds: $t(21.555) = -3.03$, $p < .01$). One-year-olds, however, only made anticipatory gaze shifts marginally above chance for turn transitions overall and for question turns alone (overall: $t(24.784) = -2.049$, $p = .051$; questions: $t(25.009) = -2.03$, $p = .053$).

We also tested the two baseline linguistic conditions against each other—*no speech* and *normal speech*—to find out when linguistic information made a difference in children's anticipations. Because, as we have seen, children primarily show linguistic effects in question–answer turn transitions, we investigated the use of linguistic cues across age by testing anticipation separately for question and non-question turns. Compared to the *no speech* condition, children made significantly more anticipatory switches in the *normal* speech condition for questions at ages 6, 4, and 3, and also marginally at age 2 (6-year-olds: $t(36.919) = 3.8019$, $p < .001$; 4-year-olds: $t(41.449) = 2.9777$, $p < .01$; 3-year-olds: $t(35.724) = 2.4286$, $p < .05$; 2-year-olds: $t(41.078) = 1.8018$, $p = .079$). Children's anticipatory switches for questions did not significantly differ in the *no speech* and *normal* speech conditions at ages 5 or 1 (5-year-olds: $t(29.406) = 1.2783$, $p = .211$; 1-year-olds: $t(35.907) = 0.4961$, $p = .623$). In contrast, children's anticipatory switch rates for non-question turns were not significantly different between the *no speech* and *normal* speech conditions at any age (all $p > .09$). Thus, consistent with the regression results, children were more likely to show an effect of linguistic content as they got older, but only for question transitions.

The regression models for the children's data also revealed two significant interactions with age. The first was a significant interaction of age and language condition (for *prosody only* compared to the *no speech* reference level), suggesting a different age effect between the two linguistic conditions. As in Experiment 1, we explored each age interaction by extracting an average difference score over participants for the effect of language condition (*no speech* vs. *prosody only*) within each random permutation of the data, making pairwise comparisons between the six age groups. These tests revealed that children's anticipation in the *prosody only* condition significantly improved at ages five and six compared to the *no speech* baseline (with difference scores greater than 95% of the random data scores; $p < .05$). See Supplementary material Fig. B.2 for these *prosody only* difference score distributions.

The second age-based interaction was a three-way interaction of age, transition type, and language condition (for *normal* speech compared to the *no speech* baseline). We again created pairwise comparisons of the average difference scores for the transition type-language condition interaction across age groups in each random permutation of the data, finding that the effect of transition type in the *normal* speech condition became larger with age, with significant improvements by age 4 over ages 1 and 2 (99.9% and 98.86%, respectively), by age 5 over age 4 (97.54%), and by age 6 over ages 1, 2, and 5 (99.5%, 97.36%, and 95.04%), all significantly different from chance ($p < .05$). See Supplementary material Fig. B.3 for these *normal* speech difference score distributions.

Discussion

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 find the age of emergence for spontaneous turn anticipation. Many of our results replicate the findings from Experiment 1: participants often made more anticipatory switches when they had access to linguistic information and, when they did, tended to make more anticipatory switches for questions compared to non-questions.

As in Experiment 1, children and adults spontaneously tracked the turn structure of the conversations. Participants made anticipatory gaze switches at above-chance rates starting at age two for both questions and non-questions. Longer gaps had a broader impact on participants' anticipations in this second experiment; we saw that, overall, longer inter-turn gaps resulted in more anticipatory switches, with the *no speech* condition showing equal or stronger effects of gap duration than all other conditions.

As before, participants made far more anticipations for questions than for non-question turns—at least for those two years old and older. But these effects were different for the conditions with partial linguistic information: *prosody only* and *words only*. In the *prosody only* condition, performance was initially low for young children and increased significantly with age. In the *words only* condition, children age two and older showed robust switching for questions (much like in *normal* speech), but never rose above chance for non-question turns (Fig. 6), with no significant differences from the *no speech* baseline. These findings do not support an early role for prosody or lexical information alone in children's spontaneous predictions about turn structure. They also give no support for the idea that lexical information is sufficient on its own to support children's anticipatory switching. They do underscore the developing relationship between the online use of linguistic cues, inter-turn silence, and speech act in spontaneous predictions about upcoming turn structure.

