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

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, we measured children's anticipatory gaze to upcoming responders while controlling linguistic cues to upcoming 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's predictive looking behavior improved from ages one to six, but even one-year-olds made more anticipatory looks than would be expected by chance. In both experiments, children and adults anticipated more often after hearing questions. Like adults, prosody alone did not improve children's predictive gaze shifts. But, unlike adults, lexical cues alone were also not sufficient to improve prediction—children's performance was best overall with access to lexicosyntax and prosody together. Our findings support an account in which turn prediction emerges in infancy, but becomes fully integrated with linguistic processing only gradually.

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

1. Introduction

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

*Corresponding author

by the roles and goals of its participants. But what sets conversation apart is its structure: Sequences of interconnected, communicative actions that take place across alternating turns at talk. Sequential, turn-based structures in conversation are strikingly uniform across language communities and linguistic modalities. Turn-taking behaviors are also cross-culturally consistent in their basic features and the details of their implementation (De Vos et al., 2015; Dingemanse et al., 2013; Stivers et al., 2009).

Children participate in sequential coordination with their caregivers starting at three months of age—before they can rely on any linguistic cues in taking turns (see, among others, Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001; Snow, 1977). Infant turn taking is different from adult turn taking in several ways, however. Infant turn taking is heavily scaffolded by caregivers, has different timing from adult turn taking, 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: Early non-verbal interactions likely establish the protocol by which children come to use language with others. How do children integrate linguistic knowledge with these preverbal turn-taking abilities, and how does this integration change over the course of childhood?

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 the remainder of the introduction, 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. Adults' turn taking

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

Conversation distinguishes itself from non-conversational turn-taking behaviors by the complexity of the turn sequencing involved. In the examples above (traffic, games, and monkeys) the set of sequence and action types is far more limited and predictable than what we find in everyday talk. 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!").

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

Adults have a lot of information at their disposal to help make accurate predictions about upcoming turn content. Lexical, syntactic, and prosodic information (e.g., *wh-* words, subject-auxiliary inversion, and list intonation) can all inform addressees about upcoming linguistic structure (De Ruiter et al., 2006; Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 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 et al., 2009; Stivers and Rossano, 2010). Additionally, the sequential context of a turn can make it clear what will come next: Answers after questions, thanks or denial after compliments, et cetera (Schegloff, 2007).

Prior work suggests that adult listeners primarily use lexicosyntactic information to accurately predict upcoming turn structure (De Ruiter et al., 2006). De Ruiter and colleagues (2006) asked participants to listen to snippets of spontaneous conversation and to press a button whenever they anticipated that the current speaker was about to finish 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' 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 (Magyari and De Ruiter, 2012; see also Magyari et al., 2014). They can also identify speech acts within the first word of an utterance (Gísladóttir et al., 2015), allowing them to start planning their response at the first moment possible (Bögels et al., 2015).

Despite this body of evidence, 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. And in addition, prosody is tied closely to the syntax of an utterance, so the two linguistic signals are difficult to control independently (Ford and Thompson, 1996). Torreira, Bögels and Levinson (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 in multi-utterance turns, participants were tricked into pressing the "turn-end" button 29% of the time. Their results suggest that listeners actually do rely on prosodic cues to execute a response (see also de De Ruiter et al. (2006):525). These experimental findings corroborate other corpus and experimental work promoting a combination of cues (lexicosyntactic, prosodic,

[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 pragmatic) as key for accurate turn-end prediction (Duncan, 1972; Ford and Thompson, 1996; Hirvenkari et al., 2013).

1.2. Children's turn prediction

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

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) that, by nine months, slows down considerably, only gradually speeding up again after 12 months (Hilbrink et al., 2015). Taking turns with brief transitions between speakers is difficult for children; while their 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 (Casillas et al., In press; Garvey, 1984; Ervin-Tripp, 1979). This puzzling pattern is likely due to their linguistic development: Taking turns on time is easier when the response is a simple vocalization rather than a linguistic utterance. Integrating language into the turn-taking system may be one major factor in children's delayed responses (Casillas et al., In press).

While children, like adults, could use linguistic cues in the ongoing turn to make 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 and Ramus, 2003); they show preference for the typical stress patterns of their native language over others by 6–9 months (e.g., iambic vs. trochaic), and can use prosodic information to segment the speech stream into smaller chunks from 8 months onward (Johnson and Jusczyk, 2001; Morgan and Saffran, 1995). Four- to five-month-olds also 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, both of which are useful for extracting turn structure from ongoing speech (Jusczyk et al., 1995; Soderstrom et al., 2003). In comparison, children show at best a very limited lexical inventory before their first birthday (Bergelson and Swingley, 2013; Shi and Melancon, 2010).

Keitel and colleagues (2013) were one of the first to explore how children use linguistic cues to predict upcoming turn structure. They asked 6-, 12-, 24-, 36-month-old, 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 conversation videos—one normal and one with flattened pitch (i.e., with flattened intonation contours)—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. Among those, only 36-month-olds were affected by a lack of intonation contours, leading 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 suggested that intonation might play a secondary role in turn prediction, but only after children acquire more sophisticated, adult-like language comprehension abilities.

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 perhaps the period in which switching is least likely to occur (Hirvenkari et al., 2013)— thus, this particular baseline maximizes the chance of finding a difference between gaze frequency at turn transitions and baseline. 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, including turn transitions. We follow this conservative approach in our work.

Second, the conversation stimuli Keitel et al. (2013) used were somewhat unusual. The average gap between turns was 900 msec, which is much longer than typical adult timing, where gaps average around 200 msec (Stivers et al., 2009). The speakers in the videos were also asked to minimize their movements while performing a scripted and adult-directed conversation, which would have created a somewhat unnatural stimulus. 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. (2013) study still demonstrates intriguing age-based differences in children's ability to predict upcoming turn structure. Our current work thus takes this paradigm as a starting point.[2]

1.3. 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 across development. We present two experiments in which we measure children's anticipatory gaze to responders while watching conversation videos with natural (people using 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.

Because the results of our experiments are complex, we highlight three primary findings. First, although children and adults use linguistic cues to make predictions about upcoming turn structure, they do so primarily in predict speaker transitions after questions (a speech act effect). This speech act effect, which we did not initially predict, is intriguing and suggests that previous work may have neglected an important dimension of linguistic cues. Second, we find that children make more predictions than expected by chance starting at age two, but that this effect is small with our stimuli. Third, we find no evidence of an early prosody advantage in children's anticipations and, further, no evidence that prosodic or lexical cues alone can substitute for their combination in the full linguistic signal, as is proposed for adults

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

(De Ruiter et al., 2006); instead, anticipation is strongest for a stimulus with the full range of cues. In sum, our findings support an account in which turn prediction emerges in infancy, but becomes fully integrated with linguistic processing only gradually across development.

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 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,

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

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

After recording, we combined the audio and video files by hand, and cropped each recording to the 30-second interval with the most turn activity. Because we recorded the conversations in stereo, the male and female voices

[3] 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.

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, stdev=112 msec) were slightly longer than for non-English videos (mean=286, median=251, stdev=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 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, stdev=118 msec). Non-English videos did not show a large difference in transition time for questions (mean=270, median=257, stdev=116 msec) and non-questions (mean=302, median=252, stdev=134 msec).

