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

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

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

Abstract

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

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

1. Introduction

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

*Corresponding author

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

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

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

1.1. Turn taking

Turn taking itself is not unique to conversation. Many other human activities are organized around sequential turns at action. Traffic intersections and computer network communication both use turn-taking systems. Children's early games (e.g., give-and-take, peek-a-boo) have built-in, predictable turn structure (**??**). 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 (**??**). In all these cases, turn taking serves as a protocol for interaction, allowing the participants to coordinate with one another through sequences of contingent action.

Conversation distinguishes itself from non-conversational turn-taking behaviors by the complexity of the turn sequencing involved. In the examples above (traffic, games, and monkeys) the set of sequence and action types is

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

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

1.2. Adults' turn prediction

Adults have a lot of information at their disposal to help make accurate predictions about upcoming turn content. Lexical, syntactic, and prosodic information (e.g., *wh-* words, subject-auxiliary inversion, and list intonation) can all inform addressees about upcoming linguistic structure (**????**). 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 (**??**). Additionally, the sequential context of a turn can make it clear what will come next: Answers after questions; thanks or denial after compliments, et cetera (**?**).

Prior work suggests that adult listeners primarily use lexicosyntactic information to accurately predict upcoming turn structure (**?**). De Ruiter and

colleagues (**?**) 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. De Ruiter and colleagues found that, with pitch-flattened speech, the timing of participants' button responses was comparable to their timing with the full linguistic signal. But, when no lexical information was available, participants' responses were significantly earlier. 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 (**?**; see also **?**). They can also identify speech acts within the first word of an utterance (**?**), allowing them to start planning their response at the first moment possible (**?**).

The role of prosody for adult turn-prediction is still a matter of debate. De Ruiter and colleagues' (2006) experiment focused on the role of intonation, which is only a partial index of prosody. Prosodic structure is also tied closely to the syntax of an utterance, and so the two linguistic signals are difficult to control independently (**?**). **?** 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 in multi-utterance turns, participants were tricked into pressing the "turn-end" button 29% of the time. Their results suggest that listeners actually do rely on prosodic cues to execute a response (see also de **?**:525). These exper-

[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.

imental findings corroborate other corpus and experimental work promoting a combination of cues (lexicosyntactic, prosodic, and pragmatic) as key for accurate turn-end prediction (**???**).

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

1.3. Children's turn prediction

1.3.1. Observational studies

The majority of work on children's early turn taking has focused on observations of spontaneous interaction. Children's first turn-like structures appear as early as two to three months in proto-conversation with their caregivers (**??**). During proto-conversations, caregivers interact with their infants as if they were capable of making meaningful contributions; they take every look, vocalization, arm flail, and burp as "utterances" in the joint discourse (**???**). 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 (**??**). Infants at this age also notice changes to social contingency outside of turn structure. In the Still Face paradigm, caregivers interact with their infants and then suddenly halt, taking on a neutral expression with a sustained gaze. When faced with this sudden disappearance of social contingency, infants three months and older try a range of methods to reinitiate the interaction, such as vocalization, reaching, and smiling before looking away or getting upset (**??**).

The timing of children's responses to their caregivers' speech shows a non-linear pattern of fall-rise-fall from early infancy to middle childhood. A recent study by Hilbrink et al. (2015) finds that infants' turn timing at three months is often too early or too late: They start vocalizing in overlap on 40% of their caregivers' turns, and their non-overlapped vocalizations come after an average inter-turn silent gap of 350–900 msec (adult average: 200 msec). Between four and nine months, children begin to reduce the number of turns happening in overlap while also improving on their average response latency.

But then, later on, children's response latencies slow down again, peaking at average gaps of more than 1000 msec at nine months, with only very gradual improvement after that (**?**). While children's avoidance of overlap is nearly adult-like by nine months, the timing of their non-overlapped responses stays much longer than the 200 msec standard for the next few years (**??**).

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

1.3.2. Experimental studies

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

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

Keitel and colleagues (**?**) addressed the random baseline issue in their study of 6-, 12-, 24-, and 36-month-olds. They asked participants to watch short videos of conversation, and tracked their eye movements at points of speaker change. They found that children's anticipatory gaze frequency was

only greater than chance for 36-month-olds and adults. Their study was the first to focus on the role of linguistic processing in children's turn predictions. They showed their participants two types of conversation videos: One normal and one with flattened pitch (i.e., with flattened intonation contours), finding that only 36-month-olds were affected by a lack of intonation contours. The adult control group made equal numbers of anticipatory looks in the videos, with and without intonation contours, consistent with prior adult findings (**?**). Keitel and colleagues concluded that children's ability to predict upcoming turn structure relies on their ability to comprehend the stimuli (emerging around 36 months), especially with respect to semantic access. They also suggest that intonation takes a secondary role in turn prediction, but only *after* children acquire more sophisticated, adult-like language comprehension systems (sometime after 36 months).

Although the Keitel et al. (**?**) study constitutes a substantial advance over previous work, it has its own 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, during which switching is least likely to occur (**?**) and thereby maximizing their chances of finding a difference between gaze frequency at turn transitions and their baseline rate. A more conservative baseline would be to compare participants' looking behavior at turn transitions to their looking behavior during randomly-selected windows of time throughout the stimulus. We follow this conservative approach in our work.

Second, the conversation stimuli they used were somewhat unusual. The average gap between turns was 900 msec, which is much longer than typical adult timing, where gaps average around 200 msec (**?**). The speakers in the videos were also asked to minimize their movements while performing a scripted and adult-directed conversation, which would have created a somewhat unnatural stimulus. Additionally, in order 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 more difficult to distinguish the two talkers. Again, we attempt to address these issues in our current study.