General discussion

Children begin to develop conversational turn-taking skills long before their first words emerge (Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001; Snow, 1977). As they become fast and knowledgeable language users, they also become able 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, when spontaneous predictions begin, they are primarily driven by participants' expectation of an immediate response in the next turn (e.g., after questions). In making predictions about upcoming turn structure, children used a combination of lexical and prosodic cues; neither signal alone was sufficient to support increased anticipatory gaze. We also found no early advantage for prosody over lexicosyntax; children's anticipatory switch rates in the *prosody only* condition were initially low, but showed significant gains by age five. We discuss these findings with respect to the role of linguistic processing and inter-turn silence for predicting upcoming turn structure, the importance of questions in predictions about conversation, and children's developing competence as conversationalists.

Predicting upcoming turn structure

Prior work with adults has found a consistent role for lexicosyntax in predicting upcoming turn structure (De Ruiter et al., 2006; Magyari & De Ruiter, 2012), whereas the role of prosody is still under debate (Duncan, 1972; Ford & Thompson, 1996; Bögels & Torreira, 2015). Knowing that children comprehend more about prosody than lexicosyntax early on (see Speer & Ito (2009) for a review), we thought it possible that young children would instead show an advantage for prosody in their predictions about turn structure in conversation. Our results suggest that, on the contrary, exclusively presenting prosodic information to children limits their spontaneous predictions about upcoming turn structure until age five.

Thus, using prosody alone to accurately predict turn boundaries in conversation appears to be difficult for adults and children. Prosodic information is continuous, multidimensional, and can index multiple meanings at once—it encodes syntactic structure, speech act, and extralinguistic information without clear one-to-one mappings between form and meaning (Cutler, Dahan, & Van Donselaar, 1997; Shriberg et al., 1998; Lammertink, Casillas, Benders, Post, & Fikkert, 2015). For these reasons, prosodic information alone may not be enough for young children to easily make precise temporal predictions about turn structure, and identify question turns in unfolding speech. Therefore, although children show early facility with prosodic discrimination (Nazzi & Ramus, 2003; Soderstrom et al., 2003; Johnson & Jusczyk, 2001; Jusczyk et al., 1995; Morgan & Saffran, 1995; Mehler et al., 1988), using prosodic knowledge for turn prediction may be difficult without additional information from lexical or syntactic cues.

Our findings suggest that there is one prosodic cue that is an exception to this rule: inter-turn silence. Generally speaking, participants showed a greater anticipatory switches for longer inter-turn gaps, but the effect of inter-turn gap duration is strongest in our data when upcoming responses are less predictable, whether due to the asymmetrical response expectations for questions vs. non-questions (Experiment 1) or the lack of non-verbal cues and any linguistic information (Experiment 2). Notably, there were no significant interactions of gap duration with participant age. This pattern of results suggests that, when predictive information about upcoming responses is absent, long silences may increase participants' expectation for a speaker change and promote more anticipatory gaze switches. Pauses are detected and related to phrasal structure from early on; 5-month-old infants use pauses to parse intonational phrases (Männel & Friederici, 2009). The lack of interactions between age and gap duration suggests that the use of inter-turn silence remains important for older speakers and the interactions between transition type and gap duration (Experiment 1) and condition and gap duration (Experiment 2; marginal), suggest that this effect is not simply the result of having more time to make a gaze switch. These findings thus suggest that silence is an early and lasting cue for identifying turn structure online when other predictive information is not adequate.

Notably, many other non-linguistic cues encode information about transition type, including gaze and gesture. We did not systematically test those cues here but, like inter-turn silence, they may play a critical role in parsing and making predictions about turn structure when other linguistic information is not sufficient to make accurate predictions.