[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" (Sacks et al., 1974). As such, they would still be fixated on the prior speaker when the overlap started, and then would have to switch their gaze *reactively* to the responder.

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, 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 msec to plan an eye movement, following standards from adult anticipatory processing studies (e.g., Kamide et al., 2003).

We checked each participant's gaze at each turn transition for three characteristics (Figure 2): (1) That the participant fixated on the prior speaker for at least 100 msec at the end of the prior turn, (2) that sometime thereafter the participant switched to fixate on the upcoming speaker for at least 100 ms, and (3) that the switch in gaze was initiated within the first 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 of speech itself.

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

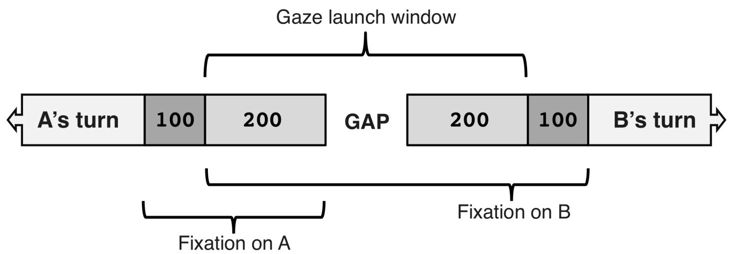


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition.

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 development, 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

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.

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 demonstrated that they performed basic turn tracking during the videos, regardless of language. Figure 3 shows participants' anticipatory gaze rates across age, language condition, and transition type.

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 (Bates et al., 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)[5], transition type (question vs.

[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

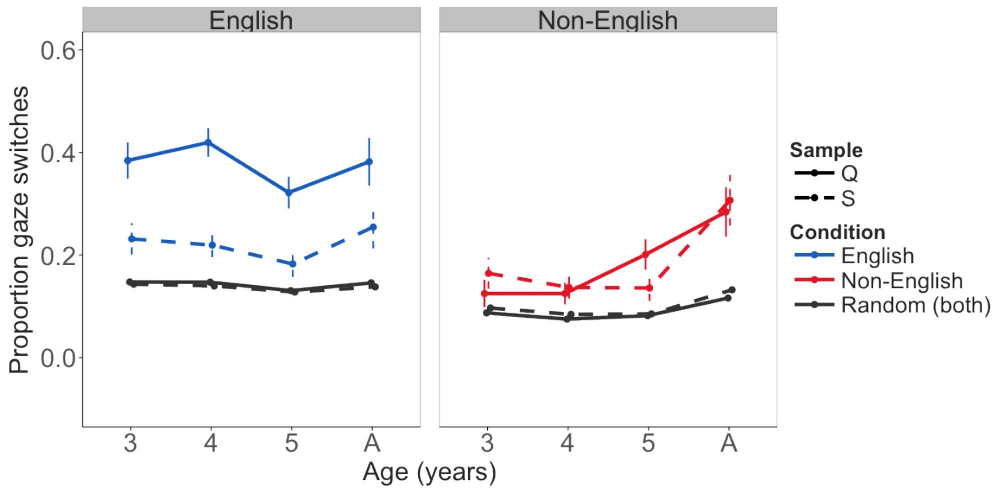


Figure 3: Anticipatory gaze rates across language condition and transition type for the real (red and blue) and randomly permuted baseline (gray). Vertical bars represent the standard error.

non-question), age (3, 4, 5; numeric), 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 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 (Barr et al., 2013).[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,

as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). We would need multiple items from each of the 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.

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.

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=-

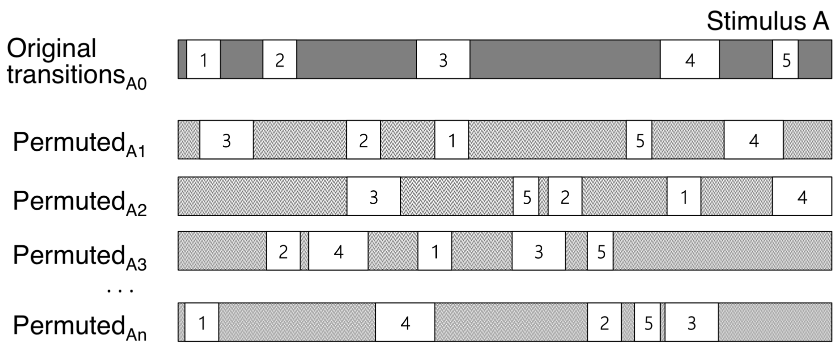


Figure 4: Example of analysis window permutations for a stimulus with five turn transitions. The windows were ±300 msec around the inter-turn gap.

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 involved: (1) Randomizing the order and temporal placement of the analysis windows within each stimulus (Figure 4; "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 Section 2.1.4), 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 random permutation.

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. Because these analyses are complex, we report their full details in Appendix 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 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). 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 94.6% of random t-values).

2.2.3. Developmental effects

The models reported above revealed a significant interaction of age and language condition (Table 2) that was unlikely be due to random looking (Figure 3). To further explore this effect, we compared the average effect of language condition for each age group: Using the permutation analyses above, we extracted the average difference score for the two language conditions (English minus non-English) for each subject, computing 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 pariticipant age groups. Details are given in Appendix B.

These analyses showed that, while 3- and 4-year olds showed similarly large effects of language condition, 5-year-olds showed a significantly smaller effect of language condition, compared to both younger age groups. In other words, the difference in the effect of language condition for 5-year-olds compared to younger children was larger than would be expected by chance in 99.52% of the randomly permuted data sets for 3-year-olds and 99.96% of the data sets for 4-year-olds—differences of $p<.01$ and $p<.001$, respectively (see Figure B.1 for difference score distributions).

When does spontaneous turn prediction emerge developmentally during natural speech? To test whether the youngest age group (3-year-olds) already exceeded chance in their anticipatory gaze switches, we used two-tailed t-tests to compare their real gaze rates to the random baseline in the English condition. Although the overall effect was small, 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, $p<.001$).

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. 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 A.1a). 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. 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. Prior work using online, metalinguistic tasks has shown that participants can use linguistic cues to accurately predict upcoming turn ends (Torreira et al., 2015; Magyari and 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 upcoming responses.

Second, we saw developmental effects such that older children anticipated more reliably, but only in the non-English videos. In the English videos all children anticipated, especially foro questions. But the language condition 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 rely primarily on lexicosyntactic 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 are likely to find alternative cues to turn taking behavior.

In Experiment 2 we follow up on these findings, improving on two aspects of the design. Our language manipulation in this first experiment was too coarse to provide data regarding specific linguistic cues (e.g., prosody vs. lexicosyntax). In Experiment 2, to compare lexicosyntactic and prosodic cues directly, controlling for the presence of non-verbal cues, wed use artificial stimuli. In addition, we saw above-chance anticipation in the youngest children we tested in Experiment 1. Although there were developmental changes in prediction for non-English stimuli, all children anticipated successfully after questions. In Experiment 2 we explore a wider developmental range.