Despite these minor methodological issues, the Keitel et al. (**?**) study still demonstrates intriguing age-based differences in children's ability to predict upcoming turn structure, and the results suggest that both semantic and

intonational development *do* play a role in children's looking patterns. Our current work thus takes this paradigm as our starting point.[2]

1.3.3. Prosodic development

The roles of prosody and lexicosyntax in children's turn predictions are currently unknown, but children understand more about prosody than lexicosyntax early in life. Children begin to acquire prosody in the womb, and can distinguish their native language's rhythm type from others (e.g., syllable-timed vs. stress-timed) 2–5 days after birth (**???**). Beginning between four and five months, infants prefer pauses in speech to be inserted at prosodic boundaries, and by 6 months they can start using prosodic markers to pick out sub-clausal syntactic units (**??**). They show preference for the typical stress patterns of their native language over others by 6–9 months (e.g., iambic vs. trochaic), and can use prosodic information to segment the speech stream into smaller chunks from 8 months onward (**???**). In comparison, children show only a limited lexical inventory at six months, and begin to recognize function words just before their first birthdays, with syntactic categorization beginning around 14 months (**??**). Two-month-olds also notice changes in word order, but this ability appears to rely on prosodic cueing (**?**). Generally speaking then, our current knowledge about children's linguistic development points to a possible early advantage for prosody in children's turn-taking predictions.

1.4. The Current Study

We report here on the role of linguistic processing in children's predictions about upcoming turn structure. We focus in particular on how children use prosodic and lexicosyntactic information to make their predictions. Prior work has focused mainly on lexicosyntax and intonation, and not on prosody proper (**?**; **?**, but see **?**), even though infants seem to acquire the basic rhythmic properties of the prosodic signal first (**???**).

In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling for the amount of lexicosyntactic and prosodic information available. In Experiment 1, English-speaking participants viewed video clips of naturalistic conversation from several different languages. We used multiple languages to control for the presence of lexicosyntactic information while keeping prosodic and non-linguistic cues intact.

[2]See also **??**.

The results showed minimal differences between the predictive looking behavior of preschoolers and adults. In Experiment 2, we created artificial (puppet) visual scenes, enabling us to separately control lexicosyntactic and prosodic cues in the conversational stimuli. In this more controlled paradigm, we found that children's predictive looking behavior improved from ages one to six, but that even one-year-olds made more anticipatory looks than would be expected by chance.

In both experiments children consistently looked faster to responders after hearing questions, compared to non-questions. Both prosodic and lexicosyntactic information played a role in children's predictions about turn structure, but the two information sources were used differently at different ages. Our findings overall support an account in which predictive processes for turn taking in conversation are present early, but their integration with linguistic information takes substantial practice.

2. Experiment 1

We recorded participants' eye movements as they watched six short videos of two-person (dyadic) conversation 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-native) prosody, gaze, breath, phrase final lengthening). Using this method, we compared children and adult's anticipatory looks from the current speaker to the upcoming speaker at points of turn transition in English and non-English videos.

2.1. Methods

2.1.1. Participants

We recruited 74 children between ages 3;0–5;11 and 11 undergraduate adults to participate in the experiment. Our child sample included 19 three-year-olds, 32 four-year-olds, and 23 five-year-olds, all enrolled in a local nursery school. All participants were native English speakers. Approximately one-third (N=25) of the children's parents and teachers reported that their child regularly heard a second (and sometimes third or further) language, but

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

only one child frequently heard a language that was used in our non-English video stimuli, and we excluded his data from analyses. None of the adult participants reported fluency in a second language.

2.1.2. Materials

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

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

taneous, but were as close as possible while still remaining child-oriented in topic, prosodic pattern, and lexicosyntactic construction.[3]

After recording, we combined the audio and video files by hand, and cropped each recording to the 30-second interval with the most turn activity. Because we recorded the conversations in stereo, the male and female voices came out of separate speakers during video playback. This gave each voice in the videos a localized source (from the left or right loudspeaker). We coded each turn transition in the videos for language condition (English vs. non-English), inter-turn gap duration (in milliseconds), and speech act (question vs. non-question). The non-English stimuli were coded for speech act from a monolingual English-speaker's perspective, i.e., which turns "sound like" questions, and which don't: we asked five native American English speakers to listen to the audio signal for each turn and judge whether it sounded like a question. We then coded turns with at least 80% "yes" responses as questions.

Because the conversational stimuli were recorded semi-spontaneously, the duration of turn transitions and the number of speaker transitions in each video was variable. We measured the duration of each turn transition from the audio recording associated with each video. We excluded turn transitions longer than 550 msec and shorter than 90 msec, including overlapped transitions, from analysis.[4] 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, stdv=112 msec) were slightly longer than for non-English videos (mean=286, median=251, stdv=122 msec). The longer gaps in the English videos could give them a slight advantage: Our definition of an "anticipatory gaze shift" includes shifts that are initiated during the gap between turns (Figure 2), so participants had slightly more time to make

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

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

anticipatory shifts in the English videos.

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 msec) than non-questions (mean=325, median=315, stdv=118 msec). Non-English videos did not show a large difference in transition time for questions (mean=270, median=257, stdv=116 msec) and non-questions (mean=302, median=252, stdv=134 msec).

2.1.3. Procedure

Participants sat in front of an SMI 120Hz corneal reflection eye-tracker mounted beneath a large flatscreen display. The display and eye-tracker were secured to a table with an ergonomic arm that allowed the experimenter to position the whole apparatus at a comfortable height, approximately 60 cm from the viewer. We placed stereo speakers on the table, to the left and right of the display.

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

Children were less inclined to simply sit and watch videos of conversation in languages they didn't speak, so we used a different procedure to keep them engaged: The experimenter started each session by asking the child about what languages he or she could speak, and about what other languages he or she had heard of. Then the experimenter expressed her own enthusiasm for learning about new languages, and invited the child to watch a video about "new and different languages" together. If the child agreed to watch, the experimenter and the child sat together in front of the display, with the child centered in front of the tracker and the experimenter off to the side. Each conversation video was preceded and followed by a 15–30 second 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-second 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 minutes.

2.1.4. Data preparation and coding

To determine whether participants predicted upcoming turn transitions, we needed to define a set of criteria for what counted as an anticipatory gaze shift. Prior work using similar experimental procedures has found that adults and children make anticipatory gaze shifts to upcoming talkers within a wide time frame; the earliest shifts occur before the end of the prior turn, and the latest occur after the onset of the response turn, with most shifts occurring in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, **?**). 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 msec to plan an eye movement, following standards from adult anticipatory processing studies (e.g., **?**).