Perhaps surprisingly, we found no evidence that lexical information alone is equivalent to the full linguistic signal in driving children's predictions, as has been shown previously for adults (Magyari & De Ruiter, 2012; De Ruiter et al., 2006) and as is replicated with adult participants in the current study. Unlike prosodic cues, lexicosyntactic cues are discreet and have much clearer form-to-meaning mappings, with clear lexicosyntactic cues to questionhood that occur early within turns (e.g., *wh*-words, *do*-insertion, and subject-auxiliary inversion). That said, children's lexical and syntactic knowledge is limited for quite some time (Tomasello & Brooks, 1999, but see also Bergelson & Swingley, 2013; Shi & Melancon, 2010). Although our stimuli were made in a child-friendly style, they are still other-directed and fairly complex, with 20–30 s of continuous conversational speech.

It is perhaps for this reason that children's performance was always best with the full signal, where lexicosyntactic information was supported by prosodic information and vice versa. Even in adults, Bögels and Torreira (2015) showed that the trade-off in informativity between lexical and prosodic cues is more subtle in semi-natural speech. The present findings are the first to show evidence of a similar effect developmentally.

The question effect

In both experiments, anticipatory looking was primarily driven by question transitions, a pattern that has not been previously reported in other anticipatory gaze studies, on children or adults (Hirvenkari et al., 2013; Keitel et al., 2013; Tice (Casillas) and Henetz, 2011). Questions make an upcoming speaker switch immediately relevant, helping the listener to predict with high certainty what will happen next (i.e., an answer from the addressee), and are often easily identifiable by overt prosodic and lexicosyntactic cues.

Prior work on children's acquisition of questions indicates that they may already have some knowledge of question–answer sequences by the time they begin to speak: questions make up approximately one third of the utterances children hear, before and after the onset of speech, and even into their preschool years, though the type and complexity of questions changes throughout development (Casillas et al., 2016; Fitneva, 2012; Henning, Striano, & Lieven, 2005; Shatz, 1979).[13] For the first few years, many of the questions directed to children are "test" questions—questions that the caregiver already has the answer to (e.g., "What does a cat say?"), but this changes as children get older. Questions help caregivers to get their young children's attention and to ensure that information is in common ground, even if the responses are non-verbal or infelicitous (Bruner, 1985; Fitneva, 2012; Snow, 1977). Moreover, because of their high frequency and relatively limited number of formats, questions, especially *wh*-questions, may be more identifiable and predictable compared to other types of speech acts. So, in addition to having a special interactive status, questions are a frequent, predictable, and core characteristic of many caregiver-child interactions, motivating a general benefit for questions in turn structure anticipation.

Two important routes for future work are then: (1) how does children's ability to monitor for questions in conversation relate to their prior experience with questions? and (2) what is it about questions that makes children and adults more likely to anticipatorily switch their gaze to addressees? If this "question" effect exists for all turns that require an immediate response ("adjacency pairs"; Schegloff, 2007), other turn types, such as imperatives, compliments, and complaints should show similar patterns. If the effect is instead about overall predictability of the syntactic frame, children would instead show similar patterns for other frequent frames from child-directed speech (e.g., "Look at the X"; Mintz, 2003). The recognizability and predictability of syntactic frames is likely to play a role in turn prediction as children become more sophisticated language users, even if the effect is truly about adjacency pairs; for example, rhetorical and tag questions take a very similar form to prototypical polar questions, but usually do not require an answer. So, though it is clear that adults and children anticipate responses more often for questions than non-questions, we do not yet know whether their predictive action is limited to turns formatted as questions, turns with high recognizability and predictability, or turns that project an immediate response from the addressee.

A question effect suggests that participants' spontaneous predictions may be driven by what lies *beyond* the end of the current turn—not just by the upcoming end of the turn itself, as has been focused on in prior work (Bögels & Torreira, 2015; De Ruiter et al., 2006; Keitel et al., 2013; Magyari & De Ruiter, 2012). In future work, it will be crucial to measure prediction from a first-person perspective to find out what kinds of predictions are most relevant to addressees in conversation.

