

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 participants, finding minimal differences between the predictive looking behavior of preschoolers and adults. In Experiment 2, we phonetically controlled lexicosyntactic and prosodic cues in English-only speech, finding that children's predictive looking behavior improved from ages one to six, but that even one-year-olds made more anticipatory looks than would be expected by chance. In both experiments, children and adults anticipated more often after hearing questions. Like adults, prosody alone

*Corresponding author

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 takes several years before becoming fully integrated with linguistic processing.

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

¹ 1. Introduction

² Spontaneous conversation is a universal context for using and learning
³ language. Like other types of human interaction, it is organized at its core
⁴ by the roles and goals of its participants. But what sets conversation apart is
⁵ its structure: Sequences of interconnected, communicative actions that take
⁶ place across alternating turns at talk. Sequential, turn-based structures in
⁷ conversation are strikingly uniform across language communities and linguis-
⁸ tic modalities. Turn-taking behaviors are also cross-culturally consistent in
⁹ their basic features and the details of their implementation (De Vos et al.,
¹⁰ 2015; Dingemanse et al., 2013; Stivers et al., 2009). How does this ability
¹¹ develop?

¹² Children participate in sequential coordination with their caregivers start-
¹³ ing at three months of age—before they can rely on any linguistic cues in
¹⁴ taking turns (see, among others, Bateson, 1975; Hilbrink et al., 2015; Jaffe
¹⁵ et al., 2001; Snow, 1977). Of course, infant turn taking is different from
¹⁶ adult turn taking in several ways. Infant turn taking is heavily scaffolded by
¹⁷ caregivers, has different timing from adult turn taking, and lacks semantic
¹⁸ content (Hilbrink et al., 2015; Jaffe et al., 2001). However, children’s early,
¹⁹ turn-structured social interactions are presumably a critical precursor to their
²⁰ later conversational turn taking: Early non-verbal interactions likely estab-
²¹ lish the protocol by which children come to use language with others. How
²² do children integrate linguistic knowledge with these preverbal turn-taking
²³ abilities?

²⁴ In this study, we investigate when children begin to make predictions
²⁵ about upcoming turn structure in conversation, and how they integrate lan-
²⁶ guage into their predictions as they grow older. In what follows, we first give
²⁷ a basic review of turn-taking research and the state of current knowledge
²⁸ about adult turn prediction. We then discuss recent work on the develop-
²⁹ ment of turn-taking skills before turning to the details of our own study.

³⁰ *1.1. Adults' turn taking*

³¹ Turn taking itself is not unique to conversation. Many other human activi-
³² ties are organized around sequential turns at action. Traffic intersections and
³³ computer network communication both use turn-taking systems. Children's
³⁴ early games (e.g., give-and-take, peek-a-boo) have built-in, predictable turn
³⁵ structure (Ratner and Bruner, 1978; Ross and Lollis, 1987). Even monkeys
³⁶ take turns: Non-human primates such as marmosets and Campbell's mon-
³⁷ keys vocalize contingently with each other in both natural and lab-controlled
³⁸ environments (Lemasson et al., 2011; Takahashi et al., 2013). In all these
³⁹ cases, turn taking serves as a protocol for interaction, allowing the partici-
⁴⁰ pants to coordinate with one another through sequences of contingent action.

⁴¹ Conversation distinguishes itself from non-conversational turn-taking be-
⁴² haviors by the complexity of the turn sequencing involved. In the examples
⁴³ above (traffic, games, and monkeys) the set of sequence and action types is
⁴⁴ far more limited and predictable than what we find in everyday talk. Con-
⁴⁵ versational turns come grouped into semantically-contingent sequences of
⁴⁶ action. The groups can span turn-by-turn exchanges (e.g., simple question-
⁴⁷ response, "How are you?"—"Fine.") or sequence-by-sequence exchanges (e.g.,
⁴⁸ reciprocals, "How are you?"—"Fine, and you?"—"Great!").

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

63 Adults have a lot of information at their disposal to help make accurate
64 predictions about upcoming turn content. Lexical, syntactic, and prosodic
65 information (e.g., *wh*- words, subject-auxiliary inversion, and list intonation)
66 can all inform addressees about upcoming linguistic structure (De Ruiter
67 et al., 2006; Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015).

68 Non-verbal cues (e.g., gaze, posture, and pointing) often appear at turn-
69 boundaries and can sometimes act as late indicators of an upcoming speaker
70 switch (Rossano et al., 2009; Stivers and Rossano, 2010). Additionally, the
71 sequential context of a turn can make it clear what will come next: An-
72 swers after questions, thanks or denial after compliments, et cetera (Schegloff,
73 2007).

74 Prior work suggests that adult listeners primarily use lexicosyntactic in-
75 formation to accurately predict upcoming turn structure (De Ruiter et al.,
76 2006). De Ruiter and colleagues (2006) asked participants to listen to snip-
77 pets of spontaneous conversation and to press a button whenever they antici-
78 pated that the current speaker was about to finish his or her turn. The speech
79 snippets were controlled for the amount of linguistic information present;
80 some were normal, but others had flattened pitch, low-pass filtered speech,
81 or further manipulations. With pitch-flattened speech, the timing of par-
82 ticipants' button responses was comparable to their timing with the full
83 linguistic signal. But when no lexical information was available, partici-
84 pants' responses were significantly earlier. The authors concluded that lex-
85 icosyntactic information¹ was necessary and possibly sufficient for turn-end

¹The “lexicosyntactic” condition only included flattened pitch and so was not exclu-

⁸⁶ projection, while intonation was neither necessary nor sufficient. Congru-
⁸⁷ ent evidence comes from studies varying the predictability of lexicosyntactic
⁸⁸ and pragmatic content: Adults anticipate turn ends better when they can
⁸⁹ more accurately predict the exact words that will come next (Magyari and
⁹⁰ De Ruiter, 2012; see also Magyari et al., 2014). They can also identify speech
⁹¹ acts within the first word of an utterance (Gísladóttir et al., 2015), allowing
⁹² them to start planning their response at the first moment possible (Bögels
⁹³ et al., 2015).

⁹⁴ Despite this evidence, the role of prosody for adult turn prediction is
⁹⁵ still a matter of debate. De Ruiter and colleagues' (2006) experiment fo-
⁹⁶ cused on the role of intonation, which is only a partial index of prosody.
⁹⁷ Prosodic structure is also tied closely to the syntax of an utterance, and
⁹⁸ so the two linguistic signals are difficult to control independently (Ford and
⁹⁹ Thompson, 1996). Torreira, Bögels and Levinson (2015) used a combination
¹⁰⁰ of button-press and verbal responses to investigate the relationship between
¹⁰¹ lexicosyntactic and prosodic cues in turn-end prediction. Critically, their
¹⁰² stimuli were cross-spliced so that each item had full prosodic cues to accom-

sively lexicosyntactic—the speech would still have residual prosodic structure, including syllable duration and intensity.

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

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:

₁₂₂ They take every look, vocalization, arm flail, and burp as “utterances” in the
₁₂₃ joint discourse (Bateson, 1975; Jaffe et al., 2001; Snow, 1977). Infants catch
₁₂₄ onto the structure of proto-conversations quickly. By three to four months
₁₂₅ they notice disturbances to the contingency of their caregivers’ response and,
₁₂₆ in reaction, change the rate and quality of their vocalizations (Bloom, 1988;
₁₂₇ Masataka, 1993).

₁₂₈ The timing of children’s responses to their caregivers’ speech shows a
₁₂₉ non-linear pattern. Infants’ contingent vocalizations in the first few months
₁₃₀ of life show very fast timing (though with a lot of vocal overlap) that, by nine
₁₃₁ months, slows down considerably, only gradually speeding up again after 12
₁₃₂ months (Hilbrink et al., 2015). Taking turns with brief transitions between
₁₃₃ speakers is difficult for children; while their avoidance of overlap is nearly
₁₃₄ adult-like by nine months, the timing of their non-overlapped responses stays
₁₃₅ much longer than the 200 msec standard for the next few years (Casillas
₁₃₆ et al., In press; Garvey, 1984; Ervin-Tripp, 1979). This puzzling pattern is
₁₃₇ likely due to their linguistic development: Taking turns on time is easier
₁₃₈ when the response is a simple vocalization rather than a linguistic utterance.
₁₃₉ Integrating language into the turn-taking system may be one major factor in
₁₄₀ children’s delayed responses (Casillas et al., In press).

¹⁴¹ While children, like adults, could use linguistic cues in the ongoing turn to
¹⁴² make predictions about upcoming turn structure, studies of early linguistic
¹⁴³ development point to a possible early advantage for prosody over lexicosyn-
¹⁴⁴ tax in children's turn-taking predictions. Infants can distinguish their native
¹⁴⁵ language's rhythm type from others soon after birth (Mehler et al., 1988;
¹⁴⁶ Nazzi and Ramus, 2003); they show preference for the typical stress pat-
¹⁴⁷ terns of their native language over others by 6–9 months (e.g., iambic vs.
¹⁴⁸ trochaic), and can use prosodic information to segment the speech stream
¹⁴⁹ into smaller chunks from 8 months onward (Johnson and Jusczyk, 2001;
¹⁵⁰ Morgan and Saffran, 1995). Four- to five-month-olds also prefer pauses in
¹⁵¹ speech to be inserted at prosodic boundaries, and by 6 months they can start
¹⁵² using prosodic markers to pick out sub-clausal syntactic units, both of which
¹⁵³ are useful for extracting turn structure from ongoing speech (Jusczyk et al.,
¹⁵⁴ 1995; Soderstrom et al., 2003). In comparison, children show at best a very
¹⁵⁵ limited lexical inventory before their first birthday (Bergelson and Swingley,
¹⁵⁶ 2013; Shi and Melancon, 2010).

¹⁵⁷ Keitel and colleagues (2013) were one of the first to explore how children
¹⁵⁸ use linguistic cues to predict upcoming turn structure. They asked 6-, 12-, 24-
¹⁵⁹, 36-month-old, and adult participants to watch short videos of conversation

160 and tracked their eye movements at points of speaker change. They showed
161 their participants two types of conversation videos—one normal and one with
162 flattened pitch (i.e., with flattened intonation contours)—to test the role of
163 intonation in participants’ anticipatory predictions about upcoming speech.
164 Comparing children’s anticipatory gaze frequency to a random baseline, they
165 found that only 36-month-olds and adults made anticipatory gaze switches
166 more often than expected by chance. Among those, only 36-month-olds were
167 affected by a lack of intonation contours, leading Keitel and colleagues to
168 conclude that children’s ability to predict upcoming turn structure relies on
169 their ability to comprehend the stimuli lexicosemantically. They also suggest
170 that intonation takes a secondary role in turn prediction, following their
171 and other studies findings with adults (e.g., De Ruiter et al., 2006), but only
172 *after* children acquire more sophisticated, adult-like language comprehension
173 systems (sometime after 36 months in their data).