We checked each participant's gaze at each turn transition for three characteristics (Figure 2): (1) That the participant fixated on the prior speaker for at least 100 msec at the end of the prior turn, (2) that sometime thereafter the participant switched to fixate on the upcoming speaker for at least 100 ms, and (3) that the switch in gaze was initiated within the first 200 msec of the response turn, or earlier. These criteria guarantee that we only counted gaze shifts when: (1) Participants were tracking the previous speaker, (2) switched their gaze to track the upcoming speaker, and (3) did so before they could have simply reacted to the onset of speech in the response. Under this assumption, a gaze shift that was initiated within the first 200 msec of the response (or earlier) was planned *before* the child could react to the onset

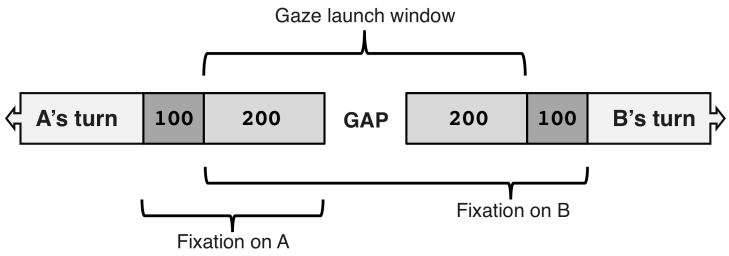


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition.

of speech itself.

As mentioned, most anticipatory switches happen in the inter-turn gap, but we also allowed anticipatory gaze switches that occurred in the final syllables of the prior turn. Early switches are consistent with the distribution of responses in explicit turn-boundary prediction tasks. For example, in a button press task, adult participants anticipate turn ends approximately 200 msec in advance of the turn's end, and anticipatory responses to pitch-flattened stimuli come even earlier (**?**). We therefore allowed switches to occur as early as 200 msec before the end of the prior turn. For very early and very late switches, our requirement for 100 msec of fixation on each speaker would sometimes extend outside of the transition window boundaries (200 msec before and after the inter-turn gap). The maximally available fixation window was 100 msec before and after the earliest and latest possible switch point (300 msec before and after the inter-turn gap). We did not count switches made during the fixation window as anticipatory. We *did* count switches made during the inter-turn gap. The period of time from the beginning of the possible fixation window on the prior speaker to the end of the possible fixation window on the responder was our total analysis window (300 msec + the inter-turn gap + 300 msec).

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 devel-

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

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

opment, along with children's increased linguistic competence.

2.2. 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 the videos were in a language they could understand or not. Children looked at the current speaker less than adults did during the non-English videos. Despite this, their looks to the addressee did not increase substantially in the non-English videos, indicating that their looks away were probably related to boredom rather than confusion about ongoing turn structure. Overall, participants' pattern of gaze to current speakers indicated that they performed basic turn tracking during the videos, regardless of language.

2.2.1. Statistical models

We identified anticipatory gaze switches for all 36 usable turn transitions, based on the criteria outlined in Section 2.1.4, and analyzed them for effects of language, transition type, and age with two mixed-effects logistic regressions (**??**). We built one model each for children and adults. We modeled children and adults separately because effects of age are only pertinent to

Children

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.96146	0.84901	-1.132	0.257446	
Age	-0.18268	0.17507	-1.043	0.296725	
LgCond=*non-English*	-3.29347	0.96045	-3.429	0.000606	***
Type=*non-Question*	-1.10129	0.86494	-1.273	0.202925	
Duration	3.40169	1.22826	2.770	0.005614	**
Age*LgCond=*non-English*	0.52065	0.21190	2.457	0.014008	**
Age*TypeS=*non-Question*	-0.01628	0.19437	-0.084	0.933232	
LgCond=*non-English** Type=*non-Question*	2.68166	1.35016	1.986	0.047013	*
Age*LgCond=*non-English** Type=*non-Question*	-0.45632	0.30163	-1.513	0.130315	

Adults

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.1966	0.6942	-0.283	0.776988	
LgCond=*non-English*	-0.8812	0.9602	-0.918	0.358754	
Type=*non-Question*	-4.4953	1.3139	-3.421	0.000623	***
Duration	-1.1227	1.9880	-0.565	0.572238	
LgCond=*non-English** Type=*non-Question*	3.2972	1.6101	2.048	0.040581	*
LgCond=*non-English** Duration	1.3626	3.0077	0.453	0.650527	
Type=*non-Question** Duration	10.5107	3.3459	3.141	0.001682	**
LgCond=*non-English** Type=*non-Question** Duration	-6.3156	4.4926	-1.406	0.159790	

Table 2: Model output for children and adults' anticipatory gaze switches.

the children's data. The child model included condition (English vs. non-

English)[5], transition type (question vs. non-question), age (3, 4, 5), and duration of the inter-turn gap (seconds, e.g., 0.441) as predictors, with full interactions between condition, transition type, and age. We included the duration of the inter-turn gap as a predictor since longer gaps also provide more opportunities to make anticipatory switches (Figure 2). We additionally included random effects of item (turn transition) and participant, with random slopes of condition, transition type, and their interaction for participants (**?**).[6] The adult model included condition, transition type, duration, and their interactions as predictors with participant and item included as random effects and random slopes of condition, transition type, and their interaction for participant.

Children's anticipatory gaze switches showed effects of language condition (β=-3.29, SE=0.961, t=-3.43, p<.001) and gap duration (β=3.4, SE=1.229, t=2.77, p<.01) with additional effects of an age-by-language condition interaction (β=0.52, SE=0.212, t=2.46, p<.05) and a language condition-by-transition type interaction (β=2.68, SE=1.35, t=1.99, p<.05). There were no significant effects of age or transition type alone (β=-0.18, SE=0.175, t=-1.04, p=.3 and β=-1.10, SE=0.865, t=-1.27, p=.2, respectively).

Adults' anticipatory gaze switches shows an effect of transition type (β=-4.5, SE=1.314, t=-3.42, p<.001) and significant interactions between language condition and transition type (β=3.3, SE=1.61, t=2.05, p<.05) and transition type and gap duration (β=10.51, SE=3.346, t=3.141, p<.01).

2.2.2. Random baseline comparison

We estimated the probability that these patterns were the result of random looking by running the same regression models on participants' real eye-tracking data, only this time calculating their anticipatory gaze switches with respect to randomly permuted turn transition windows. This process

[5]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 (**?**). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

[6]The models we report 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.

involved: (1) randomizing the order and temporal placement of the analysis windows within each stimulus (Figure 3; the analysis window is defined in Figure 2), thereby randomly redistributing the analysis windows across the eye-tracking signal, (2) re-running each participant's eye tracking data through switch identification (described in 2.1.4), but this time using the randomly permuted analysis windows, and (3) modeling the anticipatory gazes from the randomly permuted data with the same statistical models we used for the original data (Section 2.2.1; 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 randomly permuted version of the data.