One possible scenario is that listeners in spontaneous, first-person conversation use multiple strategies to make predictions about upcoming turn structure: they could semi-passively attend to incoming speech for cues to upcoming speaker transition (e.g., questions and other adjacency pairs) and, when possible upcoming transition is detected, switch into a more precise turn-end prediction mode (àla De Ruiter et al., 2006). A flexible prediction system like this one allows listeners to continuously monitor ongoing conversation for turn-related cues at a low cost while still managing to plan their responses and come in quickly when needed.

To test this hypothesis, we would need to look at prediction from a first-person perspective, which very little work so far has accomplished (present work included). Although third-party measures enable us to measure participants' predictions without any interference from language production, they also limit our knowledge about how the need to give a response might itself play an important role in addressees' prediction strategies. Recent work has shown that shifts in addressee gaze similar to those measured here indeed occur in spontaneous conversation (Holler & Kendrick, 2015), but much more work is needed to determine how participants make predictions about turn structure in first-person contexts and whether those mechanisms shift at points of imminent speaker change.

Early competence for turn taking?

One of the core aims of our study was to test whether children show an early competence for turn taking, as is

[13] There is substantial variation in question frequency by individual and socioeconomic class (Hart & Risley, 1992; Weisleder, 2012).

proposed by studies of spontaneous mother-infant protoconversation and theories about the mechanisms underlying human interaction in general (Hilbrink et al., 2015; Levinson, 2006). We found evidence that young children make spontaneous predictions about upcoming turn structure: definitely by age two and marginally by age one.

These results contrast with Keitel et al.'s (2013) finding that children cannot anticipate upcoming turn structure at above-chance rates until age three. The current study used an appreciably more conservative random baseline than the one used in Keitel and colleagues' study. Therefore, this difference in age of emergence more likely stems from our use of a more engaging speech style, stereo speech playback, and more typical turn transition durations. The child-friendly style of speech in particular may have helped in two ways: keeping children more engaged with the stimuli and using less syntactically complex and more prosodically exaggerated speech (Fernald et al., 1989; Snow, 1977; Werker & McLeod, 1989) compared to what they would get with adult-adult conversation.

To be clear, young children's "above chance" performance was often still far from adult-like predictive behavior—turn prediction (and the concurrent use of linguistic cues from unfolding speech) increased only gradually with age. Children at ages one and two were still very close to chance in their anticipations and, even at age six, children were not fully adult-like in their predictions. This indicates that young children may at first rely primarily on non-verbal cues, like inter-turn silence, to anticipate turn transitions but that, by adulthood, listeners use both verbal and non-verbal cues to make predictions. Relatedly, adult listeners may be more expert in flexibly adapting to the turn-relevant cues present at any moment, e.g., responding to non-English prosodic cues in Experiment 1.

Taken together, our data suggest that turn-taking skills do begin to emerge in infancy, but that children cannot consistently make effective predictions until they can identify question turns in unfolding speech and react to them quickly. This finding leads us to wonder how participant role (first- instead of third-person) and differences in early interactional experience (e.g., frequent vs. infrequent question-asking from caregivers) feed into this early predictive skill. It also bridges prior work showing a predisposition for turn taking in infancy (e.g., Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001; Snow, 1977) with children's apparently *late* acquisition of adult-like competence for turn taking in spontaneous conversation (Casillas et al., 2016; Ervin-Tripp, 1979; Garvey, 1984; Garvey & Berninger, 1981). It also reinforces the idea that it takes children several years to fully integrate linguistic information into their turn-taking systems (Casillas et al., 2016; Garvey & Berninger, 1981).

What makes the integration of linguistic information so gradual? We suspect that two slow-developing processes—children's linguistic knowledge (e.g., *wh*-words, subject-auxiliary inversion) and their speed of processing for linguistic information (e.g., parsing and retrieval)—both contribute to their ability to make predictions about turn structure in unfolding speech. Children may be able to integrate predictive cues for turn taking from the start, but their knowledge of these cues and their speed in parsing and recognizing them may be too slow at first for use in online prediction. This account falls in line with the early and continued use of non-verbal cues found in the current study, but more work is needed to tease these developmental threads apart.