174 Although the Keitel et al. (2013) study constitutes a substantial advance
175 over previous work in this domain, it has its own limitations. Because these
176 limitations directly inform our own study design, we review them in some
177 detail. First, their estimates of baseline gaze frequency (“random” in their
178 terminology) were not random. Instead, they used gaze switches during

¹⁷⁹ ongoing speech as a baseline. Ongoing speech is perhaps the period in which
¹⁸⁰ switching is least likely to occur (Hirvenkari et al., 2013), thus maximizing
¹⁸¹ chances of finding a difference between gaze frequency at turn transitions
¹⁸² and their baseline rate. A more conservative baseline would be to compare
¹⁸³ participants' looking behavior at turn transitions to their looking behavior
¹⁸⁴ during randomly selected windows of time throughout the stimulus, including
¹⁸⁵ turn transitions. We follow this conservative approach in our work.

¹⁸⁶ Second, the conversation stimuli Keitel et al. (2013) used were somewhat
¹⁸⁷ unusual. The average gap between turns was 900 msec, which is much longer
¹⁸⁸ than typical adult timing, where gaps average around 200 msec (Stivers et al.,
¹⁸⁹ 2009). The speakers in the videos were also asked to minimize their move-
¹⁹⁰ ments while performing a scripted and adult-directed conversation, which
¹⁹¹ would have created a somewhat unnatural stimulus. Additionally, in order
¹⁹² to produce more naturalistic conversation, it would have been ideal to local-
¹⁹³ ize the sound sources for the two voices in the video (i.e., to have the voices
¹⁹⁴ come out of separate left and right speakers). But both voices were recorded
¹⁹⁵ and played back on the same audio channel, which may have made it more
¹⁹⁶ difficult to distinguish the two talkers (again, we attempt to address these
¹⁹⁷ issues in our current study). Despite these minor methodological issues, the

¹⁹⁸ Keitel et al. (2013) study still demonstrates intriguing age-based differences
¹⁹⁹ in children's ability to predict upcoming turn structure. Our current work
²⁰⁰ thus takes this paradigm as a starting point.²

²⁰¹ Our goal in the current study is to find out when children begin to make
²⁰² predictions about upcoming turn structure and to understand how their pre-
²⁰³ dictions are affected by linguistic cues across development. We present two
²⁰⁴ experiments in which we measure children's anticipatory gaze to respon-
²⁰⁵ ders while watching conversation videos with natural (people using English
²⁰⁶ vs. non-English; Experiment 1) and non-natural (puppets with phonetically
²⁰⁷ manipulated speech; Experiment 2) control over the presence of lexical and
²⁰⁸ prosodic cues. We tested children across a wide range of ages (Experiment
²⁰⁹ 1: 3–5 years; Experiment 2: 1–6 years), with adult control participants in
²¹⁰ each experiment.

²¹¹ Our findings suggest that children and adults use linguistic cues to make
²¹² predictions about upcoming turn structure, but that they primarily do so to
²¹³ predict speaker transitions after a question. We also find that children make
²¹⁴ more predictions than expected by chance at age two, but that it takes several
²¹⁵ years before they recruit linguistic cues to anticipate responses more often

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

₂₁₆ after questions. We found no direct evidence of an early prosody advantage
₂₁₇ in children's anticipations and, further, no evidence that prosodic or lexical
₂₁₈ cues alone can substitute for their combination in the full linguistic signal,
₂₁₉ as is proposed for adults (De Ruiter et al., 2006).

₂₂₀ **2. Experiment 1**

₂₂₁ We recorded participants' eye movements as they watched six short videos
₂₂₂ of two-person (dyadic) conversation interspersed with attention-getting filler
₂₂₃ videos. Each conversation video featured an improvised discourse in one of
₂₂₄ five languages (English, German, Hebrew, Japanese, and Korean); partici-
₂₂₅ pants saw two videos in English and one in every other language. The partic-
₂₂₆ ipants, all native English speakers, were only expected to understand the two
₂₂₇ videos in English. We showed participants non-English videos to limit their
₂₂₈ access to lexical information while maintaining their access to other cues to
₂₂₉ turn boundaries (e.g., (non-native) prosody, gaze, breath, phrase final length-
₂₃₀ ening). Using this method, we compared children and adult's anticipatory
₂₃₁ looks from the current speaker to the upcoming speaker at points of turn
₂₃₂ transition in English and non-English videos.



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

233 *2.1. Methods*

234 *2.1.1. Participants*

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

244 *2.1.2. Materials*

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

253 Each recording session began with a 20-minute warm-up period of spon-
254 taneous conversation during which the pair talked for five minutes on four
255 topics (favorite foods, entertainment, hometown layout, and pets). Then we
256 asked talkers to choose a new topic—one relevant to young children (e.g.,
257 riding a bike, eating breakfast)—and to improvise a dialogue on that topic.
258 We asked them to speak as if they were on a children’s television show in
259 order to elicit child-directed speech toward each other. We recorded until the
260 talkers achieved at least 30 seconds of uninterrupted discourse with enthu-
261 siastic, child-directed speech. Most talker pairs took less than five minutes
262 to complete the task, usually by agreeing on a rough script at the start. We

²⁶³ encouraged talkers to ask at least a few questions to each other during the
²⁶⁴ improvisation. The resulting conversations were therefore not entirely spon-
²⁶⁵ taneous, but were as close as possible while still remaining child-oriented in
²⁶⁶ topic, prosodic pattern, and lexicosyntactic construction.³

²⁶⁷ After recording, we combined the audio and video files by hand, and
²⁶⁸ cropped each recording to the 30-second interval with the most turn activity.
²⁶⁹ Because we recorded the conversations in stereo, the male and female voices
²⁷⁰ came out of separate speakers during video playback. This gave each voice in
²⁷¹ the videos a localized source (from the left or right loudspeaker). We coded
²⁷² each turn transition in the videos for language condition (English vs. non-
²⁷³ English), inter-turn gap duration (in milliseconds), and speech act (question
²⁷⁴ vs. non-question). The non-English stimuli were coded for speech act from
²⁷⁵ a monolingual English-speaker’s perspective, i.e., which turns “sound like”
²⁷⁶ questions, and which don’t: We asked five native American English speakers
²⁷⁷ to listen to the audio signal for each turn and judge whether it sounded
²⁷⁸ like a question. We then coded turns with at least 80% “yes” responses as

³All 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.

279 questions.

280 Because the conversational stimuli were recorded semi-spontaneously, the
281 duration of turn transitions and the number of speaker transitions in each
282 video was variable. We measured the duration of each turn transition from
283 the audio recording associated with each video. We excluded turn tran-
284 sitions longer than 550 msec and shorter than 90 msec, including over-
285 lapped transitions, from analysis.⁴ This left approximately equal numbers
286 of turn transitions available for analysis in the English (N=20) and non-
287 English (N=16) videos. On average, the inter-turn gaps for English videos
288 (mean=318, median=302, stdev=112 msec) were slightly longer than for non-
289 English videos (mean=286, median=251, stdev=122 msec). The longer gaps
290 in the English videos could give them a slight advantage: Our definition of
291 an “anticipatory gaze shift” includes shifts that are initiated during the gap

⁴Overlap 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.

292 between turns (Figure 2), so participants had slightly more time to make
293 anticipatory shifts in the English videos.

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

301 *2.1.3. Procedure*

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

308 Before the experiment started, we warned adult participants that they
309 would see videos in several languages and that, though they weren't expected
310 to understand the content of non-English videos, we *would* ask them to an-

311 swer general, non-language-based questions about the conversations. Then
312 after each video we asked participants one of the following randomly-assigned
313 questions: “Which speaker talked more?”, “Which speaker asked the most
314 questions?”, “Which speaker seemed more friendly?”, and “Did the speak-
315 ers’ level of enthusiasm shift during the conversation?” We also asked if the
316 participants could understand any of what was said after each video. The
317 participants responded verbally while an experimenter noted their responses.

318 Children were less inclined to simply sit and watch videos of conversation
319 in languages they didn’t speak, so we used a different procedure to keep them
320 engaged: The experimenter started each session by asking the child about
321 what languages he or she could speak, and about what other languages he
322 or she had heard of. Then the experimenter expressed her own enthusiasm
323 for learning about new languages, and invited the child to watch a video
324 about “new and different languages” together. If the child agreed to watch,
325 the experimenter and the child sat together in front of the display, with
326 the child centered in front of the tracker and the experimenter off to the
327 side. Each conversation video was preceded and followed by a 15–30 second
328 attention-getting filler video (e.g., running puppies, singing muppets, flying
329 bugs). If the child began to look bored, the experimenter would talk during

330 the fillers, either commenting on the previous conversation (“That was a neat
331 language!”) or giving the language name for the next conversation (“This
332 next one is called Hebrew. Let’s see what it’s like.”) The experimenter’s
333 comments reinforced the video-watching as a joint task.

334 All participants (child and adult) completed a five-point calibration rou-
335 tine before the first video started. We used a dancing Elmo for the children’s
336 calibration image. During the experiment, participants watched all six 30-
337 second conversation videos. The first and last conversations were in American
338 English and the intervening conversations were Hebrew, Japanese, German,
339 and Korean. The presentation order of the non-English videos was shuffled
340 into four lists, which participants were assigned to randomly. The entire
341 experiment, including instructions, took 10–15 minutes.

342 *2.1.4. Data preparation and coding*

343 To determine whether participants predicted upcoming turn transitions,
344 we needed to define a set of criteria for what counted as an anticipatory gaze
345 shift. Prior work using similar experimental procedures has found that adults
346 and children make anticipatory gaze shifts to upcoming talkers within a wide
347 time frame; the earliest shifts occur before the end of the prior turn, and the
348 latest occur after the onset of the response turn, with most shifts occurring

349 in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz,
350 2011). Following prior work, we measured how often our participants shifted
351 their gaze from the prior to the upcoming speaker *before* the shift in gaze
352 could have been initiated in reaction to the onset of the speaker's response.
353 In doing so, we assumed that it takes participants 200 msec to plan an eye
354 movement, following standards from adult anticipatory processing studies
355 (e.g., Kamide et al., 2003).

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

368 As mentioned, most anticipatory switches happen in the inter-turn gap,
369 but we also allowed anticipatory gaze switches that occurred in the final
370 syllables of the prior turn. Early switches are consistent with the distribution
371 of responses in explicit turn-boundary prediction tasks. For example, in
372 a button press task, adult participants anticipate turn ends approximately
373 200 msec in advance of the turn's end, and anticipatory responses to pitch-
374 flattened stimuli come even earlier (De Ruiter et al., 2006). We therefore
375 allowed switches to occur as early as 200 msec before the end of the prior turn.
376 For very early and very late switches, our requirement for 100 msec of fixation
377 on each speaker would sometimes extend outside of the transition window
378 boundaries (200 msec before and after the inter-turn gap). The maximally
379 available fixation window was 100 msec before and after the earliest and
380 latest possible switch point (300 msec before and after the inter-turn gap).
381 We did not count switches made during the fixation window as anticipatory.
382 We *did* count switches made during the inter-turn gap. The period of time
383 from the beginning of the possible fixation window on the prior speaker to the
384 end of the possible fixation window on the responder was our total analysis
385 window (300 msec + the inter-turn gap + 300 msec).