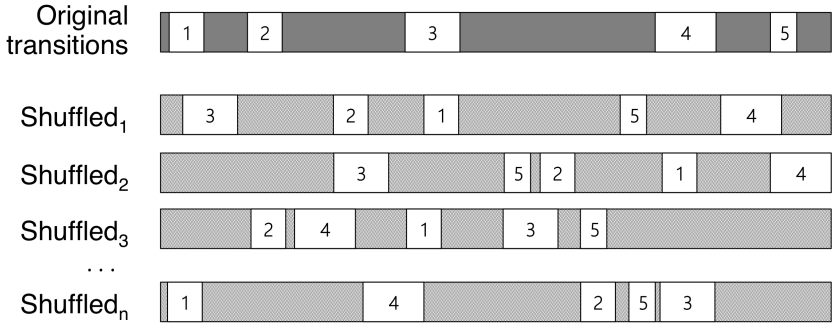
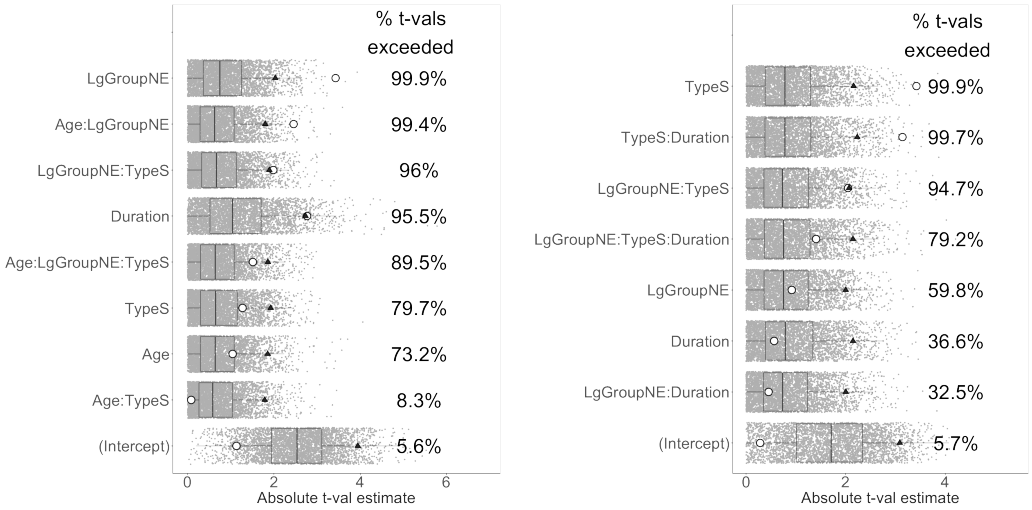


Figure 3: Example of shuffling for five turn transition analysis windows. The windows were ±300 msec around the inter-turn gap.

This procedure effectively 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 movements. The resulting anticipatory gazes from the randomly permuted analysis windows represent an average anticipatory gaze rate over all possible starting points: a random baseline. By running the real and randomly permuted data sets through identical statistical models, we can also estimate how likely it is that predictor effects in the original data (e.g., the effect of language condition; Table 2) arose from random looking.

We completed this baseline procedure on 5,000 random permutations of the original turn transition analysis windows and compared the t-values from each predictor in the original models (Table 2) to the distribution of t-values

for each predictor in the 5,000 models of the randomly permuted datasets.[7] We could then test whether significant effects from the original statistical models differed from the random baseline by calculating the proportion of random data t-values exceeded by the original t-value for each predictor, using the absolute value of all t-values for a two-tailed test. For example, children's original "language condition" t-value was $|3.429|$, which is greater than 99.9% of all $|t\text{-value}|$ estimates from the randomly-permuted data models (i.e., $p = .001$). This leads us to conclude that that the effect of language condition in the original model was highly unlikely to be the result of random gaze shifting.



(a) Children (b) Adults

Figure 4: Random-permutation and original $|t$-values$|$ for predictors of children and adults' anticipatory gaze rates. Gray dots = random model estimates, White dots = original model estimates, Triangles = 95th percentile for each t-value distribution.

Our baseline analyses revealed that none of the significant predictors from models of the original, turn-related data can be explained by random looking.

[7]We report t-values rather than beta estimates because the standard errors in the randomly permuted data models were much higher than for the original data. For those interested, plots of the beta and standard error distributions are available in the Appendix.

The children's data showed strong evidence of differentiation from the randomly permuted data for all four significant effects in the original model (Table 2: Children): the original t-values for language condition, gap duration, the age-language condition interaction, and the language condition-transition type interaction were all greater than 95% of t-values for the randomly permuted data (99.9%, 95.5%, 99.4%, and 96%, respectively; Figure 4a). Similarly, the adults' data showed significant differentiation from the randomly permuted data for two of the three originally significant predictors—transition type and the transition type-gap duration interaction (greater than 99.9% and 99.7% of random t-values, respectively)—with marginal differentiation for the interaction of language condition and transition type (greater than 94.7% of random t-values; Figure 4b). The effects of language condition and transition type for the real and randomly permuted data can also be observed in Figure 5.

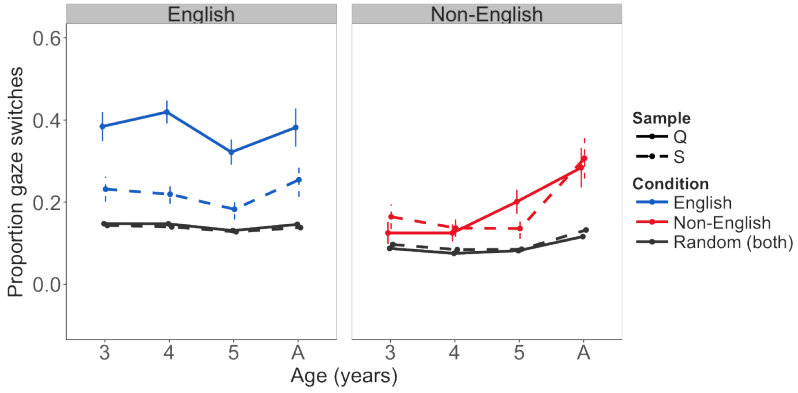


Figure 5: Anticipatory gaze rates across language condition and transition type for the real (red and blue) and randomly permuted (gray) data. Vertical bars represent the standard error.