3. Experiment 2

Experiment 2 used native-language stimuli, controlled for lexical and prosodic information, eliminating 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 one- to six-year-olds we hoped to find the developmental onset of turn-predictive gaze. We also hoped to measure changes in the relative roles of prosody and lexicosyntax across development.

Non-verbal cues in Experiment 1 (e.g., gaze and gesture) could have helped participants make predictions about upcoming turn structure (Rossano et al., 2009; Stivers and Rossano, 2010). Since our focus was on linguistic cues, we eliminated all gaze and gestural signals in Experiment 2 by replacing the videos of human actors with videos of puppets. Puppets are less realistic and expressive than human actors, but they create a natural context for having somewhat motionless talkers in the videos (thereby allowing us to eliminate gestural and gaze cues). Additionally, the prosody-controlled condition included small but global changes to syllable duration that would have required complex video manipulation or precise re-enactment with human talkers, neither of which was feasible. For these reasons, we decided to substitute puppet videos for human videos in the final stimuli.

As in the first experiment, we recorded participants' eye movements as they watched six short videos of dyadic conversation, and then analyzed

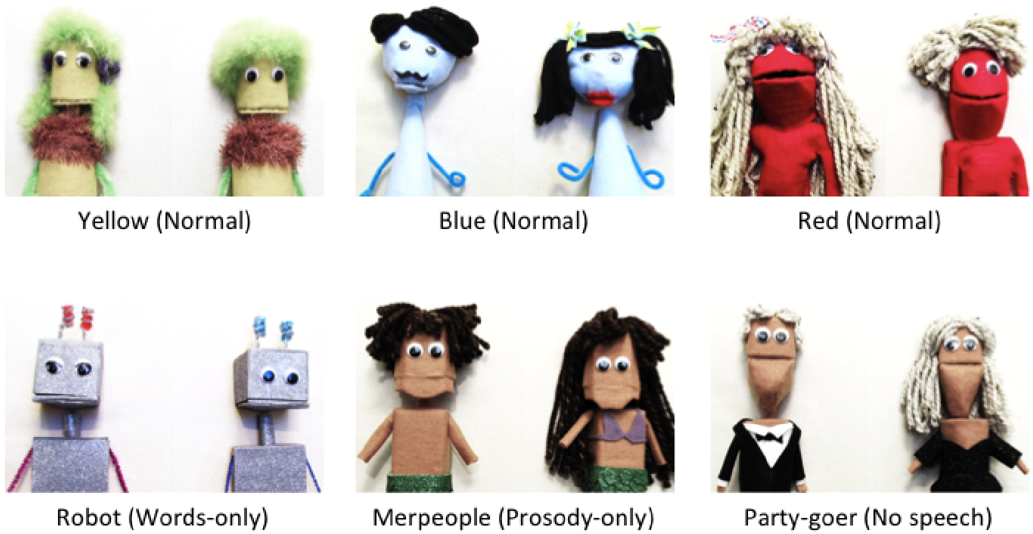


Figure 5: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

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 Children's Discovery Museum of San Jose, California, 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. 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 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 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.[7] We made duration and pitch manipulations using PSOLA resynthesis in Praat (Boersma and Weenink, 2012). Thus, the *words only* versions of the audio clips had no pitch or durational cues 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 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

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

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 (Thorgrímsson et al., 2015). 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 (Wittenburg et al., 2006).

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

3.3. Results

Participants' pattern of gaze indicated that they performed basic turn tracking across all ages and in all conditions. Participants again looked at

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

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.

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 (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. Figure 6 shows participants' anticipatory gaze rates across age, the four language conditions, and transition type.

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

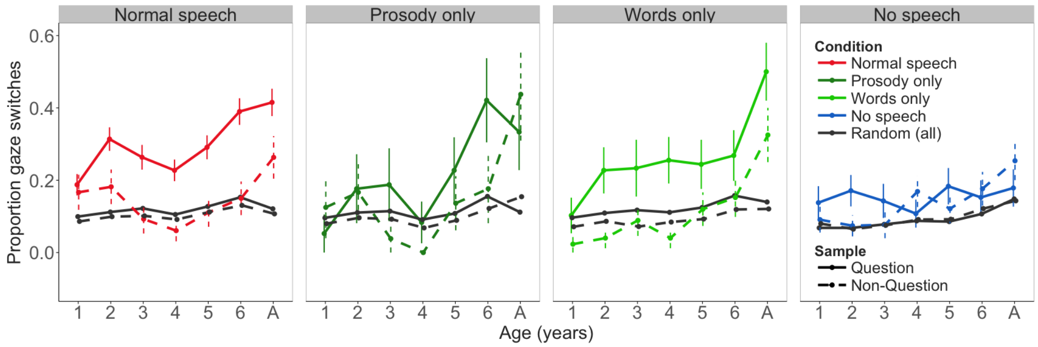


Figure 6: Anticipatory gaze rates across language condition and transition type for the real (blue, dark green, light green, and red) and randomly permuted baseline (gray). Vertical bars represent the standard error.

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), 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

	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
Type=*non-Question*	-0.81873	0.59985	-1.365	0.1723
Duration	4.17672	0.62446	6.689	2.25e-11 ***
Age*Type=*non-Question*	0.15116	0.13643	1.108	0.2679
Condition=*normal*	0.36710	0.43296	0.848	0.3965
Age*Condition=*normal*	0.12919	0.10227	1.263	0.2065
Condition=*normal** Type=*non-Question*	0.91059	0.72095	1.263	0.2066
Age*Condition=*normal** Type=*non-Question*	-0.37542	0.16963	-2.213	0.0269 *

Condition=*prosody*	-1.63429	0.86390	-1.892	0.0585 .
Age*Condition=*prosody*	0.39317	0.18907	2.080	0.0376 *
Condition=*prosody** Type=*non-Question*	1.77190	1.24864	1.419	0.1559
Age*Condition=*prosody** Type=*non-Question*	-0.47057	0.28703	-1.639	0.1011
Condition=*words*	-0.26741	0.59071	-0.453	0.6508
Age*Condition=*words*	0.13740	0.13568	1.013	0.3112
Condition=*words** Type=*non-Question*	-1.02193	1.01227	-1.010	0.3127
Age*Condition=*words** 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 ***
Type=*non-Question*	0.4292	0.6089	0.705	0.480916
Duration	4.7500	1.2480	3.806	0.000141 ***
Condition=*normal*	1.2556	0.5633	2.229	0.025805 *
Condition=*normal** Type=*non-Question*	-0.9452	0.7631	-1.239	0.215475
Condition=*prosody*	0.3349	0.8965	0.374	0.708692
Condition=*prosody** Type=*non-Question*	0.6627	1.2138	0.546	0.585108
Condition=*words*	1.5938	0.7208	2.211	0.027023 *
Condition=*words** Type=*non-Question*	-1.1265	0.9109	-1.237	0.216201

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

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* reference level; 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* reference level; β=-0.375, *SE*=0.17, t=-2.213, p<.05). There were no significant effects of age or transition type alone (Table 3.3.1), with only a marginal effect of language condition (for *prosody only* compared to the *no speech* reference level; 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* reference level). 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 random permutations of participants' anticipatory gaze. These analyses revealed that none of the significant predictors from models of the original, turn-related data could be explained by random looking. In the children's data, the original model's t-values for language condition (*prosody only*), 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 95% of the randomly permuted t-values (96.2%, 100%, 95.1%, and 95.1%, respectively). Similarly, the adults' data showed significant differentiation from the randomly permuted data for all originally significant predictors: gap duration and language condition for *normal* speech and words-only speech (greater than 100%, 96.8%, and 98.7% of random t-values, respectively).