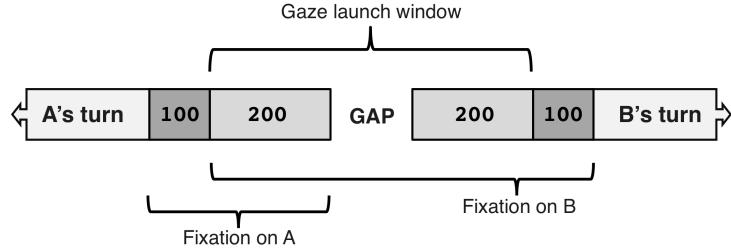


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition.

³⁸⁶ *Predictions.* We expected participants to show greater anticipation in the
³⁸⁷ English videos than in the non-English videos because of their increased
³⁸⁸ access to linguistic information in English. We also predicted that anticipa-
³⁸⁹ tion would be greater following questions compared to non-questions; ques-
³⁹⁰ tions have early cues to upcoming turn transition (e.g., *wh-* words, subject-
³⁹¹ auxiliary inversion), and also make a next response immediately relevant.
³⁹² Our third prediction was that anticipatory looks would increase with devel-
³⁹³ opment, along with children’s increased linguistic competence.

³⁹⁴ *2.2. Results*

³⁹⁵ Participants looked at the screen most of the time during video playback
³⁹⁶ (81% and 91% on average for children and adults, respectively). They pri-
³⁹⁷ marily kept their eyes on the person who was currently speaking in both

398 English and non-English videos: They gazed at the current speaker between
399 38% and 63% of the time, looking back at the addressee between 15% and
400 20% of the time (Table 1). Even three-year-olds looked more at the current
401 speaker than anything else, whether the videos were in a language they could
402 understand or not. Children looked at the current speaker less than adults
403 did during the non-English videos. Despite this, their looks to the addressee
404 did not increase substantially in the non-English videos, indicating that their
405 looks away were probably related to boredom rather than confusion about
406 ongoing turn structure. Overall, participants' pattern of gaze to current
407 speakers demonstrated that they performed basic turn tracking during the
408 videos, regardless of language. Figure 3 shows participants' anticipatory gaze
409 rates across age, language condition, and transition type, in the real data and
410 the random baseline (described below).

411 *2.2.1. Statistical models*

412 We identified anticipatory gaze switches for all 36 usable turn transitions,
413 based on the criteria outlined in Section 2.1.4, and analyzed them for effects
414 of language, transition type, and age with two mixed-effects logistic regres-
415 sions (Bates et al., 2014; R Core Team, 2014). We built one model each
416 for children and adults. We modeled children and adults separately because

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.

⁴¹⁷ effects of age are only pertinent to the children’s data. The child model

⁴¹⁸ included condition (English vs. non-English)⁵, transition type (question vs.

⁵Because each non-English language was represented by a single stimulus, we cannot treat individual languages as factors. Gaze behavior might be best for non-native languages that have the most structural overlap with participants’ native language: English speakers can make predictions about the strength of upcoming Swedish prosodic boundaries nearly as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same

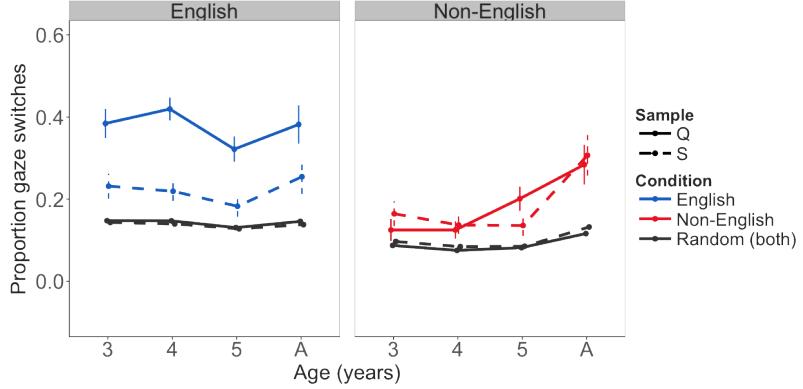


Figure 3: Anticipatory gaze rates across language condition and transition type for the real (red and blue) and randomly permuted baseline (gray). Vertical bars represent the standard error.

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,

task (Carlson et al., 2005). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

425 and their interaction for participants (Barr et al., 2013).⁶ The adult model
426 included condition, transition type, duration, and their interactions as pre-
427 dictors with participant and item included as random effects and random
428 slopes of condition, transition type, and their interaction for participant.

429 Children's anticipatory gaze switches showed effects of language condition
430 ($\beta=-3.29$, $SE=0.961$, $t=-3.43$, $p<.001$) and gap duration ($\beta=3.4$, $SE=1.229$,
431 $t=2.77$, $p<.01$) with additional effects of an age-by-language condition in-
432 teraction ($\beta=0.52$, $SE=0.212$, $t=2.46$, $p<.05$) and a language condition-by-
433 transition type interaction ($\beta=2.68$, $SE=1.35$, $t=1.99$, $p<.05$). There were
434 no significant effects of age or transition type alone ($\beta=-0.18$, $SE=0.175$,
435 $t=-1.04$, $p=.3$ and $\beta=-1.10$, $SE=0.865$, $t=-1.27$, $p=.2$, respectively).

436 Adults' anticipatory gaze switches shows an effect of transition type ($\beta=-$
437 4.5 , $SE=1.314$, $t=-3.42$, $p<.001$) and significant interactions between lan-
438 guage condition and transition type ($\beta=3.3$, $SE=1.61$, $t=2.05$, $p<.05$) and
439 transition type and gap duration ($\beta=10.51$, $SE=3.346$, $t=3.141$, $p<.01$).

⁶The 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.

440 2.2.2. Random baseline comparison

441 We estimated the probability that these patterns were the result of ran-
442 dom looking by running the same regression models on participants' real
443 eye-tracking data, only this time calculating their anticipatory gaze switches
444 with respect to randomly permuted turn transition windows. This process
445 involved: (1) Randomizing the order and temporal placement of the anal-
446 ysis windows within each stimulus (Figure 4; “analysis window” is defined
447 in Figure 2), thereby randomly redistributing the analysis windows across
448 the eye-tracking signal, (2) re-running each participant's eye tracking data
449 through switch identification (described in Section 2.1.4), this time using
450 the randomly permuted analysis windows, and (3) modeling the anticipatory
451 gazes from the randomly permuted data with the same statistical models we
452 used for the original data (Section 2.2.1; Table 2). Importantly, although
453 the onset time of each transition was shuffled within the eye-tracking signal,
454 the other intrinsic properties of each turn transition (e.g., prior speaker iden-
455 tity, transition type, gap duration, language condition, etc.) stayed constant
456 across each random permutation.

457 This procedure effectively de-links participants' gaze data from the turn
458 structure in the original stimulus, thereby allowing us to compare turn-

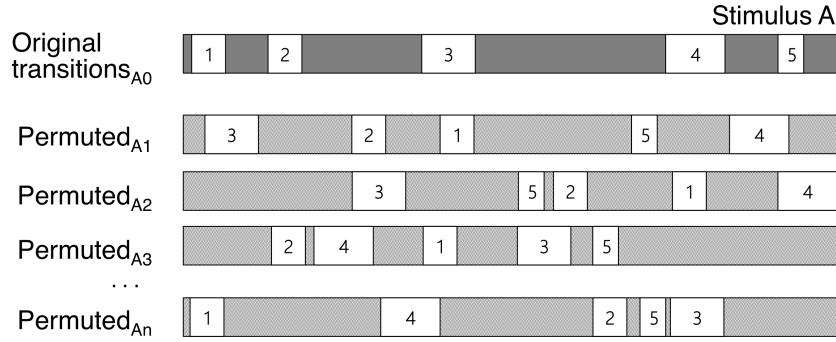


Figure 4: Example of analysis window permutations for a stimulus with five turn transitions. The windows were ± 300 msec around the inter-turn gap.

459 related (original) and non-turn-related (randomly permuted) looking behav-
 460 ior using the same eye movements. The resulting anticipatory gazes from the
 461 randomly permuted analysis windows represent an average anticipatory gaze
 462 rate over all possible starting points: a random baseline. By running the real
 463 and randomly permuted data sets through identical statistical models, we
 464 can also estimate how likely it is that predictor effects in the original data
 465 (e.g., the effect of language condition; Table 2) arose from random looking.

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

⁴⁶⁹ for each predictor in the 5,000 models of the randomly permuted datasets.⁷

⁴⁷⁰ We could then test whether significant effects from the original statistical

⁴⁷¹ models differed from the random baseline by calculating the proportion of

⁴⁷² random data t -values exceeded by the original t -value for each predictor,

⁴⁷³ using the absolute value of all t -values for a two-tailed test. For example,

⁴⁷⁴ children's original "language condition" t -value was $|3.429|$, which is greater

⁴⁷⁵ than 99.9% of all $|t\text{-value}|$ estimates from the randomly-permuted data mod-

⁴⁷⁶ els (i.e., $p=.001$). This leads us to conclude that that the effect of language

⁴⁷⁷ condition in the original model was highly unlikely to be the result of random

⁴⁷⁸ gaze shifting.

⁴⁷⁹ We excluded the output of random-permutation models that gave con-

⁴⁸⁰ vergence warnings in order to remove unreliable estimates from our analyses

⁴⁸¹ (non-converging models were 22.4% and 24.4% of all models for children and

⁴⁸² adults, respectively; see the Supplementary Materials for more information

⁴⁸³ on model exclusion).

⁷We 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.

484 Our baseline analyses revealed that none of the significant predictors from
485 models of the original, turn-related data can be explained by random look-
486 ing (see Figures A.1a and A.1b for the full distributional data). For the
487 children's data, the original t -values for language condition, gap duration,
488 the age-language condition interaction, and the language condition-transition
489 type interaction were all greater than 95% of t -values for the randomly per-
490 muted data (99.9%, 95.5%, 99.4%, and 96%, respectively). Similarly, the
491 adults' data showed significant differentiation from the randomly permuted
492 data for two of the three originally significant predictors—transition type and
493 the transition type-gap duration interaction (greater than 99.9% and 99.7%
494 of random t -values, respectively)—with marginal differentiation for the in-
495 teraction of language condition and transition type (greater than 94.6% of
496 random t -values).

497 *Developmental effects.* The original statistical model of the children's real
498 data revealed a significant interaction of age and language condition (Table
499 2) that was highly unlikely to have come from random looking (Figure 3).
500 To further explore this effect, we compared the average effect of language
501 condition for each age group: We extracted the average difference score for
502 the two language conditions (English minus non-English) for each subject,

503 computing an overall average for each random permutation of the data. For
504 each random permutation, we then made pairwise comparisons of the average
505 difference scores across participant age groups, finding that, while 3- and 4-
506 year olds showed similarly large effects of language condition, 5-year-olds
507 showed a significantly smaller effect of language condition, compared to both
508 younger age groups. In other words, the difference in the effect of language
509 condition for 5-year-olds compared to younger children was larger than would
510 be expected by chance in 99.52% of the randomly permuted data sets for 3-
511 year-olds and 99.96% of the data sets for 4-year-olds—differences of $p < .01$
512 and $p < .001$, respectively (see Figure B.1 for difference score distributions).