Limitations and future work

There are at least two major limitations to our work: speech naturalness and participant role. Following prior work (De Ruiter et al., 2006; Keitel et al., 2013), we used phonetically manipulated speech in Experiment 2. This decision resulted in 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 few of them have done so in a conversational context. Future work could instead carefully script speech or cross-splice sub-parts of turns to control for the presence of linguistic cues for turn transition (see, e.g., Bögels & Torreira, 2015).

The prediction measure used in our studies is based on an observer's view of conversation but, because participants' role in the interaction could affect their online predictions about turn taking, an ideal measure would instead capture first-person predictions. If conversational participants' predictions are partly shaped by their need to respond, first-person measures of spontaneous turn prediction will be key to revealing how participants distribute their attention over verbal and non-verbal cues while taking part in everyday interaction, the implications of which relate to theories of online language processing for both language learning and everyday talk.

That said, the third-person paradigm used in the present study still has much to tell us about turn prediction. The task is natural and intuitive in that no instruction is required, which means that it captures spontaneous predictive behavior and can be used with participants of all ages. Frequencies of anticipatory gaze switching appear to be stable across language communities where similar tasks have been tested (Keitel et al., 2013; Keitel & Daum, 2015; Holler & Kendrick, 2015; Hirvenkari et al., 2013)—even from a first-person perspective—so the task is one that measures robust predictive behavior relevant to conversational processing across languages. It also lends itself to many possibilities for controlling the presence of individual verbal and non-verbal cues and has a clear method for assessing random switching baselines across the entire stimulus. Also, if it is the case that response preparation interferes with our ability to see prediction at the ends of incoming turns (Levinson, 2016), third-person paradigms are one of the only ways to measure prediction processes in isolation.

The current findings also make predictions about what we would see in first-person paradigms. For example, a focus on possible upcoming speaker transitions is even more important when the participants themselves may need to respond; we would thus expect question-like effects to occur in first-person paradigms, and perhaps even be amplified compared to third-person paradigms. If so, participants' use of linguistic information would still subserve this goal, with prediction at a premium. Regarding development, the same facts about the complexity of prosody-based prediction and children's initial limited lexical inventories would still hold, as would the use of silence and non-verbal cues to assess and predict turn structure in the absence of clear predictive linguistic information. The paradigm presented here thus has important contributions to make in our understanding of how participants attend to and make predictions about conversational interaction.

Conclusions

Conversation plays a central role in children's language learning. It is the driving force behind what children say and what they hear. Adults use linguistic information to accurately predict turn structure in conversation, which facilitates their online comprehension and allows them to respond relevantly and on time. The present study offers new findings regarding the role of speech acts and linguistic processing in online turn prediction, and has given evidence that turn prediction emerges by age two, increases with age, and is driven by the ability to identify and react to question turns in unfolding speech. However, children's successful integration of online linguistic processing and online predictions about upcoming turn structure develops gradually. When participants can't use predictive linguistic cues (because they are absent, unfamiliar, or are processed too late), children and adults alike rely on retroactive cues such as inter-turn silence to predict upcoming speaker change. Using language to make predictions about upcoming interactive content takes time to develop and, for participants of all ages appears to be primarily driven by participants' expectations about what will happen next, beyond the end of the current turn.

Acknowledgments

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), an NSF Graduate Research Fellowship and NSF Dissertation Improvement Grant to MC, and a Merck Foundation fellowship to MCF. Earlier versions of these data and analyses were presented to conference audiences (Casillas & Frank, 2012; Casillas & Frank, 2013). We also thank Tania Henetz, Francisco Torreira, Stephen C. Levinson, and Eve V. Clark for their feedback on earlier versions of this work. The analysis code for this project can be found on GitHub at https://github.com/langcog/turn_taking/.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.06.013.

References

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* <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.

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* <http://www.praat.org> [Computer program] Version 5.3.16.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5* 12881.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52*, 46–57.

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. (2016). Turn taking, timing, and planning in early language acquisition. *Journal of Child Language*, 1–28.