3.3.3. 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 turn structure predictions in our paradigm. As in Experiment 1, we used two-tailed t-tests to compare children's real gaze rates to the random baseline 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 for three age groups: ages one, two, and three. Two- and three-year-old 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, made anticipatory gaze shifts that were not statistically significant 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).

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 further explored each age interaction by extracting an average difference score over subjects 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 (difference scores greater than 95% of the random data scores (p<.05).

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).

3.4. 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. Taken together, our results replicate the findings from Experiment 1: Participants make more anticipatory switches when they have access to lexical information and, when they do, tend to make more anticipatory switches for questions compared to non-questions.

As in Experiment 1, with normal speech, children and adults spontaneously tracked the turn structure of the conversations, making anticipatory gaze switches at an above-chance rate across all ages. And in addition, they made far more anticipations for questions than for non-question turns—at least for two-year-olds and older. But these effects were different for the two comparison conditions, *prosody-only* and *words-only*. In the *prosody-only* condition, performance was low for younger children and increased substantially for older children (especially for questions). In the *words-only* condition, anticipation performance was relatively robust for questions for two-year-olds and older (much like in normal speech), but never rose above chance for the children for non-question turns. Thus, these findings do not support an early role for prosody in children's spontaneous turn structure predictions. On the contrary, children's predictions appeared to stem more from lexical information, probably primarily from question words.

4. General Discussion

Children begin to develop conversational turn-taking skills long before their first words (Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001; Snow, 1977). As they acquire language, they also acquire the information needed to make accurate predictions about upcoming turn structure. Until recently, we have had very little data on how children weave language into their already-existing turn-taking behaviors.

In two experiments investigating children's anticipatory gaze to upcoming speakers, we found evidence that turn prediction develops early in childhood and that spontaneous predictions are primarily driven by participants' expectation of an immediate response in the next turn. In making predictions about upcoming turn structure, children used a combination of lexical and prosodic cues; neither lexical nor prosodic cues alone were sufficient to support increased anticipatory gaze. We also found no early advantage for prosody over lexicosyntax, and instead found that children were unable to make above-chance anticipatory gazes in the *prosody only* condition until age five. We discuss these findings with respect to the role of linguistic cues in predictions about upcoming turn structure, the importance of questions in spontaneous predictions about conversation, and children's developing competence as conversationalists.

4.1. Predicting turn structure with linguistic cues

Prior work with adults has found a consistent and critical role for lexicosyntax in predicting upcoming turn structure (De Ruiter et al., 2006; Magyari and De Ruiter, 2012), with the role of prosody still under debate (Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015). Knowing that children comprehend more about prosody than lexicosyntax early on (see introduction; also see Speer and Ito, 2009 for a review), we thought it possible that young children would instead show an advantage for prosody in their predictions about turn structure in conversation. Our results suggest that, on the contrary, when presented with *only* prosodic information, children's spontaneous predictions about upcoming turn structure are limited until age five.

Importantly, we also found no evidence that lexical information alone is equivalent to full linguistic information for children, as has been shown before (Magyari and De Ruiter, 2012; De Ruiter et al., 2006) and replicated in the current study for adult participants: In both experiments, children's performance was best in conditions when they had access to the full linguistic signal. Adults on the other hand, showed significant gains in anticipatory gaze switching in both conditions with lexical cues. If this effect arose in our data because children do not make predictions based on phonetically manipulated speech in conversational contexts, the current findings can be overturned by follow-up work controlling linguistic information through other means. But if not, there may be something specially informative about the combined prosodic and lexical cues to questionhood that boosts children's anticipations before they can use these cues separately. Even in adults, Torreira and colleagues (2015) were able to show that the trade-off in informativity between lexical and prosodic cues is more subtle in semi-natural (spliced) speech. The present findings are the first to show evidence of a similar effect developmentally.

4.1.1. The question effect

In both experiments, anticipatory looking was primarily driven by question transitions, a pattern that had not been previously reported in other anticipatory gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari, 2013; Tice 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.

Compared to prosodic cues (e.g., final rising intonation), lexicosyntactic cues (e.g., *wh*-words, *do*-insertion, and subject-auxiliary inversion) were frequent, categorical, and early-occurring in the utterance. Children may have therefore had an easier time picking out and interpreting lexical cues to questionhood. The question effect showed its first significant gains between ages three and four in the *normal* speech condition of Experiment 2, by which time children frequently hear and use a variety of polar *wh*-questions (Clark, 2009). Furthermore, while lexicosyntactic question cues were available on every instance of *wh*- and *yes/no* questions in our stimuli, prosodic question cues were only salient on *yes/no* questions and, even then, the mapping of prosodic contour to speech act (e.g., high final rises for polar questions) is far from one-to-one.

Prior work on children's acquisition of questions indicates that they may already have some understanding about 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, even though the types and complexity of questions change throughout development (Casillas et al., In press; Fitneva, 2012; Henning et al., 2005; Shatz, 1979).[9] 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). So, in addition to having a special interactive status (for adults and children alike), questions are a core characteristic of many caregiver-child interactions, motivating a general benefit for questions in turn structure anticipation.

Two important questions for future work are then: (a) How does children's ability to monitor for questions in conversation relate to their prior experience with questions? and (b) what is it about questions that makes children and adults more likely to anticipatorily switch their gaze to addressees? Other request formats, such as imperatives, compliments, and complaints make a response from the addressee highly likely in the next turn (Schegloff, 2007). Rhetorical and tag questions, on the other hand, take a similar form to prototypical polar questions, but often do not require an answer. So, though it is clear that adults and children anticipated responses more often for questions than non-questions, we do not yet know whether their predictive action is limited to turns formatted as questions or is generally applicable to turn structures that project an immediate response from the addressee.

The question effect itself has implications for our current theories about turn prediction. While participants may always use linguistic information to predict upcoming speaker changes (as they have in the two studies presented here), they might not always use linguistic information to predict upcoming turn ends, as is assumed in metalinguistic measures of turn-end prediction (e.g., pressing a button while listening to speech: Torreira et al.,

[9]There is substantial variation question frequency by individual and socioeconomic class (Hart and Risley, 1992).

2015; Magyari and De Ruiter, 2012; De Ruiter et al., 2006) and as has been the focus of other turn-end prediction studies (Ford and Thompson, 1996; Duncan, 1972). Our results rather suggest that participants' *spontaneous* predictions, at least while viewing third-party conversation, are the results of question-monitoring, presumably achieved by recruiting linguistic cues to questionhood from the unfolding signal. In other words, our results suggest that predictions are driven by what is *beyond* the end of the current turn—that questions, not lexical cues, are sufficient for prediction at the first level of conversation monitoring.