513 When does spontaneous turn prediction emerge developmentally during
514 natural speech? To test whether the youngest age group (3-year-olds) already
515 exceeded chance in their anticipatory gaze switches, we used two-tailed t -tests
516 to compare their real gaze rates to the random baseline in the most familiar
517 speech condition for our participants: English. We found that three-year-
518 olds made anticipatory gaze switches significantly above chance, both when
519 all transitions were considered ($t(22.824) = -4.147, p < .001$) and for question
520 transitions alone ($t(21.677) = -5.268, p < .001$).

521 2.3. Discussion

522 Children and adults spontaneously tracked the turn structure of the con-
523 versations, making anticipatory gaze switches at an above-chance rate across
524 all ages and conditions (Table 1; Figure 3). Children's anticipatory gaze rates
525 were affected by language condition, transition type, age, and gap duration
526 (Table 2), none of which could be explained by a baseline of random gaze
527 switching (Figure A.1a).

528 Language condition (English vs. non-English) affected children's antici-
529 pations in two ways (Table 2; Figure 3). First, children made more antici-
530 patory switches overall in English videos, compared to non-English videos.
531 This effect suggests that lexical access is important for children's ability
532 to anticipate upcoming turn structure; children had no lexical access to the
533 speech in the non-English videos, though they did have access to (non-native)
534 prosodic cues and non-verbal behavior. This finding is consistent with prior
535 work on turn-end prediction in adults (De Ruiter et al., 2006; Magyari and
536 De Ruiter, 2012) and children (Keitel et al., 2013). Second, children system-
537 atically made more anticipatory switches after hearing a question compared
538 to a non-question, but only in the English condition, suggesting that, when
539 children have access to lexical cues, they are more likely to make an anticipa-

540 tory gaze switch if they can expect an immediate response from the addressee.
541 If so, then children's (and adults') attention to lexical cues for turn taking
542 may primarily be in monitoring the signal for cues to questionhood (e.g.,
543 subject-auxiliary inversion, *wh*-words, etc.).

544 Children's anticipatory gaze switches were also affected by their age, but
545 only in the non-English videos: 3- and 4-year-olds made many more anticipa-
546 tory switches when watching videos in English compared to non-English, but
547 this effect of language condition had attenuated significantly by age 5 (Table
548 2; Figure 3; Figure B.1). This interaction suggests that the 5-year-olds were
549 able to leverage anticipatory cues in the non-English videos in a way that
550 3- and 4-year-olds could not, possibly by shifting more attention to the non-
551 native prosodic or non-verbal cues. Prior work on children's turn-structure
552 anticipation proposed that children's turn-end predictions rely primarily on
553 lexicosyntactic structure (and not, e.g., prosody) as they get older (Keitel
554 et al., 2013). The current results suggest more flexibility in children's pre-
555 dictions; when they do not have access to lexical information, older children
556 are likely to find alternative cues to turn taking behavior.

557 Finally, children showed an effect of gap duration (Table 2). This effect
558 is straightforward: Longer gaps resulted in longer analysis windows, yielding

559 more time for children to make an anticipatory gaze.

560 Adults' anticipatory gaze rates were also affected by transition type, lan-
561 guage condition, and gap duration (Table 2), none of which could be easily
562 explained by a baseline of random gaze switching (Figure A.1b). Like chil-
563 dren, adults made more anticipatory switches after hearing questions com-
564 pared to non-questions, suggesting that anticipation mattered more to them
565 when an immediate response was expected. Also like children, the advantage
566 for questions was driven by lexical access such that adults must have relied
567 on lexicosyntactic cues to questionhood in picking out turns that potentially
568 require an immediate response, though this effect was only marginally di-
569 vergent from the distribution of randomly permuted data ($p=.053$; Figure
570 A.1b). Finally, adults' anticipation rates were also affected by gap dura-
571 tion, but more so for questions than non-questions (Table 2), suggesting that
572 adults were less likely overall to make switches at non-questions, and so did
573 not benefit from extra time to do so.

574 *2.3.1. Summary*

575 Children and adults' predictions alike were benefited by access to lexical
576 information (English) and speech act status (questionhood), suggesting that
577 linguistic cues, particularly lexical ones, facilitate their spontaneous predic-

tions about upcoming turn structure through the identification of turns with immediate responses. Children’s anticipatory gaze rates for questions and non-questions in English was stable across ages and comparable to adult behavior (Figure 3), suggesting that they can identify questions in native stimuli with adult-like competence by age three. Although participants’ ability to recognize questions was facilitated by lexical access (i.e., English vs. non-English), the prosody in the non-English videos was non-native, and so the experimental design can not conclusively show which linguistic cues children relied on in the English videos to identify question turns. Relatedly, though lexical access clearly facilitated participants’ anticipatory gaze rate, it was not necessary for participants—especially adults—in order to exceed chance switching rates (Figure 3), suggesting that participants use non-lexical cues (e.g., prosody, non-verbal behavior) to make anticipatory eye movements at least some of the time.

Interestingly, adults and children both were strongly affected by transition type, in that they made more anticipatory switches after hearing questions, compared to non-questions. 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

597 question (Figure 3). Prior work using online, metalinguistic tasks has shown
598 that participants can use linguistic cues to accurately predict upcoming turn
599 ends (Torreira et al., 2015; Magyari and De Ruiter, 2012; De Ruiter et al.,
600 2006). The current results add a new dimension to our understanding of
601 how listeners make predictions about turn ends: Both children and adults
602 spontaneously monitor the linguistic structure of unfolding turns for cues to
603 upcoming responses.

604 Children and adults behaved relatively similarly in this first experiment,
605 and our language manipulation (English vs. non-English) was too coarse to
606 comment on when children begin to use specific linguistic cues (e.g., prosody
607 vs. lexicosyntax). We would instead need to directly compare lexicosyn-
608 tactic and prosodic cues in the participants' native language, controlling for
609 the presence of non-verbal cues. To see the emergence of anticipatory gaze
610 switching we would also need to include younger children, since participants
611 already reliably made anticipatory gaze switches at age three. We follow up
612 on both of these ideas in Experiment 2.

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= <i>non-English</i>	-3.29347	0.96045	-3.429	0.000606 ***
Type= <i>non-Question</i>	-1.10129	0.86494	-1.273	0.202925
Duration	3.40169	1.22826	2.770	0.005614 **
Age*LgCond= <i>non-English</i>	0.52065	0.21190	2.457	0.014008 **
Age*TypeS= <i>non-Question</i>	-0.01628	0.19437	-0.084	0.933232
LgCond= <i>non-English</i> *	2.68166	1.35016	1.986	0.047013 *
Type= <i>non-Question</i>				
Age*LgCond= <i>non-English</i> *	-0.45632	0.30163	-1.513	0.130315
Type= <i>non-Question</i>				

Adults

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

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

⁶¹³ **3. Experiment 2**

⁶¹⁴ We improved our design by using native-language stimuli, controlling for
⁶¹⁵ lexical and prosodic information, eliminating non-verbal cues, and testing
⁶¹⁶ children from a wider age range. All of the videos in Experiment 2 were in
⁶¹⁷ the participants' native language (American English). To tease apart the
⁶¹⁸ role of lexical and prosodic information, we phonetically manipulated the
⁶¹⁹ speech signal for pitch, syllable duration, and lexical access. By testing one-
⁶²⁰ to six-year-olds we hoped to find the developmental onset of turn-predictive
⁶²¹ gaze. We also hoped to measure changes in the relative roles of prosody and
⁶²² lexicosyntax across development.

⁶²³ Non-verbal cues in Experiment 1 (e.g., gaze and gesture) could have
⁶²⁴ helped participants make predictions about upcoming turn structure (Rossano
⁶²⁵ et al., 2009; Stivers and Rossano, 2010). Since our focus is on linguistic cues,
⁶²⁶ we eliminated all gaze and gestural signals in Experiment 2 by replacing
⁶²⁷ the videos of human actors with videos of puppets. Puppets are less real-
⁶²⁸ istic and expressive than human actors, but they create a natural context
⁶²⁹ for having somewhat motionless talkers in the videos (thereby allowing us
⁶³⁰ to eliminate gestural and gaze cues). Additionally, the prosody-controlled
⁶³¹ condition included small but global changes to syllable duration that would

632 have required complex video manipulation or precise re-enactment with hu-
633 man talkers, neither of which was feasible. For these reasons, we decided to
634 substitute puppet videos for human videos in the final stimuli.

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

639 *3.1. Methods*

640 *3.1.1. Participants*

641 We recruited 27 undergraduate adults and 129 children between ages 1;0–
642 6;11 to participate in our experiment. We recruited our child participants
643 from the Children's Discovery Museum in San Jose, California, targeting ap-
644 proximately 20 children for each of the six 1-year age groups (range=20–23).

645 All participants were native English speakers, though some parents (N=27)
646 reported that their child heard a second (and sometimes third) language at
647 home. None of the adult participants reported fluency in a second language.

648 We ran Experiment 2 at a local children's museum because it gave us access
649 to children with a more diverse range of ages.

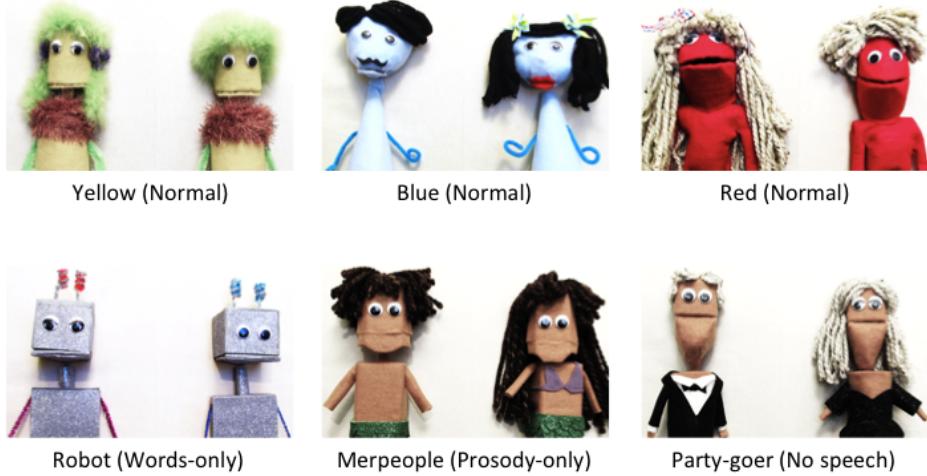


Figure 5: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

650 3.1.2. *Materials*

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

657 *Audio recordings.* The recording session was set up in the same way as

658 the first experiment, but with a shorter warm up period (5–10 minutes) and
659 a pre-determined topic for the child-friendly improvisation ('riding bikes',
660 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the
661 talkers were native English speakers, and were recorded in male-female pairs.

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

667 *Audio Manipulation.* We created four versions of each audio clip: *nor-*
668 *mal*, *words only*, *prosody only*, and *no speech*. That is, one version with a full
669 linguistic signal (*normal*), and three with incomplete linguistic information
670 (hereafter "limited cue" conditions). The *normal* clips were the unmanipu-
671 lated, original audio clips.