Developmental effects. The model of the children's data revealed a significant interaction of age and language condition (tab:E1-models) that was highly unlikely to have derived from random looking (Figure 5). To further explore this effect, we compared the average effect of language condition across all ages: we extracted the average difference score for the two language conditions (English minus non-English) for each subject, computing

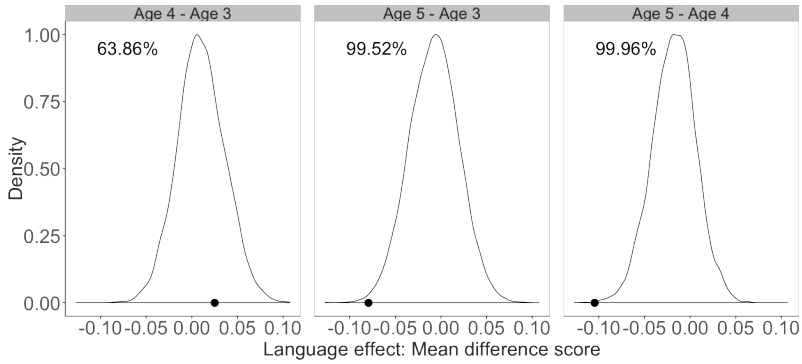


Figure 6: Pairwise comparisons of the language condition effect across ages for the original data (black dots) and the 5,000 randomly permuted datasets (distribution).

an overall average for each random permutation of the data. For each random permutation, we then made pairwise comparisons of the average difference scores across ages. Figure 6 plots the real-data difference scores against the random-data difference score distribution for each pairwise age comparison, showing that 3- and 4-year olds were affected equally by language condition, but that 5-year-olds affected less than both 3- and 4-year-olds (with 99.52% and 99.96% of difference scores greater than the randomly permuted data, respectively, i.e., differences of $p < .01$ and $p < .001$).

2.3. 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 (Table 1 ; Figure 5). 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 (Figure 4a).

Language condition (English vs. non-English) affected children's anticipations in two ways (Table 2; Figure 5). First, children made more anticipatory switches overall in English videos, compared to non-English videos. This effect suggests that lexical access is important for children's ability to anticipate upcoming turn structure; children had no lexical access to the speech in the non-English videos, though they did have access to (non-native) prosodic

cues and non-verbal behavior, consistent with prior work on turn-end prediction in adults (**?**; **?**) and children (**?**). Second, children systematically made more anticipatory switches after hearing a question compared to a non-question, but only in the English condition, suggesting that, when children have access to lexical cues, they are more likely to make an anticipatory gaze switch when they can expect an immediate response from the addressee (i.e., an answer). In itself, this explanation also suggests that children are picking up on lexical cues to questionhood in anticipating upcoming answers (e.g., subject-auxiliary inversion, *wh*-words, etc.).

Children's anticipatory gaze switches were also affected by their age, but only in the non-English videos: 3- and 4-year-olds made many more anticipatory switches when watching videos in English compared to non-English, but this effect of language condition had attenuated significantly by age 5 (Table 2; Figure 5; Figure 6). This interaction 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-native prosodic or non-verbal cues. Prior work on children's turn-structure anticipation proposed that children's turn-end predictions center on lexicosyntactic structure (and not, e.g., prosody) as they get older. The current results suggest more flexibility in children's predictions; when they do not have access to linguistic structure they are more likely to rely on alternative cues as they get older.

Finally, children showed an effect of gap duration (Table 2). This effect is straightforward: longer gaps resulted in longer analysis windows, yielding more time for children to make an anticipatory gaze.

Adults' anticipatory gaze rates were also affected by transition type, language condition, and gap duration (Table 2), none of which could be easily explained by a baseline of random gaze switching (Figure 4b). Like children, adults made more anticipatory switches after hearing questions compared to non-questions, suggesting that anticipation mattered more to them when an immediate response was expected. Also like children, the advantage for questions was driven by lexical access such that adults must have relied on lexicosyntactic cues to questionhood in picking out turns that potentially require an immediate response, though this effect was only marginally significant compared to the distribution of randomly permuted data ($p = .053$; Figure 4b). Finally, adults' anticipation rates were also affected by gap duration, but more so for questions than non-questions (Table 2). This interaction suggests that adults were less likely overall to make anticipatory switches at

non-questions (as is evident for adults and children in Figure 5), and so did not benefit from extra time to do so compared to long gaps for questions.

2.3.1. Summary

Children and adults' predictions alike were benefited by access to lexical information (English) and speech act status (questionhood), suggesting that linguistic cues, particularly lexical ones, facilitate their spontaneous predictions about upcoming turn structure through the identification of turns that will have immediate responses from the addressee. Children's anticipatory gaze rates for questions and non-questions in English was stable across ages and comparable to adult behavior (Figure 5), suggesting that they can identify questions in native stimuli with adult-like competence by age three. Although participants' ability to recognize questions was facilitated by lexical access (i.e., English vs. non-English), the prosody in the non-English videos was non-native, and so the experimental design can not conclusively show which linguistic cues children relied on in the English videos to identify question turns.

Relatedly, though lexical access clearly facilitated participants' anticipatory gaze rate, it was not necessary for participants, especially adults, in order to exceed chance switching rates (Figure 5), suggesting that participants use non-lexical cues (e.g., prosody, non-verbal behavior) to make anticipatory eye movements at least some of the time. In the case of adults there was no overall effect of language condition, only an interaction between language condition and transition type: adults spontaneously anticipated upcoming speakers whether they had lexical access or not, but when given lexical access, focused their anticipations on turns with immediate responses (questions).

Interestingly, adults and children both were strongly affected by transition type, in that they made more anticipatory switches after hearing questions, compared to non-questions. We had hypothesized that a projected response might affect participants' predictions but it in fact seems to be driving many of the effects in our data. Even in the English videos, when participants had full access to linguistic cues, their rates of anticipation were relatively low—in fact, comparable to the non-English videos—unless the turn was a question (Figure 5). Prior work using online, metalinguistic tasks has shows that participants can use linguistic cues to accurately predict upcoming turn ends. The current results suggest that, in their spontaneous predictions, both children and adults monitor the linguistic structure of unfolding turns for cues to upcoming responses. Because the form of an answer is partially

determined by its question (e.g., locative expressions for "where" questions), participants could even be predicting what type of response they will hear, and so they are looking anticipatorily at the responder for confirmation of their prediction. This interpretation is initial, but in-line with work showing early response planning and speech act recognition (**??**).