There is at least one clear hypothesis that can bridge these apparently conflicting results: Listeners in spontaneous, first-person conversation may use multiple strategies in making predictions about upcoming turn structure, using more passive prediction to detect upcoming speaker transition (e.g., questions), and then switching into precise turn-end prediction mode (à la De Ruiter et al., 2006) when necessary. A flexible prediction system like this one allows listeners to continuously monitor ongoing conversation at a low cost while still managing to plan their responses and come in quickly when needed.

To test this hypothesis, which integrates findings from turn-structure prediction with multiple measures, age groups, and styles of linguistic control, it will be crucial to look at prediction from a first-person perspective. The results we present here are based on predictions about third-party conversation, which enables participants to follow interactions with no chance of actually participating. Although recent work has shown that similar anticipatory eye gazes do occur in spontaneous conversation (Holler and Kendrick, 2015), more work is needed to determine if the same question advantage occurs, which linguistic cues seem to drive it, and whether participants switch into a "precision" mode when they detect imminent speaker change.

4.1.2. 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 proposed by studies of spontaneous mother-infant proto-conversation and theories about the mechanisms underlying human interaction in general (Hilbrink et al., 2015; Levinson, 2006). We did find evidence that young children already make spontaneous predictions about upcoming turn structure, definitely at age two and even marginally at age one. However, "above chance" performance was far from adult-like predictive behavior, and children in our studies did not show adult-like competency in their predictions, even at age six. This may indicate that children rely more on non-verbal cues in anticipating turn transitions or, alternatively, that adults are better at flexibly adapting to the turn-relevant cues present at any moment.

Taken together, the data suggest that turn-taking skills do begin to emerge in infancy, but that their predictions don't help them anticipate much until they have acquired the ability to pick out question turns. This finding leads us to wonder how participant role (first- instead of third-person) and cultural differences (e.g., high vs. low parent-infant interaction styles) might feed into this early predictive skill. It also bridges the prior work showing a predisposition for turn taking in infancy (e.g., Hilbrink et al., 2015) but late acquisition of adult-like competence when it comes to integrating linguistic information into turn-taking behaviors (Casillas et al., In press; Garvey, 1984; Ervin-Tripp, 1979).

4.2. 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, resulting 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 almost none of them have done so in a conversational context. Future work could instead carefully script or cross-splice parts of turns to control for the presence or absence of linguistic cues for turn transition.

The prediction measure used in these studies is based on an observer's view of third-party conversation but, because participants' role in the interaction could affect their online predictions about turn taking, an ideal experimental measure would capture first-person behavior. First-person measures of spontaneous turn prediction will be key to revealing how participants distribute their attention over linguistic 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.

4.3. 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 is not integrated with linguistic cues until much later. Using language to make predictions about upcoming interactive content takes time and, for both children and adults, is primarily driven by participants' orientation to what will happen beyond the end of the current turn.

Acknowledgements

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

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. URL: `https://github.com/lme4/lme4/http://lme4.r-forge.r-project.org/`. [Computer program] R package version 1.1-7.

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

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

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

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

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

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., In press. Turn taking, timing, and planning in early language acquisition. Journal of Child Language .

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

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

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

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

De Vos, C., Torreira, F., Levinson, S.C., 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.

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

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

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

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

Gísladóttir, R., Chwilla, D., Levinson, S.C., 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.

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

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

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

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

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

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

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

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

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

Magyari, L., Bastiaansen, M.C.M., De Ruiter, J.P., Levinson, S.C., 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.

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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thorgrímsson, G., Fawcett, C., Liszkowski, U., 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: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.

Torreira, F., Bögels, S., Levinson, S.C., 2015. Intonational phrasing is necessary for turn-taking in spoken interaction. Journal of Phonetics 52, 46–57.

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

Appendix A. Permutation Analyses

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.[10] 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.

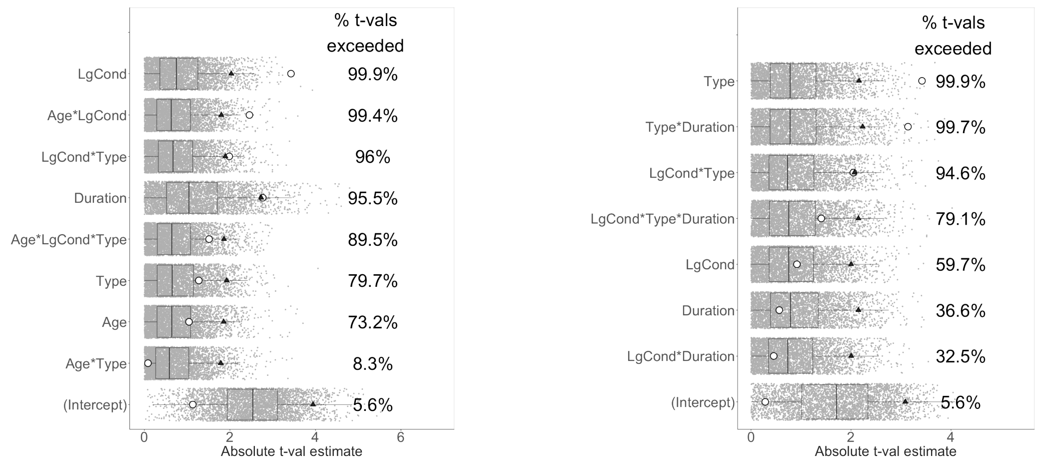
E1: We excluded the output of random-permutation models that gave convergence warnings in order to remove unreliable estimates from our analyses (non-converging models were 22.4% and 24.4% of all models for children and adults, respectively; see the Supplementary Materials for more information on model exclusion).

E2: As before, we excluded the output of random-permutation models that resulted in convergence warnings in order to remove unreliable model estimates from our analyses (non-converging models made up 69% and 70% of models for children and adults, respectively; see the Supplementary Materials for more information on model exclusion).

In all of the following plots, the gray dots represent the randomly permuted data's model estimates for the value listed (beta or standard error), the white dots represent the model estimates from the original data, and the triangles represent the 95th percentile for each distribution being shown.

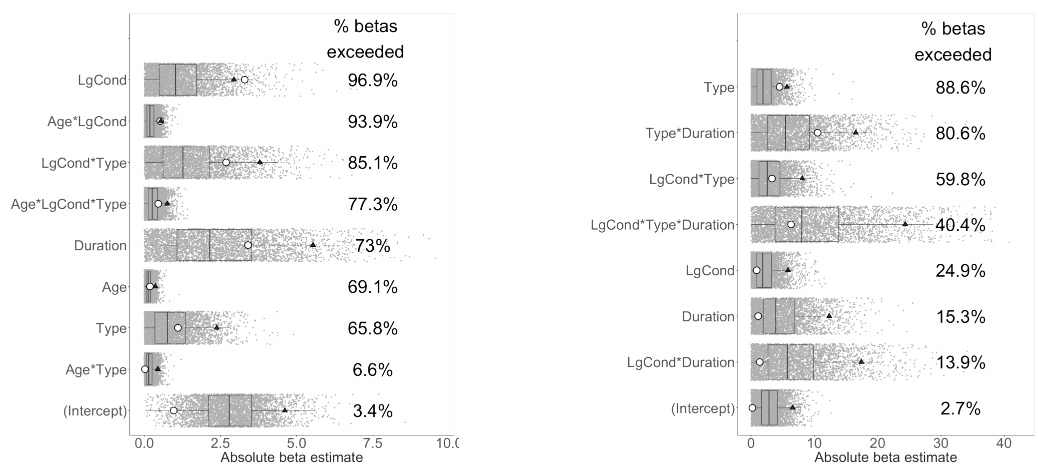
[10] We report t-values rather than beta estimates because the standard errors in the randomly permuted data models were much higher than for the original data. For those interested, plots of the beta and standard error distributions are available in the Supplementary Materials.