672 The *words only* clips were manipulated to have robot-like speech: We
673 flattened the intonation contours to each talker's average pitch (F0) and
674 we reset the duration of every nucleus and coda to each talker's average
675 nucleus and coda duration.⁸ We made duration and pitch manipulations

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

676 using PSOLA resynthesis in Praat (Boersma and Weenink, 2012). Thus,
677 the *words only* versions of the audio clips had no pitch or durational cues
678 to upcoming turn boundaries, but did have intact lexicosyntactic cues (and
679 residual phonetic correlates of prosody, e.g., intensity).

680 We created the *prosody only* clips by low-pass filtering the original record-
681 ing at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006).
682 This manipulation creates a “muffled speech” effect because low-pass filter-
683 ing removes most of the phonetic information used to distinguish between
684 phonemes. The *prosody only* versions of the audio clips lacked lexical infor-
685 mation, but retained their intonational and rhythmic cues to upcoming turn
686 boundaries.

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

clips.

694 Finally, because low-pass filtering removes significant acoustic energy, the
695 *prosody only* clips were much quieter than the other three conditions. Our
696 last step was to downscale the intensity of the audio tracks in the three other
697 conditions to match the volume of the *prosody only* clips. We referred to the
698 conditions as “normal”, “robot”, “mermaid”, and “birthday party” speech
699 when interacting with participants.

700 *Video recordings.* We created puppet video recordings to match the ma-
701 nipulated 20-second audio clips. The puppets were minimally expressive;
702 the experimenter could only control the opening and closing of their mouths;
703 their head, eyes, arms, and body stayed still. Puppets were positioned look-
704 ing forward to eliminate shared gaze as a cue to turn structure (Thorgrímsson
705 et al., 2015). We took care to match the puppets’ mouth movements to the
706 syllable onsets as closely as possible, specifically avoiding any mouth move-
707 ment before the onset of a turn. We then added the manipulated audio clips
708 to the puppet video recordings by hand.

709 We used three pairs of puppets used for the *normal* condition—‘red’,
710 ‘blue’ and ‘yellow’—and one pair of puppets for each limited cue condition:
711 “robots”, “merpeople”, and “party-goers” (Figure 8). We randomly assigned
712 half of the conversation topics (‘birthday cake’, ‘pets’, and ‘breakfast’) to the

713 *normal* condition, and half to the limited cue conditions ('riding bikes', 'rainy
714 days', and 'the library'). We then created three versions of the experiment,
715 so that each of the six puppet pairs was associated with three different con-
716 versation topics across the different versions of the experiment (18 videos
717 in total). We ensured that the position of the talkers (left and right) was
718 counterbalanced in each version by flipping the video and audio channels as
719 needed.

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

727 There were 85 turn transitions for analysis after excluding transitions
728 longer than 550 msec and shorter than 90 msec. The remaining turn tran-
729 sitions had slightly more questions than non-question (N=50 and N=35, re-
730 spectively), with transitions distributed somewhat evenly across conditions
731 (keeping in mind that there were three *normal* videos and only one lim-

732 ited cue video for each experiment version): *normal* (N=36), *words only*
733 (N=13), *prosody only* (N=17), and *no speech* (N=19). Inter-turn gaps for
734 questions (mean=365, median=427) were longer than those for non-questions
735 (mean=302, median=323) on average, but gap duration was overall com-
736 parable across conditions: *normal* (mean=334, median=321), *words only*
737 (mean=347, median=369), *prosody only* (mean=365, median=369), and *no*
738 *words* (mean=319, median=329). The longer gaps for question transitions
739 could give them an advantage because our anticipatory measure includes
740 shifts initiated during the gap between turns (Figure 2).

741 *3.2. Procedure*

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

750 3.2.1. *Data preparation and coding*

751 We coded each turn transition for its linguistic condition (*normal, words*
752 *only, prosody only*, and *no speech*) and transition type (question/non-question)⁹
753 and identified anticipatory gaze switches to the upcoming speaker using the
754 methods from Experiment 1.

755 3.3. *Results*

756 Participants' pattern of gaze indicated that they performed basic turn
757 tracking across all ages and in all conditions. Participants again looked at
758 the screen most of the time during video playback (82% and 86% average
759 for children and adults, respectively), primarily looking at the person who
760 was currently speaking (Table 2). They tracked the current speaker in every
761 condition—even one-year-olds looked more at the current speaker than at
762 anything else in the three limited cue conditions (40% for *words only*, 43%
763 for *prosody only*, and 39% for *no speech*). There was a steady overall increase
764 in looks to the current speaker with age and added linguistic information
765 (Tables 3 and 4). Looks to the addressee also decreased with age, but the

⁹We 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).

⁷⁶⁶ change was minimal. Figure 6 shows participants' anticipatory gaze rates
⁷⁶⁷ across age, the four language conditions, and transition type, in the real
⁷⁶⁸ data and the random baseline.

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.

⁷⁶⁹ *3.3.1. Statistical models*

⁷⁷⁰ We identified anticipatory gaze switches for all 85 usable turn transi-
⁷⁷¹ tions, and analyzed them for effects of language condition, transition type,
⁷⁷² and age with two mixed-effects logistic regressions. We again built separate

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.

773 models for children and adults because effects of age were only pertinent to
 774 the children’s data. The child model included condition (normal/prosody
 775 only/words only/no speech; with no speech as the reference level), transition
 776 type (question vs. non-question), age (1, 2, 3, 4, 5, 6; numeric), and duration
 777 of the inter-turn gap (in seconds) as predictors, with full interactions between
 778 language condition, transition type, and age. We again included the dura-
 779 tion of the inter-turn gap as a control predictor and added random effects of
 780 item (turn transition) and participant, with random slopes of transition type
 781 for participants. The adult model included condition, transition type, their
 782 interactions, and duration as a control predictor, with participant and item

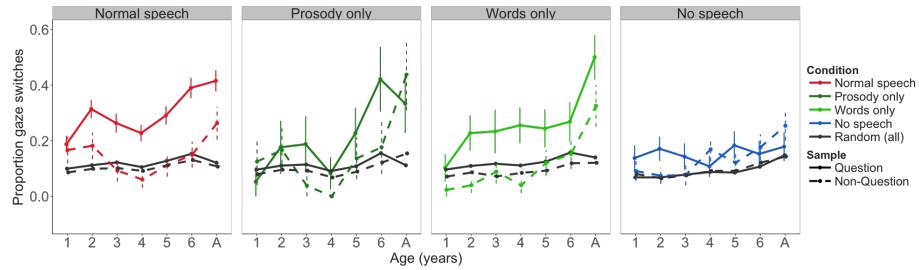


Figure 6: Anticipatory gaze rates across language condition and transition type for the real (blue, dark green, light green, and red) and randomly permuted baseline (gray). Vertical bars represent the standard error.

783 included as random effects and random slopes of condition and transition
 784 type.

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

Condition= <i>prosody</i>	-1.63429	0.86390	-1.892	0.0585 .
Age*Condition= <i>prosody</i>	0.39317	0.18907	2.080	0.0376 *
Condition= <i>prosody</i> *	1.77190	1.24864	1.419	0.1559
Type= <i>non-Question</i>				
Age*Condition= <i>prosody</i> *	-0.47057	0.28703	-1.639	0.1011
Type= <i>non-Question</i>				
Condition= <i>words</i>	-0.26741	0.59071	-0.453	0.6508
Age*Condition= <i>words</i>	0.13740	0.13568	1.013	0.3112
Condition= <i>words</i> *	-1.02193	1.01227	-1.010	0.3127
Type= <i>non-Question</i>				
Age*Condition= <i>words</i> *	0.08946	0.22349	0.400	0.6890
Type= <i>non-Question</i>				

Adults

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

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

785 Children's anticipatory gaze switches showed an effect of gap duration
786 ($\beta=4.18$, $SE=0.624$, $t=6.689$, $p<.001$), a two-way interaction of age and
787 language condition (for *prosody only* speech compared to the *no speech* refer-
788 ence level; $\beta=0.393$, $SE=0.189$, $t=2.08$, $p<.05$), and a three-way interaction
789 of age, transition type, and language condition (for *normal* speech compared
790 to the *no speech* reference level; $\beta=-0.375$, $SE=0.17$, $t=-2.213$, $p<.05$). There
791 were no significant effects of age or transition type alone (Table 3.3.1), with
792 only a marginal effect of language condition (for *prosody only* compared to
793 the *no speech* reference level; $\beta=-1.634$, $SE=0.864$, $t=-1.89$, $p=.06$)

794 Adults' anticipatory gaze switches showed effects of gap duration ($\beta=4.75$,
795 $SE=1.248$, $t=3.806$, $p<.001$) and language condition (for *normal* speech
796 $\beta=1.256$, $SE=0.563$, $t=2.229$, $p<.05$ and *words only* speech $\beta=1.594$, $SE=0.721$,
797 $t=2.211$, $p<.05$ compared to the *no speech* reference level). There were no
798 effects of transition type ($\beta=0.429$, $SE=0.609$, $t=0.705$, $p=.48$).

799 *3.3.2. Random baseline comparison*

800 Using the same technique described in Experiment 1 (Section 2.2.2), we
801 created and modeled 5,000 random permutations of participants' anticipa-

tory gaze (see Figures A.4a and A.4b for the full distributional data). Our baseline analyses revealed that none of the significant predictors from models of the original, turn-related data (Table 5: Children) can be explained by random looking. In the children’s data, the original 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). The effects of language condition and transition type for the real and randomly permuted data can also be observed in Figure 3.

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).

821 *Developmental effects.* The model of the children’s data revealed two signif-
822 icant interactions with age, neither of which derived from random looking
823 (Table 5; Figure 6). The first was a significant interaction of age and lan-
824 guage condition (for *prosody only* compared to the *no speech* reference level),
825 suggesting a different age effect between the two linguistic conditions. As
826 in Experiment 1, we further explored each age interaction by extracting an
827 average difference score over subjects for the effect of language condition (*no*
828 *speech* vs. *prosody only*) within each random permutation of the data, mak-
829 ing pairwise comparisons between the six age groups. These tests revealed
830 that children’s anticipation in the *prosody only* condition significantly im-
831 proved at ages five and six (difference scores greater than 95% of the random
832 data scores; $p < .05$; see Figure B.2 for the full distributions).

833 The second age-based interaction was a three-way interaction of age, tran-
834 sition type, and language condition (for *normal* speech compared to the *no*
835 *speech* baseline). We again created pairwise comparisons of the average dif-
836 ference scores for the transition type-language condition interaction across
837 age groups in each random permutation of the data, finding that the effect
838 of transition type in the *normal* speech condition became larger with age,
839 with significant improvements by age 4 over ages 1 and 2 (99.9% and 98.86%,

840 respectively), by age 5 over age 4 (97.54%), and by age 6 over ages 1, 2, and 5
841 (99.5%, 97.36%, and 95.04%), all significantly different from chance ($p < .05$;
842 see Figure B.3 for the full distributions).