Overall, children and adults behaved relatively similarly and our language manipulation (English vs. non-English) was too coarse to comment on when children begin to use different types of native linguistic cues (e.g., prosody vs. lexicosyntax); we would instead need to directly compare lexicosyntactic and prosodic cues in the participants' native language, controlling for the presence of non-verbal cues. To see the emergence of anticipatory gaze switching we would also need to include younger children since participants already reliably made anticipatory gaze switches at age three. In sum, Experiment 1 lays the analytic groundwork for a method that allows for greater experimental control, which we introduce in Experiment 2.

3. Experiment 2

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

Non-verbal cues in Experiment 1 (e.g., gaze and gesture) could have helped participants make predictions about upcoming turn structure (**??**). Since our focus is on linguistic cues, we eliminated all gaze and gestural signals in Experiment 2 by replacing the videos of human actors with videos of puppets. Puppets are less realistic and expressive than human actors, but they create a natural context for having somewhat motionless talkers in the videos (thereby allowing us to eliminate gestural and gaze cues). Additionally, the prosody-controlled condition included small but global changes to syllable duration that would have required complex video manipulation or precise re-enactment with human talkers, neither of which was feasible. For

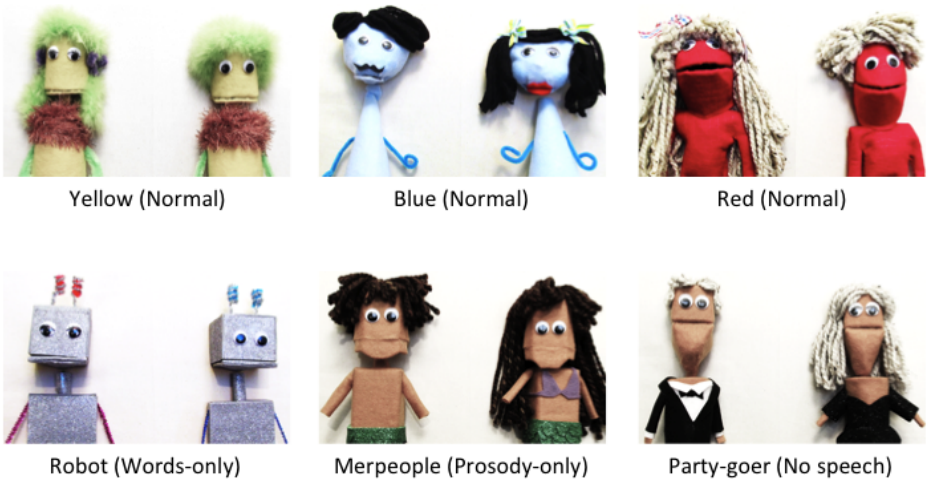


Figure 7: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

these reasons, we decided to substitute puppet videos for human videos in the final stimuli.

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

3.1. Methods

3.1.1. Participants

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

3.1.2. Materials

We created 18 short videos of improvised, child-friendly conversation (Figure 7). 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 minutes) and a pre-determined topic for the child-friendly improvisation ('riding bikes', 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the talkers were native English speakers, and were recorded in male-female pairs. As before, we asked talkers to speak "as if they were on a children's television show" and to ask at least a few questions during the improvisation. We cut each audio recording down to the 20-second interval with the most turn activity. The 20-second clips were then phonetically manipulated and used in the final video stimuli.

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

The *words only* clips were manipulated to have robot-like speech: We flattened the intonation contours to each talker's average pitch (F0) and we reset the duration of every nucleus and coda to each talker's average nucleus and coda duration.[8] We made duration and pitch manipulations using PSOLA resynthesis in Praat (**?**). Thus, the *words only* versions of the audio clips had no pitch or durational cues to upcoming turn boundaries, but did have intact lexicosyntactic cues (and residual phonetic correlates of prosody, e.g., intensity).

We created the *prosody only* clips by low-pass filtering the original recording at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006). This manipulation creates a "muffled speech" effect because low-pass filtering removes most of the phonetic information used to distinguish between

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

phonemes. The *prosody only* versions of the audio clips lacked lexical information, but retained their intonational and rhythmic cues to upcoming turn boundaries.

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

Finally, because low-pass filtering removes significant acoustic energy, the *prosody only* clips were much quieter than the other three conditions. Our last step was to downscale the intensity of 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-second audio clips. The puppets were minimally expressive; the experimenter could only control the opening and closing of their mouths; their head, eyes, arms, and body stayed still. Puppets were positioned looking forward to eliminate shared gaze as a cue to turn structure (**?**). We took care to match the puppets' mouth movements to the syllable onsets as closely as possible, specifically avoiding any mouth movement before the onset of a turn. We then added the manipulated audio clips to the puppet video recordings by hand.

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

The duration of turn transitions and the number of speaker changes across videos was variable because the conversations were recorded semi-

spontaneously. We measured turn transitions from the audio recording of the *normal*, *words only*, and *prosody only* conditions. There was no audio from the original conversation in the *no speech* condition videos, so we measured turn transitions from the video recording, using ELAN video editing software (**?**).

There were 85 turn transitions for analysis after excluding transitions longer than 550 msec and shorter than 90 msec. The remaining turn transitions had slightly more questions than non-question (N=50 and N=35, respectively), with transitions distributed somewhat evenly across conditions (keeping in mind that there were three *normal* videos and only one limited cue video for each experiment version): *Normal* (N=36), *words only* (N=13), *prosody only* (N=17), and no speech (N=19). Inter-turn gaps for questions (mean=365, median=427) were longer than those for non-questions (mean=302, median=323) on average, but gap duration was overall comparable across conditions: *Normal* (mean=334, median=321), *words only* (mean=347, median=369), *prosody only* (mean=365, median=369), and *no words* (mean=319, median=329). The longer gaps for question transitions could give them an advantage because our anticipatory measure includes shifts initiated during the gap between turns (Figure 2).

3.2. Procedure

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

3.2.1. 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)[9] and identified anticipatory gaze switches to the upcoming speaker using the methods from Experiment 1.