Appendix A.1. Experiment 1



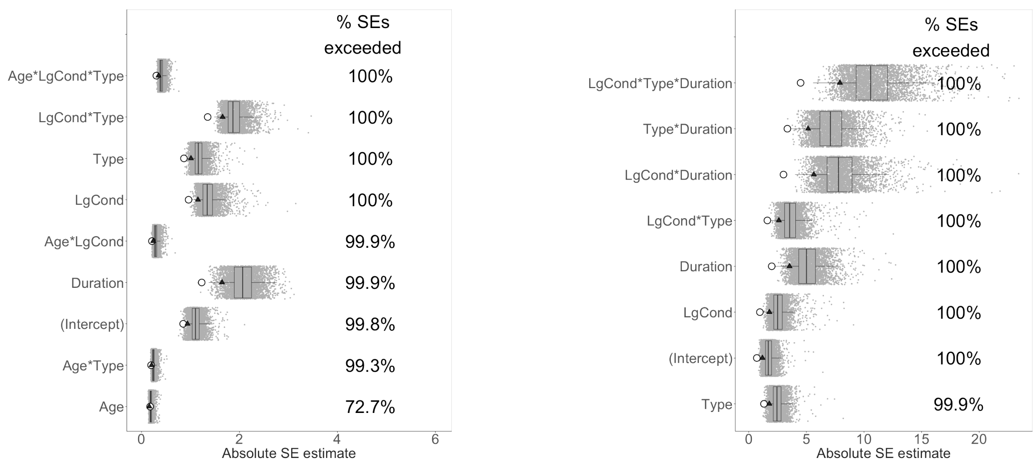
(a) Children (b) Adults

Figure A.1: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.



(a) Children (b) Adults

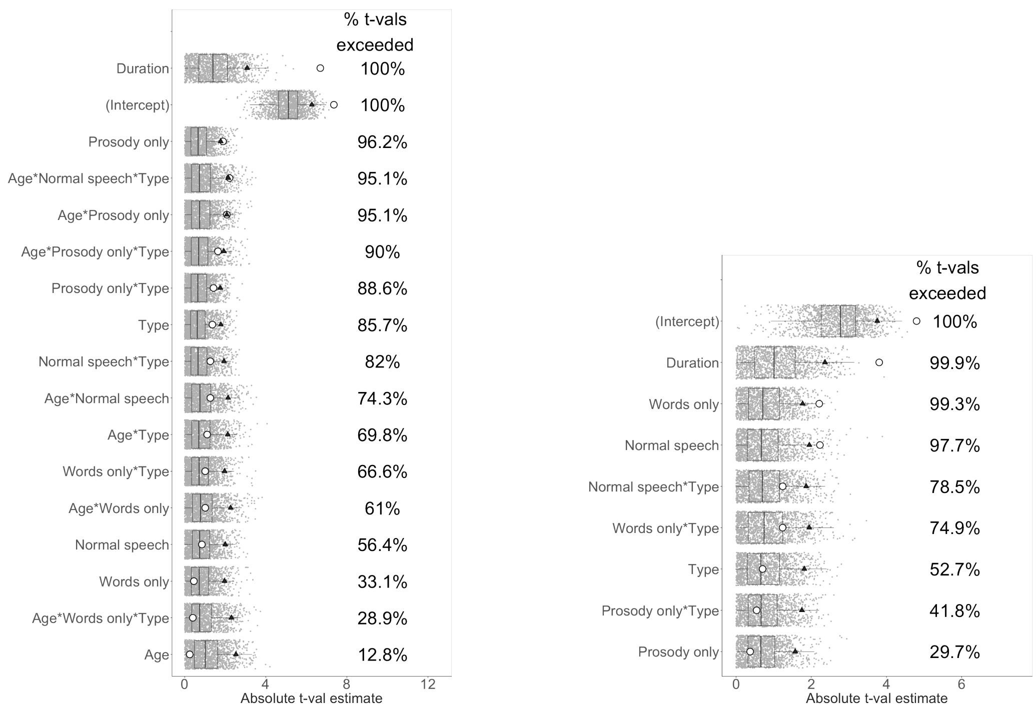
Figure A.2: Random-permutation and original $|\beta\text{-values}|$ for predictors of gaze rates in Experiment 1.



(a) Children (b) Adults

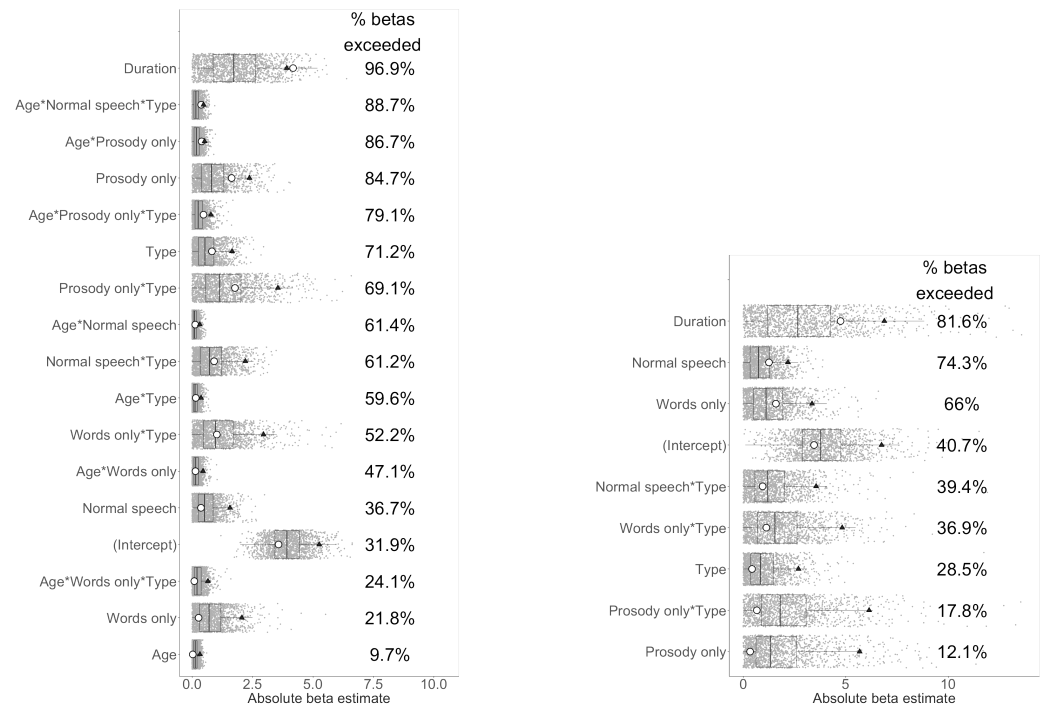
Figure A.3: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.