843 Our main goal in extending the age range to 1- and 2-year-olds in Ex-
844 periment 2 was to find the age of emergence for spontaneous turn structure
845 predictions. As in Experiment 1, we used two-tailed t -tests to compare chil-
846 dren's real gaze rates to the random baseline in the *normal* speech condition,
847 in which the speech stimulus is most like what children hear every day. We
848 tested real gaze rates against baseline for three age groups: ages one, two,
849 and three. Two- and three-year-old children made anticipatory gaze switches
850 significantly above chance both when all transitions were considered (2-year-
851 olds: $t(26.193) = -4.137$, $p < .001$; 3-year-olds: $t(22.757) = -2.662$, $p < .05$) and
852 for question transitions alone (2-year-olds: $t(25.345) = -4.269$, $p < .001$; 3-year-
853 olds: $t(21.555) = -3.03$, $p < .01$). One-year-olds, however, made anticipatory
854 gaze shifts that were only marginally above chance for turn transitions overall
855 and for question turns alone (overall: $t(24.784) = -2.049$, $p = .051$; questions:
856 $t(25.009) = -2.03$, $p = .053$).

857 3.4. Discussion

858 As in Experiment 1, children and adults spontaneously tracked the turn
859 structure of the conversations, making anticipatory gaze switches at an above-
860 chance rate across all ages, but not in all conditions (Table 3; Figure 6). How-
861 ever, when children had access to full linguistic information (in the *normal*
862 speech condition), they made more anticipatory gaze switches than expected
863 by chance at age two (and even marginally at age one). Across conditions,
864 children's anticipatory gaze rates were affected by gap duration, plus interac-
865 tions of language condition, transition type and age (Table 5), none of which
866 could be explained by a baseline of random gaze switching (Figure A.4a).

867 Two of the three linguistic conditions affected children's anticipatory gaze
868 switches, compared to the no-speech reference level: prosody-only speech and
869 *normal* speech. *Prosody only* speech resulted in fewer anticipatory gazes over-
870 all, compared to the no-speech baseline, though the effect was only marginal
871 ($p=.06$). However, prosody-only speech *did* significantly differ from the *no*
872 *speech* condition in its interaction with age: Whereas age gains were neg-
873 ligible in the *no speech* condition, 5- and 6-year-olds in the prosody-only
874 condition showed significantly more anticipatory gaze switches than younger
875 children (Figure B.2), going from at-chance anticipatory gaze rates to rates

⁸⁷⁶ well above chance (Figure 6).

⁸⁷⁷ Anticipatory gaze for questions increased significantly more with age in
⁸⁷⁸ the *normal* speech condition compared to the no-speech baseline. As in Ex-
⁸⁷⁹ periment 1, children consistently made more anticipatory gazes after hearing
⁸⁸⁰ questions when they had access to lexical material. But now, with a larger
⁸⁸¹ age span in Experiment 2, we can begin to see a developmental path for
⁸⁸² the question effect. At least for *normal* speech, greater anticipatory looking
⁸⁸³ for questions is not present from the start: 4-year-olds are the first to make
⁸⁸⁴ significant gains on 1- and 2-year-olds, but then there are further significant
⁸⁸⁵ gains at age 5, and again age 6 (Figure B.3). While children at ages five
⁸⁸⁶ and six showed adult-like differentiation of questions and non-questions in
⁸⁸⁷ the *normal* speech condition, 1-year-olds had nearly identical switch rates
⁸⁸⁸ for the two transition types. This suggests that the participants' tendency
⁸⁸⁹ to make more anticipatory switches for questions emerges after age one and
⁸⁹⁰ continues developing, with rapid gains in ages three through age six. Fi-
⁸⁹¹nally, children showed a straightforward effect of gap duration (Table 5), as
⁸⁹² in Experiment 1.

⁸⁹³ Adults' anticipatory gaze rates were affected by gap duration and two
⁸⁹⁴ of the language conditions (Table 5), none of which could be explained by

895 a baseline of random gaze switching (Figure A.4b). Adults made more an-
896 ticipatory switches overall for the *normal* speech and words-only conditions
897 compared to the no-speech condition, falling in-line with past work showing
898 that adults primarily use lexical information in making predictions about up-
899 coming turn structure (De Ruiter et al., 2006). Though adults did make more
900 anticipatory switches for questions than non-questions on average in the two
901 lexical conditions (Figure 6), the effect of transition type was not significant
902 in either one (Table 5). Like children, adults also showed a straightforward
903 effect of gap duration.

904 *3.4.1. Summary*

905 Children and adults both showed more anticipatory gaze switches with
906 increased linguistic information, but only for a subset of the linguistic con-
907 ditions and transition types. We had expected to see the most anticipatory
908 switches in the *normal* condition and the least anticipatory switches in the
909 *no speech* condition because they contained the most and least linguistic in-
910 formation, respectively. We had also expected to replicate our finding from
911 Experiment 1 that questions result in more anticipatory switches than non-
912 questions, with the added hypothesis that the question effect is driven by
913 lexicosyntactic cues. We additionally anticipated an overall increase in an-

914 ticipatory switches with age. Finally, since the development of prosodic skills
915 partially precedes the development of lexicosyntax, we expected to see more
916 switches in the *prosody only* condition compared to the *words only* condition
917 in the youngest age groups.

918 In fact, children and adults did show more anticipatory gaze switching
919 in the *normal* speech condition, compared to the no-speech condition, but
920 for children this effect only emerged in the form of anticipations following
921 question turns, which increased with age significantly faster than it did when
922 there were no linguistic cues present at all. Adults also showed significantly
923 higher anticipatory switch rates in the *words only* condition but no effects
924 of transition type, alone or within linguistic conditions. Taken together,
925 these results partly replicate the findings from Experiment 1: Participants
926 make more anticipatory switches when they have access to lexical information
927 and, when they do, tend to make more anticipatory switches for questions
928 compared to non-questions.

929 We had anticipated significant gains in anticipatory switching with age,
930 but children only showed significant developmental increases in the *prosody*
931 *only* condition and the *normal* condition (for question transitions). Rather
932 than showing an early advantage for prosody over lexical information, chil-

933 dren's anticipation were at chance with prosody-only speech and did not show
934 significant improvement until age five. In contrast, their anticipatory gaze
935 switches were already marginally above chance at age one and significantly
936 so at age two for *normal* speech. These findings therefore do not support an
937 early role for prosody in children's spontaneous turn structure predictions.
938 On the contrary, children's predictions were best when lexical information
939 was present, especially when following a question.

940 However, these results do not support the idea that lexical information is
941 sufficient for children, as has been proposed for adults previously (De Ruiter
942 et al., 2006) and replicated in the present study. On the contrary, children
943 showed significant gains in the *normal* speech condition, but *not* in the *words*
944 *only* condition. Notably, the *normal* speech condition, in addition to having
945 both lexical and prosodic information, is also the one most likely to occur
946 in our participants' daily lives (compared to muffled or robotic speech) and
947 therefore may have an extra advantage of familiarity over the other condi-
948 tions.

949 Finally, though participants were often able to make anticipatory gaze
950 switches more often than would be expected by chance with no linguistic in-
951 formation (in the *no speech* condition), the advantage for lexical cues emerg-

952 ing in the children's data with age and in the adult's data overall suggests
953 that participants were using linguistic information in making their predic-
954 tions when it was available.

955 The core aims of Experiment 2 were to gain better traction on the indi-
956 vidual roles of prosody and lexicosyntax in children's turn predictions, and
957 to find the age of emergence for spontaneous turn anticipation. Replicating
958 many of our findings from Experiment 1, we found that children and adults
959 both use linguistic cues, in children's case, primarily to anticipate responses
960 after question turns and primarily when the full speech signal was available.
961 Adults, on the other hand, showed broadly increased anticipation for lexical
962 conditions, with no overall effects of transition type.

963 Rather than finding an early advantage for prosodic cues, we found that
964 children struggled to make above-chance anticipations in the *prosody only*
965 condition until age five. They didn't show significant overall effects for lexical
966 information alone either, suggesting that children need the full linguistic
967 signal to extract predictive benefits from ongoing conversational speech. This
968 finding is surprising given children's earlier development of prosodic skills
969 but unsurprising in the light of their overall question bias—compared to
970 prosodic cues, lexicosyntactic cues to questionhood are categorical units, have

971 clearer form-to-meaning mappings, and occur early in the utterance. This
972 finding has implications for our conceptions about adults and children's use
973 of linguistic cues for spontaneous turn prediction, which we expand on in the
974 General Discussion.

975 **4. General Discussion**

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

982 In two experiments investigating children's anticipatory gaze to upcom-
983 ing speakers, we found evidence that turn prediction develops early in child-
984 hood and that spontaneous predictions are primarily driven by participants'
985 expectation of an immediate response in the next turn. In making predic-
986 tions about upcoming turn structure, children used a combination of lexical
987 and prosodic cues; neither lexical nor prosodic cues alone were sufficient to
988 support increased anticipatory gaze. We also found no early advantage for

989 prosody over lexicosyntax, and instead found that children were unable to
990 make above-chance anticipatory gazes in the *prosody only* condition until age
991 five. We discuss these findings with respect to the role of linguistic cues in
992 predictions about upcoming turn structure, the importance of questions in
993 spontaneous predictions about conversation, and children's developing com-
994 petence as conversationalists.

995 *4.1. Predicting turn structure with linguistic cues*

996 Prior work with adults has found a consistent and critical role for lex-
997 icosyntax in predicting upcoming turn structure (De Ruiter et al., 2006;
998 Magyari and De Ruiter, 2012), with the role of prosody still under debate
999 (Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015). Knowing
1000 that children comprehend more about prosody than lexicosyntax early on
1001 (see introduction; also see Speer and Ito, 2009 for a review), we thought it
1002 possible that young children would instead show an advantage for prosody in
1003 their predictions about turn structure in conversation. Our results suggest
1004 that, on the contrary, when presented with *only* prosodic information, chil-
1005 dren's spontaneous predictions about upcoming turn structure are limited
1006 until age five.

1007 Importantly, we also found no evidence that lexical information alone is

1008 equivalent to full linguistic information for children, as has been shown be-
1009 fore (Magyari and De Ruiter, 2012; De Ruiter et al., 2006) and replicated
1010 in the current study for adult participants: In both experiments, children's
1011 performance was best in conditions when they had access to the full linguistic
1012 signal. Adults on the other hand, showed significant gains in anticipatory
1013 gaze switching in both conditions with lexical cues. If this effect arose in
1014 our data because children do not make predictions based on phonetically
1015 manipulated speech in conversational contexts, the current findings can be
1016 overturned by follow-up work controlling linguistic information through other
1017 means. But if not, there may be something specially informative about the
1018 combined prosodic and lexical cues to questionhood that boosts children's an-
1019 ticipations before they can use these cues separately. Even in adults, Torreira
1020 and colleagues (2015) were able to show that the trade-off in informativity
1021 between lexical and prosodic cues is more subtle in semi-natural (spliced)
1022 speech. The present findings are the first to show evidence of a similar effect
1023 developmentally.

1024 *4.1.1. The question effect*

1025 In both experiments, anticipatory looking was primarily driven by ques-
1026 tion transitions, a pattern that had not been previously reported in other an-

₁₀₂₇ ticipatory gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari,
₁₀₂₈ 2013; Tice and Henetz, 2011). Questions make an upcoming speaker switch
₁₀₂₉ immediately relevant, helping the listener to predict with high certainty what
₁₀₃₀ will happen next (i.e., an answer from the addressee), and are often easily
₁₀₃₁ identifiable by overt prosodic and lexicosyntactic cues.