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

3.3. 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). Children and adults primarily kept their eyes on the person who was currently speaking: They gazed at the current speaker between 44% and 69% of the time, looking back at the addressee between 11% and 14% of the time (Table 2). They tracked the current speaker in every condition—even one-year-olds looked more at the current speaker than at anything else in the three limited cue conditions (40% for *words only*, 43% for *prosody only*, and 39% for *no speech*). There was a steady overall increase in looks to the current speaker with age and added linguistic information (Tables 3 and 4). Looks to the addressee also decreased with age, but the change was minimal.

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

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

Table 4: Average proportion of gaze to the current speaker and addressee during periods of talk across conditions.

3.3.1. Statistical models

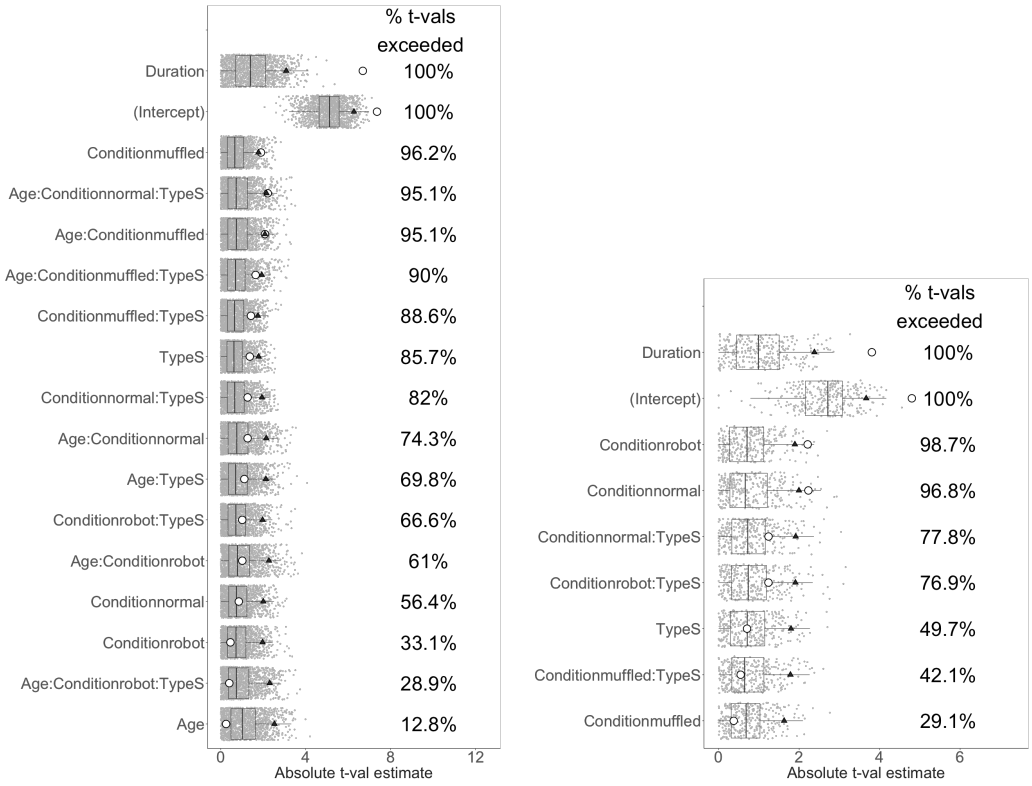
We identified anticipatory gaze switches for all 85 usable turn transitions, and analyzed them for effects of language condition, transition type, and age with two mixed-effects logistic regressions (**??**). We again built separate models for children and adults because effects of age were only pertinent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

Children's anticipatory gaze switches showed an effect of gap duration (β=4.18, SE=0.624, t=6.689, p<.001), a two-way interaction of age and language condition (for prosody only speech compared to the no speech intercept; β=0.393, SE=0.189, t=2.08, p<.05), and a three-way interaction of age, transition type, and language condition (for normal speech compared to the no speech intercept; β=-0.375, SE=0.17, t=-2.213, p<.05). There were no significant effects of age or transition type alone (Table **??**), with only a marginal effect of language condition (for prosody only compared to the no speech intercept; β=-1.634, SE=0.864, t=-1.89, p=.06)

Adults' anticipatory gaze switches showed effects of gap duration (β=4.75, SE=1.248, t=3.806, p<.001) and language condition (for normal speech β=1.256, SE=0.563, t=2.229, p<05. and words only speech β=1.594, SE=0.721, t=2.211, p<.05 compared to the no speech intercept). There were no effects of transition type (β=0.429, SE=0.609, t=0.705, p=.48).

3.3.2. Random baseline comparison

Using the same technique described in experiment 1 (Section 2.2.2), we created and modeled 5,000 random permutations of participants' anticipatory gaze. Our baseline analyses revealed that none of the significant predictors from models of the original, turn-related data (Table 5: Children) can be explained by random looking. In the children's data, the original t-values for language condition (prosody only), gap duration, the two-way interaction of



(a) Children (b) Adults

Figure 8: Random-permutation and original $|t$-values$|$ for predictors of children and adults' anticipatory gaze rates. Gray dots = random model estimates, White dots = original model estimates, Triangles = 95th percentile for each t-value distribution.

age and language condition (prosody only) and the three-way interaction of age, transition type, and language condition (normal only) were all greater than 95% of the randomly permuted t-value (96.2%, 100%, 95.1%, and 95.1%, respectively; Figure 8a). Similarly, the adults' data showed significant differentiation from the randomly permuted data for all originally significant predictors: language condition for normal and words only speech and gap duration (greater than 96.8%, 98.7%, and 100% of random t-values, respectively; Figure 8b). The effects of language condition and transition type for the real and randomly permuted data can also be observed in Figure 5.

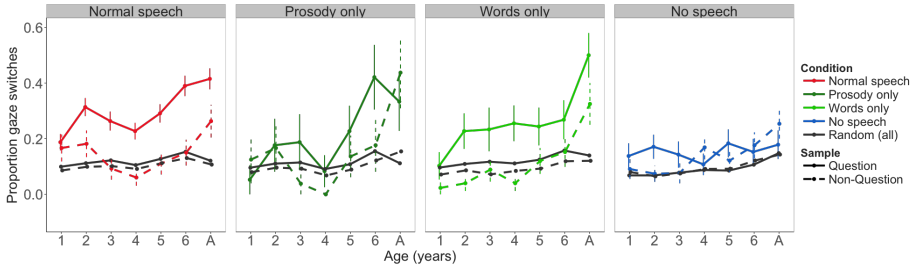


Figure 9: Anticipatory gaze rates across language condition and transition type for the real (blue, dark green, light green, and red) and randomly permuted (gray) data. Vertical bars represent the standard error.