Casillas, M., & Frank, M. C. (2012). Cues to turn boundary prediction in adults and preschoolers. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of SemDial (SeineDial): The 16th workshop on the semantics and pragmatics of dialogue* (pp. 61–69).

Casillas, M., & Frank, M. C. (2013). The development of predictive processes in children's discourse understanding. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 299–304).

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141–201.

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. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology, 6*.

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*, 283–292.

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

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*, 477–501.

Fitneva, S. (2012). Beyond answers: questions and children's learning. In J.-P. De Ruiter (Ed.), *Questions: Formal, functional, and interactional perspectives* (pp. 165–178). Cambridge, UK: Cambridge University Press.

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* (Vol. 21) : . Harvard University Press.

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

Gísladóttir, R., Chwilla, D., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PloS One, 10*, e0120068.

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*, 1096–1105.

Hedberg, N., Sosa, J. M., Görgülü, E., & Mameni, M. (2010). The prosody and meaning of Wh-questions in American English. In *The proceedings of speech prosody* (pp. 100045:1–100045:4).

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. (2015). Early developmental changes in the timing of turn-taking: A longitudinal study of mother-infant interaction. *Frontiers in Psychology, 6*.

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.

Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction. *Frontiers in Psychology, 6*.

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., 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.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin, 109*, 490.

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., & Daum, M. M. (2015). The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology, 6*.

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.

Lammertink, I., Casillas, M., Benders, T., Post, B., & Fikkert, P. (2015). Dutch and english toddlers' use of linguistic cues in predicting upcoming turn transitions. *Frontiers in Psychology, 6*.

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*, 1–4 Article number 22.

Levelt, W. J. (1989). *Speaking: From intention to articulation* : . MIT Press.

Levinson, S. C. (2006). On the human "interaction engine". In N. Enfield & S. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 39–69). Oxford: Berg..

Levinson, S. C. (2013). Action formation and ascriptions. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). Malden, MA: Wiley-Blackwell.

Levinson, S. C. (2016). Turn-taking in human communication – Origins and implications for language processing. *Trends in Cognitive Sciences, 20*, 6–14.

Magyari, L., Bastiaansen, M. C. M., De Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience, 26*, 2530–2539.

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

Männel, C., & Friederici, A. D. (2009). Pauses and intonational phrasing: ERP studies in 5-month-old German infants and adults. *Journal of Cognitive Neuroscience, 21*, 1988–2006.

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.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91–117.

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.

Nomikou, I., & Rohlfing, K. J. (2011). Language does something: Body action and language in maternal input to three-month-olds. *IEEE Transactions on Autonomous Mental Development, 3*, 113–128.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria <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.

Reddy, V., Markova, G., & Wallot, S. (2013). Anticipatory adjustments to being picked up in infancy. *PloS One, 8*, e65289.

Ross, H. S., & Lollis, S. P. (1987). Communication within infant social games. *Developmental Psychology, 23*, 241–248.

Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. In J. Sidnell (Ed.), *Conversation analysis: Comparative perspectives* (pp. 187–249). Cambridge: Cambridge University Press.

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. A primer in conversation analysis* (Vol. 1) : . Cambridge University Press.

Shatz, M. (1978). On the development of communicative understandings: An early strategy for interpreting and responding to messages. *Cognitive Psychology, 10*, 271–301.

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.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41*, 443–492.

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. (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. (2015). 1- and 2-year-olds' expectations about third-party communicative actions. *Infant Behavior and Development, 39*, 53–66.

Tice (Casillas), M., & Henetz, T. (2011). Turn-boundary projection: Looking ahead. In L. Carlson, C. Hlscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual meeting of the cognitive science society* (pp. 838–843).

Toda, S., & Fogel, A. (1993). Infant response to the still-face situation at 3 and 6 months. *Developmental Psychology, 29*, 532–538.

Tomasello, M., & Brooks, P. J. (1999). Early syntactic development: A construction grammar approach. In M. Barrett (Ed.), *The development of language* (pp. 161–190). Psychology Press.

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

Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 43*, 230–246.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *Proceedings of LREC*. .