₁₀₃₂ Compared to prosodic cues (e.g., final rising intonation), lexicosyntactic
₁₀₃₃ cues (e.g., *wh*-words, *do*-insertion, and subject-auxiliary inversion) were fre-
₁₀₃₄ quent, categorical, and early-occurring in the utterance. Children may have
₁₀₃₅ therefore had an easier time picking out and interpreting lexical cues to ques-
₁₀₃₆ tionhood. The question effect showed its first significant gains between ages
₁₀₃₇ three and four in the *normal* speech condition of Experiment 2, by which
₁₀₃₈ time children frequently hear and use a variety of polar *wh*-questions (Clark,
₁₀₃₉ 2009). Furthermore, while lexicosyntactic question cues were available on
₁₀₄₀ every instance of *wh*- and *yes/no* questions in our stimuli, prosodic question
₁₀₄₁ cues were only salient on *yes/no* questions and, even then, the mapping of
₁₀₄₂ prosodic contour to speech act (e.g., high final rises for polar questions) is
₁₀₄₃ far from one-to-one.

₁₀₄₄ Prior work on children's acquisition of questions indicates that they may
₁₀₄₅ 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; Hening et al., 2005; Shatz, 1979).¹⁰ 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 ad-

¹⁰There is substantial variation in question frequency by individual and socioeconomic class (Hart and Risley, 1992).

₁₀₆₃ dressees? Other request formats, such as imperatives, compliments, and
₁₀₆₄ complaints make a response from the addressee highly likely in the next turn
₁₀₆₅ (Schegloff, 2007). Rhetorical and tag questions, on the other hand, take a
₁₀₆₆ similar form to prototypical polar questions, but often do not require an an-
₁₀₆₇ swer. So, though it is clear that adults and children anticipated responses
₁₀₆₈ more often for questions than non-questions, we do not yet know whether
₁₀₆₉ their predictive action is limited to turns formatted as questions or is gener-
₁₀₇₀ ally applicable to turn structures that project an immediate response from
₁₀₇₁ the addressee.

₁₀₇₂ The question effect itself has implications for our current theories about
₁₀₇₃ turn prediction. While participants may always use linguistic information
₁₀₇₄ to predict upcoming speaker changes (as they have in the two studies pre-
₁₀₇₅ sented here), they might not always use linguistic information to predict
₁₀₇₆ upcoming turn ends, as is assumed in metalinguistic measures of turn-end
₁₀₇₇ prediction (e.g., pressing a button while listening to speech: Torreira et al.,
₁₀₇₈ 2015; Magyari and De Ruiter, 2012; De Ruiter et al., 2006) and as has been
₁₀₇₉ the focus of other turn-end prediction studies (Ford and Thompson, 1996;
₁₀₈₀ Duncan, 1972). Our results rather suggest that participants' *spontaneous*
₁₀₈₁ predictions, at least while viewing third-party conversation, are the results

1082 of question-monitoring, presumably achieved by recruiting linguistic cues to
1083 questionhood from the unfolding signal. In other words, our results suggest
1084 that predictions are driven by what is *beyond* the end of the current turn—
1085 that questions, not lexical cues, are sufficient for prediction at the first level
1086 of conversation monitoring.

1087 There is at least one clear hypothesis that can bridge these apparently
1088 conflicting results: Listeners in spontaneous, first-person conversation may
1089 use multiple strategies in making predictions about upcoming turn structure,
1090 using more passive prediction to detect upcoming speaker transition (e.g.,
1091 questions), and then switching into precise turn-end prediction mode (à la
1092 De Ruiter et al., 2006) when necessary. A flexible prediction system like this
1093 one allows listeners to continuously monitor ongoing conversation at a low
1094 cost while still managing to plan their responses and come in quickly when
1095 needed.

1096 To test this hypothesis, which integrates findings from turn-structure pre-
1097 diction with multiple measures, age groups, and styles of linguistic control,
1098 it will be crucial to look at prediction from a first-person perspective. The
1099 results we present here are based on predictions about third-party conver-
1100 sation, which enables participants to follow interactions with no chance of

1101 actually participating. Although recent work has shown that similar anticipatory eye gazes do occur in spontaneous conversation (Holler and Kendrick,
1102 2015), more work is needed to determine if the same question advantage occurs, which linguistic cues seem to drive it, and whether participants switch
1103 into a “precision” mode when they detect imminent speaker change.
1104

1105 *4.1.2. Early competence for turn taking?*

1106 One of the core aims of our study was to test whether children show an
1107 early competence for turn taking, as is proposed by studies of spontaneous
1108 mother-infant proto-conversation and theories about the mechanisms under-
1109 lying human interaction in general (Hilbrink et al., 2015; Levinson, 2006). We
1110 did find evidence that young children already make spontaneous predictions
1111 about upcoming turn structure, definitely at age two and even marginally
1112 at age one. However, “above chance” performance was far from adult-like
1113 predictive behavior, and children in our studies did not show adult-like com-
1114 petency in their predictions, even at age six. This may indicate that children
1115 rely more on non-verbal cues in anticipating turn transitions or, alternatively,
1116 that adults are better at flexibly adapting to the turn-relevant cues present
1117 at any moment.

1118 Taken together, the data suggest that turn-taking skills do begin to

1120 emerge in infancy, but that their predictions don't help them anticipate much
1121 until they have acquired the ability to pick out question turns. This finding
1122 leads us to wonder how participant role (first- instead of third-person) and
1123 cultural differences (e.g., high vs. low parent-infant interaction styles) might
1124 feed into this early predictive skill. It also bridges the prior work showing a
1125 predisposition for turn taking in infancy (e.g., Hilbrink et al., 2015) but late
1126 acquisition of adult-like competence when it comes to integrating linguistic
1127 information into turn-taking behaviors (Casillas et al., In press; Garvey, 1984;
1128 Ervin-Tripp, 1979).

1129 *4.2. Limitations and future work*

1130 There are at least two major limitations to our work: Speech naturalness
1131 and participant role. Following prior work (De Ruiter et al., 2006; Keitel et al.,
1132 2013), we used phonetically manipulated speech in Experiment 2, resulting in
1133 speech sounds that children don't usually hear in their natural environment.
1134 Many prior studies have used phonetically-altered speech with infants and
1135 young children (cf. Jusczyk, 2000), but almost none of them have done so
1136 in a conversational context. Future work could instead carefully script or
1137 cross-splice parts of turns to control for the presence or absence of linguistic
1138 cues for turn transition.

1139 The prediction measure used in these studies is based on an observer's
1140 view of third-party conversation but, because participants' role in the inter-
1141 action could affect their online predictions about turn taking, an ideal exper-
1142 imental measure would capture first-person behavior. First-person measures
1143 of spontaneous turn prediction will be key to revealing how participants dis-
1144 tribute their attention over linguistic and non-verbal cues while taking part
1145 in everyday interaction, the implications of which relate to theories of online
1146 language processing for both language learning and everyday talk.

1147 *4.3. Conclusions*

1148 Conversation plays a central role in children's language learning. It is
1149 the driving force behind what children say and what they hear. Adults use
1150 linguistic information to accurately predict turn structure in conversation,
1151 which facilitates their online comprehension and allows them to respond rel-
1152 evantly and on time. The present study offers new findings regarding the
1153 role of speech acts and linguistic processing in online turn prediction, and
1154 has given evidence that turn prediction emerges by age two is not integrated
1155 with linguistic cues until much later. Using language to make predictions
1156 about upcoming interactive content takes time and, for both children and
1157 adults, is primarily driven by participants' orientation to what will happen

1158 beyond the end of the current turn.

1159 **Acknowledgements**

1160 We gratefully acknowledge the parents and children at Bing Nursery
1161 School and the Children’s Discovery Museum of San Jose. This work was
1162 supported by an ERC Advanced Grant to Stephen C. Levinson (269484-
1163 INTERACT), NSF graduate research and dissertation improvement fellow-
1164 ships to the first author, and a Merck Foundation fellowship to the second
1165 author. Earlier versions of these data and analyses were presented to confer-
1166 ence audiences (Casillas and Frank, 2012, 2013). We also thank Tania Henetz,
1167 Francisco Torreira, Stephen C. Levinson, Eve V. Clark, and the First Lan-
1168 guage Acquisition group at Radboud University for their feedback on earlier
1169 versions of this work. The analysis code and raw data for this project can
1170 be found on GitHub at https://github.com/langcog/turn_taking/.

1171 **References**

1172 Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure
1173 for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory*
1174 and Language

68, 255–278.

- 1175 Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4:
1176 Linear mixed-effects models using Eigen and S4. URL:
1177 <https://github.com/lme4/lme4><http://lme4.r-forge.r-project.org/>.
1178 [Computer program] R package version 1.1-7.
- 1179 Bateson, M.C., 1975. Mother-infant exchanges: The epigenesis of conver-
1180 sational interaction. Annals of the New York Academy of Sciences 263,
1181 101–113.
- 1182 Bergelson, E., Swingley, D., 2013. The acquisition of abstract words by young
1183 infants. Cognition 127, 391–397.
- 1184 Bloom, K., 1988. Quality of adult vocalizations affects the quality of infant
1185 vocalizations. Journal of Child Language 15, 469–480.
- 1186 Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer. URL:
1187 <http://www.praat.org>. [Computer program] Version 5.3.16.
- 1188 Bögels, S., Magyari, L., Levinson, S.C., 2015. Neural signatures of response
1189 planning occur midway through an incoming question in conversation. Sci-
1190 entific Reports 5.