Appendix A.2. Experiment 2



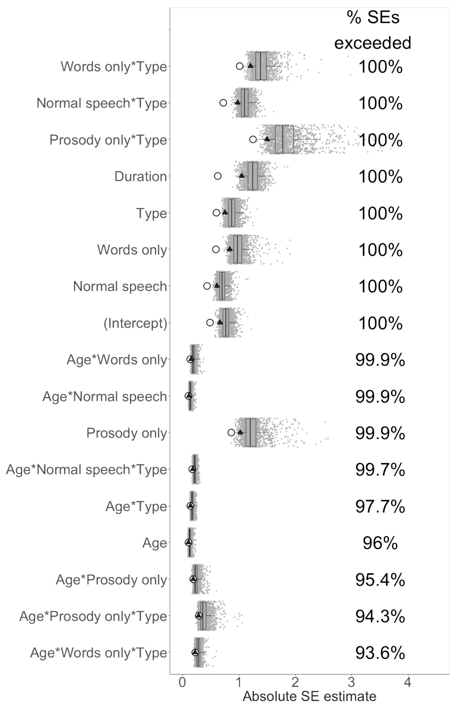
(a) Children (b) Adults

Figure A.4: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

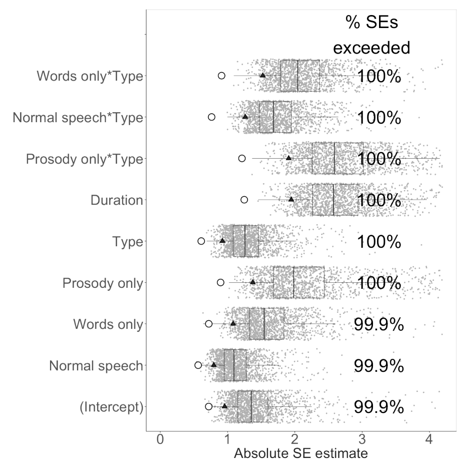


(a) Children (b) Adults

Figure A.5: Random-permutation and original $|\beta\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.



(a) Children



(b) Adults

Figure A.6: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

Appendix B. Pairwise developmental tests: Real vs. randomly permuted effects

In each of the plots below, the dot represents the original data value for the effect and the 5,000 randomly permuted data effect sizes are shown in the distribution. The percentage shown is the percentage of random permutation values exceeded by the original data value (taking the absolute value of all data points for a two-tailed test.)

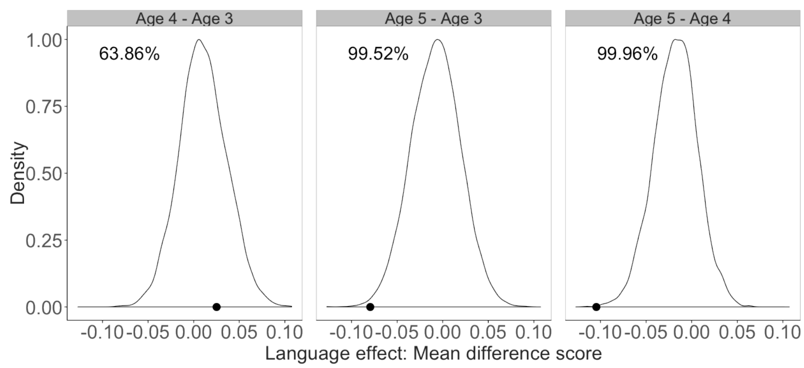


Figure B.1: Pairwise comparisons of the language condition effect across ages in Experiment 1.

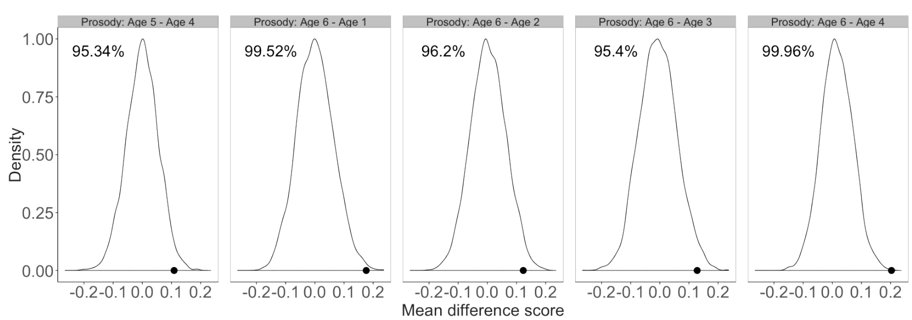


Figure B.2: Significant pairwise comparisons of the *prosody only*-*no speech* linguisitic condition effect, across ages in Experiment 2

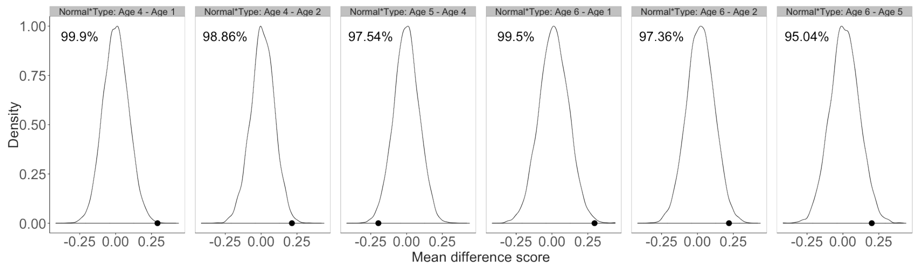


Figure B.3: Significant pairwise comparisons of the *normal speech-no speech* language condition effect for transition type, across ages, in Experiment 2.

Appendix C. Non-convergent models

Non-convergent models made up 22–24% of the 5,000 models of randomly permuted data in Experiment 1 and 69–70% of the 5,000 for Experiment 2. We excluded these non-convergent models because they displayed erratic β and SE estimates, as summarized below in the table of t-values from convergent and non-convergent models in Experiment 1. The non-convergent models from Experiment 2 showed similar patterns. The high frequency of problematic models persisted even when we changed optimizers and we suspect the issue derives from data sparsity in some of the random runs.