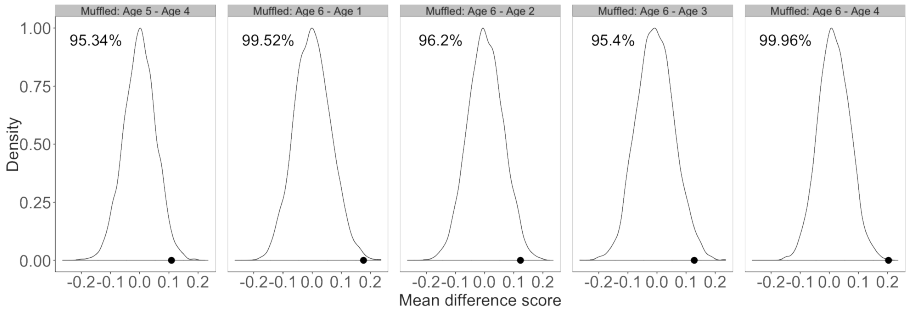


Figure 10: Significant pairwise comparisons of the prosody only-no speech linguisitic condition effect, across ages, for the original data (black dots) and the 5,000 randomly permuted datasets (distribution).

Developmental effects. The model of the children's data revealed two significant interactions with age, neither of which derived from random looking (Figure 9). The first was a significant interaction of age and language condition (for prosody only compared to the no speech baseline), suggesting a different age effect between the two linguistic conditions. As in Experiment 1, we further explored the age effect by extracting the average difference score over subjects for the interaction in each random permutation of the data, making pairwise comparisons between the six age groups. Figure 10 shows the comparisons that demonstrate a significant difference in age for the no speech baseline vs. the prosody only condition, revealing that anticipatory

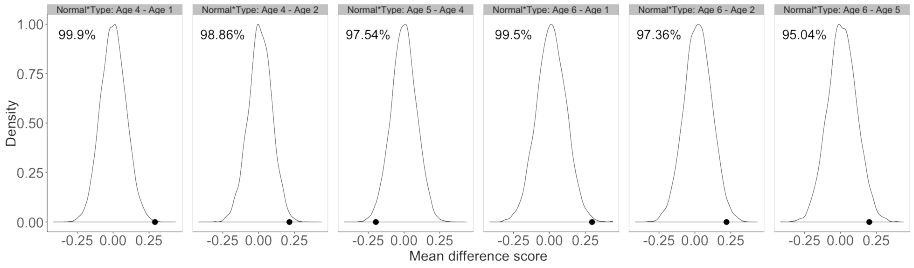


Figure 11: Significant pairwise comparisons of the normal speech-no speech language condition effect for questions status, across ages, for the original data (black dots) and the 5,000 randomly permuted datasets (distribution).

gaze in the prosody only condition significantly improves with age, especially at ages 5 and 6, compared to the no-speech baseline (all with difference scores greater than 95% of the random data scores; $p < .05$).

The second interaction with age 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 interaction in each random permutation of the data, with significant differences shown in Figure 11. These pairwise comparisons showed 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 equating to significant differences from chance ($p < .05$).

3.4. Discussion

3.4.1. Summary

4. General Discussion

Acknowledgements

We gratefully acknowledge the parents and children at Bing Nursery School and the Children's Discovery Museum of San Jose. This work was supported by an ERC Advanced Grant to Stephen C. Levinson (269484-INTERACT), NSF graduate research and dissertation improvement fellowships to the first author, and a Merck Foundation fellowship to the second author. Earlier

versions of these data and analyses were presented to conference audiences (**??**). We also thank Tania Henetz, Francisco Torreira, Stephen C. Levinson, Eve V. Clark, and the First Language Acquisition group at Radboud University for their feedback on earlier versions of this work.

Children

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.57414	0.48576	-7.358	1.87e-13	***
Age	0.02543	0.10260	0.248	0.8042	
Condition=*muffled*	-1.63429	0.86390	-1.892	0.0585	.
Condition=*normal*	0.36710	0.43296	0.848	0.3965	
Condition=*robot*	-0.26741	0.59071	-0.453	0.6508	
Type=*non-Question*	-0.81873	0.59985	-1.365	0.1723	
Duration	4.17672	0.62446	6.689	2.25e-11	***
Age*Condition=*muffled*	0.39317	0.18907	2.080	0.0376	*
Age*Condition=*normal*	0.12919	0.10227	1.263	0.2065	
Age*Condition=*robot*	0.13740	0.13568	1.013	0.3112	
Age*Type=*non-Question*	0.15116	0.13643	1.108	0.2679	
Condition=*muffled** Type=*non-Question*	1.77190	1.24864	1.419	0.1559	
Condition=*normal** Type=*non-Question*	0.91059	0.72095	1.263	0.2066	
Condition=*robot** Type=*non-Question*	-1.02193	1.01227	-1.010	0.3127	
Age*Condition=*muffled** Type=*non-Question*	-0.47057	0.28703	-1.639	0.1011	
Age*Condition=*normal** Type=*non-Question*	-0.37542	0.16963	-2.213	0.0269	*
Age*Condition=*robot** Type=*non-Question*	0.08946	0.22349	0.400	0.6890	

Adults

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.4557	0.7199	-4.800	1.58e-06	***
Condition=*muffled*	0.3349	0.8965	0.374	0.708692	
Condition=*normal*	1.2556	0.5633	2.229	0.025805	*
Condition=*robot*	1.5938	0.7208	2.211	0.027023	*
Type=*non-Question*	0.4292	0.6089	0.705	0.480916	
Duration	4.7500	1.2480	3.806	0.000141	***
Condition=*muffled** Type=*non-Question*	0.6627	1.2138	0.546	0.585108	
Condition=*normal** Type=*non-Question*	-0.9452	0.7631	-1.239	0.215475	
Condition=*robot** Type=*non-Question*	-1.1265	0.9109	-1.237	0.216201	

Table 5: Model output for children and adults' anticipatory gaze switches.