Variable	Mean	Median	*SD*	Minimum	Maximum
Children					
(Intercept)	-2.52 (-458.42)	-2.54 (-2.86)	0.87 (1319.22)	-5.53 (-8185.36)	0.41 (0.97)
Age	-0.51 (-17.83)	-0.49 (-0.53)	0.79 (83.78)	-3.71 (-672.2)	2.3 (342.8)
LgCond	-0.53 (-109.91)	-0.55 (-0.63)	0.93 (564.42)	-3.93 (-4418.74)	3.23 (2296.19)
Type	-0.10 (-29.66)	-0.09 (-0.1)	0.98 (515.12)	-4.06 (-4383.92)	3.23 (2296.19)
Duration	0.99 (345.53)	0.98 (1.15)	1.07 (1323.13)	-2.44 (-5048.24)	3.36 (3416.68)
Age*LgCond	0.19 (10.64)	0.2 (0.18)	0.9 (109.6)	-3.31 (-581.61)	5.78 (9985.16)
Age*Type	0.02 (-1.8)	0.001 (-0.05)	0.9 (98.27)	-3.36 (-884.36)	3.59 (946.81)
LgCond*Type	0.2 (45.32)	0.2 (0.27)	0.96 (691.3)	-3.12 (-4160.06)	3.45 (640.43)
Age*LgCond*Type	-0.12 (-14.23)	-0.12 (-0.15)	0.93 (156.72)	-2.98 (-1318.26)	3.39 (5107.64)
Adults					
(Intercept)	-1.63 (-126.14)	-1.71 (-1.73)	0.97 (713.39)	-4.08 (-12111.22)	2.15 (649.55)
LgCond	-0.26 (-679.6)	-0.3 (-0.53)	1.02 (15894.33)	-3.45 (-494979.7)	3.35 (88581.58)
Type	-0.11 (6.29)	-0.13 (-0.04)	1.11 (501.5)	-3.85 (-6420.76)	3.28 (8177.88)
Duration	0.25 (84.09)	0.27 (0.26)	1.1 (1152.94)	-3.25 (-10864.51)	3.46 (18540.62)
LgCond*Type	0.12 (-242.27)	0.1 (0.34)	1.07 (26836.7)	-3.41 (-62264.27)	3.81 (509198.4)
LgCond*Duration	0.15 (780.03)	0.16 (0.39)	1.04 (44105.02)	-3.84 (-798498.6)	3.55 (1145951)
Type*Duration	0.05 (-6.56)	0.05 (0.02)	1.13 (1389.9)	-3.54 (-15979.22)	3.87 (16419.46)
LgCond*Type*Duration	-0.06 (1083.63)	-0.08 (-0.21)	1.1 (63116.54)	-4.21 (-1201895)	4.02 (1284965)

Table C.1: Estimated t-values for each predictor in in the adult and child models from Experiment 1. Converging ('C') and non-converging ('non-C') values are shown as: 'C (non-C)'.

Appendix D. Miscellaneous

One alternative hypothesis for children's anticipatory gazes is that they simply grow bored and start looking away at a constant rate after a turn begins. This data plotted here show a hypothetical group of participants who begin to lose interest (at a linear rate) after one second of a turn (gray dots) compared to participants' real data from Experiment 2 (black dots). This pattern suggests that, though children do look away with time, their looks away are not simply driven by boredom.

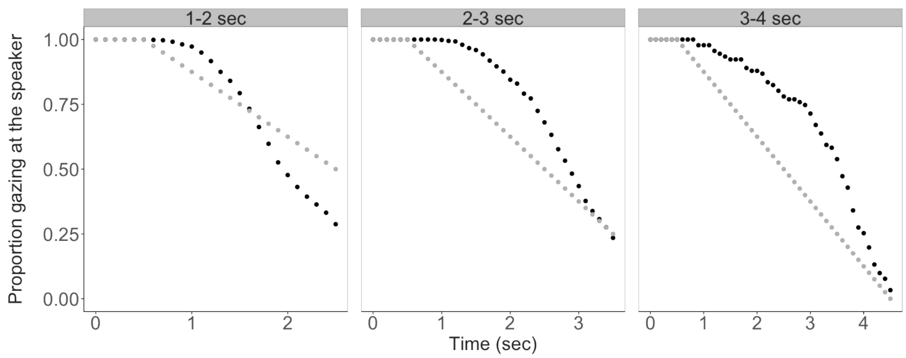


Figure D.1: Proportion participants looking at the current speaker: hypothetical boredom-driven data (gray dots) versus real data from Experiment 2 (black dots).

The design for Experiment 2 does not fully cross puppet pair (e.g., robots, blue puppets) with linguistic condition (e.g., "words only" and "no speech"). Even though each puppet pair is associated with different conversation clips across children (e.g. robots talking about kitties, birthday parties, or pancakes), robots were only associated with "words only" speech, merpeople were only associated with "prosody only" speech, and the puppets with fancy clothes were only associated with the "no speech" condition. We did this to increase the pragmatic felicity of the experiments for the older children (i.e., robots make robot sounds, merpeople's voices are muffled under water, the fancy-clothed puppets are in a room with main other voices). It is therefore fair to point out a possible confound between linguistic condition and puppet pair. Thankfully, we also ran a short follow-up study at the museum with 3–5-year-olds in which each child only saw one video—the normal speech conversation about birthday parties—with a randomly assigned puppet pair performing the conversation. Five children watched each puppet pair, for a total of 30 children across the six pairs. This experiment holds all things constant except for the appearance of the puppets. We then used a mixed effects logistic regression of children's anticipatory switches (yes or no at each transition), with puppet pair (robots/merpeople/fancy dress/other-3; Figure 1) as a fixed effect and participant and turn transition as random effects. In four versions of this model we systematically varied the reference level to check for differences between every puppet pair, finding no significant affects of puppet type on switching rate. We take this as evidence that, although we did not fully cross puppet pairs and linguistic conditions in Experiment 2, it was unlikely to have had strong effects on children's looking rates above and beyond the intended effects of linguistic condition.

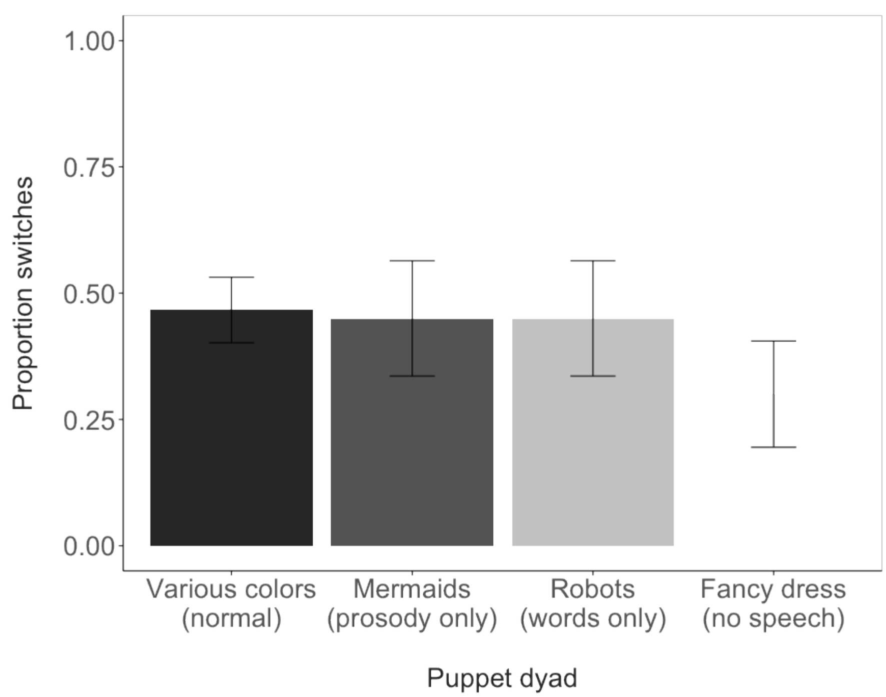


Figure D.2: Proportion gaze switches across puppet pairs when linguistic condition and conversation are held constant.