- ₁₁₉₁ Bruner, J., 1985. Child's talk: Learning to use language. *Child Language*
- ₁₁₉₂ Teaching and Therapy 1, 111–114.
- ₁₁₉₃ Bruner, J.S., 1975. The ontogenesis of speech acts. *Journal of Child Language*
- ₁₁₉₄ 2, 1–19.
- ₁₁₉₅ Carlson, R., Hirschberg, J., Swerts, M., 2005. Cues to upcoming swedish
- ₁₁₉₆ prosodic boundaries: Subjective judgment studies and acoustic correlates.
- ₁₁₉₇ *Speech Communication* 46, 326–333.
- ₁₁₉₈ Casillas, M., Bobb, S.C., Clark, E.V., In press. Turn taking, timing, and
- ₁₁₉₉ planning in early language acquisition. *Journal of Child Language* .
- ₁₂₀₀ Casillas, M., Frank, M.C., 2012. Cues to turn boundary prediction in adults
- ₁₂₀₁ and preschoolers, in: *Proceedings of SemDial*.
- ₁₂₀₂ Casillas, M., Frank, M.C., 2013. The development of predictive processes
- ₁₂₀₃ in children's discourse understanding, in: *Proceedings of the 35th Annual*
- ₁₂₀₄ *Meeting of the Cognitive Science Society*.
- ₁₂₀₅ Clark, E.V., 2009. First language acquisition. Cambridge University Press.
- ₁₂₀₆ De Ruiter, J.P., Mitterer, H., Enfield, N.J., 2006. Projecting the end of

- 1207 a speaker's turn: A cognitive cornerstone of conversation. *Language* 82,
1208 515–535.
- 1209 De Vos, C., Torreira, F., Levinson, S.C., 2015. Turn-timing in signed con-
1210 versations: coordinating stroke-to-stroke turn boundaries. *Frontiers in*
1211 *Psychology* 6.
- 1212 Dingemanse, M., Torreira, F., Enfield, N., 2013. Is “Huh?” a universal word?
1213 Conversational infrastructure and the convergent evolution of linguistic
1214 items. *PloS one* 8, e78273.
- 1215 Duncan, S., 1972. Some signals and rules for taking speaking turns in con-
1216 versations. *Journal of Personality and Social Psychology* 23, 283.
- 1217 Ervin-Tripp, S., 1979. Children’s verbal turn-taking, in: Ochs, E., Schieffelin,
1218 B.B. (Eds.), *Developmental Pragmatics*. Academic Press, New York, pp.
1219 391–414.
- 1220 Fitneva, S., 2012. Beyond answers: questions and children’s learning, in:
1221 De Ruiter, J.P. (Ed.), *Questions: Formal, Functional, and Interactional*
1222 Perspectives. Cambridge University Press, Cambridge, UK, pp. 165–178.
- 1223 Ford, C.E., Thompson, S.A., 1996. Interactional units in conversation: Syn-

- 1224 tactic, intonational, and pragmatic resources for the management of turns.
- 1225 Studies in Interactional Sociolinguistics 13, 134–184.
- 1226 Garvey, C., 1984. Children's Talk. volume 21. Harvard University Press.
- 1227 Gísladóttir, R., Chwilla, D., Levinson, S.C., 2015. Conversation electrified:
- 1228 ERP correlates of speech act recognition in underspecified utterances. PloS
1229 one 10, e0120068.
- 1230 Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. Psycho-
- 1231 logical science 11, 274–279.
- 1232 Hart, B., Risley, T.R., 1992. American parenting of language-learning chil-
- 1233 dren: Persisting differences in family-child interactions observed in natural
- 1234 home environments. Developmental Psychology 28, 1096.
- 1235 Hedberg, N., Sosa, J.M., Görgülü, E., Mameni, M., 2010. The prosody and
- 1236 meaning of Wh-questions in American English, in: Speech Prosody 2010–
- 1237 Fifth International Conference.
- 1238 Henning, A., Striano, T., Lieven, E.V., 2005. Maternal speech to infants at
- 1239 1 and 3 months of age. Infant Behavior and Development 28, 519–536.
- 1240 Hilbrink, E., Gattis, M., Levinson, S.C., 2015. Early developmental changes

- 1241 in the timing of turn-taking: A longitudinal study of mother-infant inter-
1242 action. *Frontiers in Psychology* 6.
- 1243 Hirvenkari, L., Ruusuvuori, J., Saarinen, V.M., Kivioja, M., Peräkylä, A.,
1244 Hari, R., 2013. Influence of turn-taking in a two-person conversation on
1245 the gaze of a viewer. *PLoS one* 8, e71569.
- 1246 Holler, J., Kendrick, K.H., 2015. Unaddressed participants' gaze in multi-
1247 person interaction. *Frontiers in Psychology* 6.
- 1248 Jaffé, J., Beebe, B., Feldstein, S., Crown, C.L., Jasnow, M.D., Rochat, P.,
1249 Stern, D.N., 2001. Rhythms of dialogue in infancy: Coordinated timing in
1250 development. *Monographs of the Society for Research in Child Develop-
1251 ment*. JSTOR.
- 1252 Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds:
1253 When speech cues count more than statistics. *Journal of Memory and
1254 Language* 44, 548–567.
- 1255 Jusczyk, P.W., 2000. *The Discovery of Spoken Language*. MIT press.
- 1256 Jusczyk, P.W., Hohne, E., Mandel, D., Strange, W., 1995. Picking up reg-
1257 ularities in the sound structure of the native language. *Speech perception*

- 1258 and linguistic experience: Theoretical and methodological issues in cross-
- 1259 language speech research , 91–119.
- 1260 Kamide, Y., Altmann, G., Haywood, S.L., 2003. The time-course of predic-
- 1261 tion in incremental sentence processing: Evidence from anticipatory eye
- 1262 movements. *Journal of Memory and Language* 49, 133–156.
- 1263 Keitel, A., Prinz, W., Friederici, A.D., Hofsten, C.v., Daum, M.M., 2013.
- 1264 Perception of conversations: The importance of semantics and intonation
- 1265 in childrens development. *Journal of Experimental Child Psychology* 116,
- 1266 264–277.
- 1267 Lemasson, A., Glas, L., Barbu, S., Lacroix, A., Guilloux, M., Remeuf, K.,
- 1268 Koda, H., 2011. Youngsters do not pay attention to conversational rules:
- 1269 is this so for nonhuman primates? *Nature Scientific Reports* 1.
- 1270 Levelt, W.J., 1989. Speaking: From intention to articulation. MIT press.
- 1271 Levinson, S.C., 2006. On the human “interaction engine”, in: Enfield, N.,
- 1272 Levinson, S. (Eds.), Roots of human sociality: Culture, cognition and
- 1273 interaction. Oxford: Berg, pp. 39–69.
- 1274 Levinson, S.C., 2013. Action formation and ascriptions, in: Stivers, T., Sid-

- 1275 nell, J. (Eds.), *The Handbook of Conversation Analysis*. Wiley-Blackwell,
1276 Malden, MA, pp. 103–130.
- 1277 Magyari, L., Bastiaansen, M.C.M., De Ruiter, J.P., Levinson, S.C., 2014.
1278 Early anticipation lies behind the speed of response in conversation. *Jour-*
1279 *nal of Cognitive Neuroscience* 26, 2530–2539.
- 1280 Magyari, L., De Ruiter, J.P., 2012. Prediction of turn-ends based on antici-
1281 pation of upcoming words. *Frontiers in Psychology* 3:376, 1–9.
- 1282 Masataka, N., 1993. Effects of contingent and noncontingent maternal stimu-
1283 lation on the vocal behaviour of three-to four-month-old Japanese infants.
1284 *Journal of Child Language* 20, 303–312.
- 1285 Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoni, J., Amiel-
1286 Tison, C., 1988. A precursor of language acquisition in young infants.
1287 *Cognition* 29, 143–178.
- 1288 Morgan, J.L., Saffran, J.R., 1995. Emerging integration of sequential and
1289 suprasegmental information in preverbal speech segmentation. *Child De-*
1290 *velopment* 66, 911–936.

- 1291 Nazzi, T., Ramus, F., 2003. Perception and acquisition of linguistic rhythm
1292 by infants. *Speech Communication* 41, 233–243.
- 1293 R Core Team, 2014. R: A Language and Environment for Statistical Com-
1294 puting. R Foundation for Statistical Computing. Vienna, Austria. URL:
1295 <http://www.R-project.org>. [Computer program] Version 3.1.1.
- 1296 Ratner, N., Bruner, J., 1978. Games, social exchange and the acquisition of
1297 language. *Journal of Child Language* 5, 391–401.
- 1298 Ross, H.S., Lollis, S.P., 1987. Communication within infant social games.
1299 *Developmental Psychology* 23, 241.
- 1300 Rossano, F., Brown, P., Levinson, S.C., 2009. Gaze, questioning and culture,
1301 in: Sidnell, J. (Ed.), *Conversation Analysis: Comparative Perspectives*.
1302 Cambridge University Press, Cambridge, pp. 187–249.
- 1303 Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the
1304 organization of turn-taking for conversation. *Language* 50, 696–735.
- 1305 Schegloff, E.A., 2007. *Sequence organization in interaction: Volume 1: A
1306 primer in conversation analysis*. Cambridge University Press.
- 1307 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.

- 1324 Stivers, T., Rossano, F., 2010. Mobilizing response. *Research on Language*
1325 and Social Interaction 43, 3–31.
- 1326 Takahashi, D.Y., Narayanan, D.Z., Ghazanfar, A.A., 2013. Coupled oscillator
1327 dynamics of vocal turn-taking in monkeys. *Current Biology* 23, 2162–2168.
- 1328 Thorgrímsson, G., Fawcett, C., Liszkowski, U., 2015. 1- and 2-year-olds'
1329 expectations about third-party communicative actions. *Infant Behavior*
1330 and Development 39, 53–66.
- 1331 Tice (Casillas), M., Henetz, T., 2011. Turn-boundary projection: Looking
1332 ahead, in: *Proceedings of the 33rd Annual Meeting of the Cognitive Science*
1333 Society.
- 1334 Torreira, F., Bögels, S., Levinson, S.C., 2015. Intonational phrasing is neces-
1335 sary for turn-taking in spoken interaction. *Journal of Phonetics* 52, 46–57.
- 1336 Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006.
1337 Elan: a professional framework for multimodality research, in: *Proceedings*
1338 of LREC.

1339 **Appendix A. Real vs. randomly permuted model outcomes: t , β ,**
 1340 **and SE**

1341 In all of the following plots, the gray dots represent the randomly per-
 1342 muted data's model estimates for the value listed (beta or standard error),
 1343 the white dots represent the model estimates from the original data, and the
 1344 triangles represent the 95th percentile for each distribution being shown.

1345 *Appendix A.1. Experiment 1*

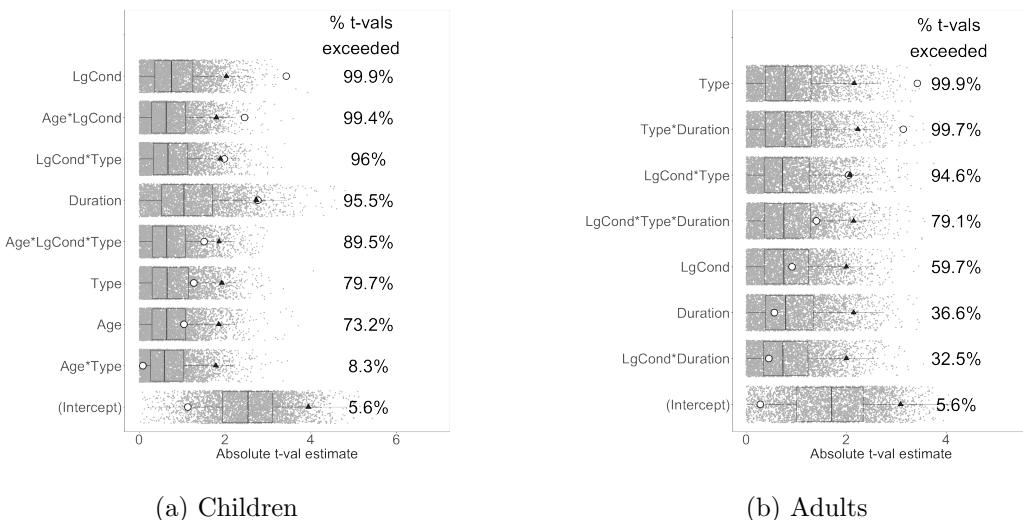


Figure A.1: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.

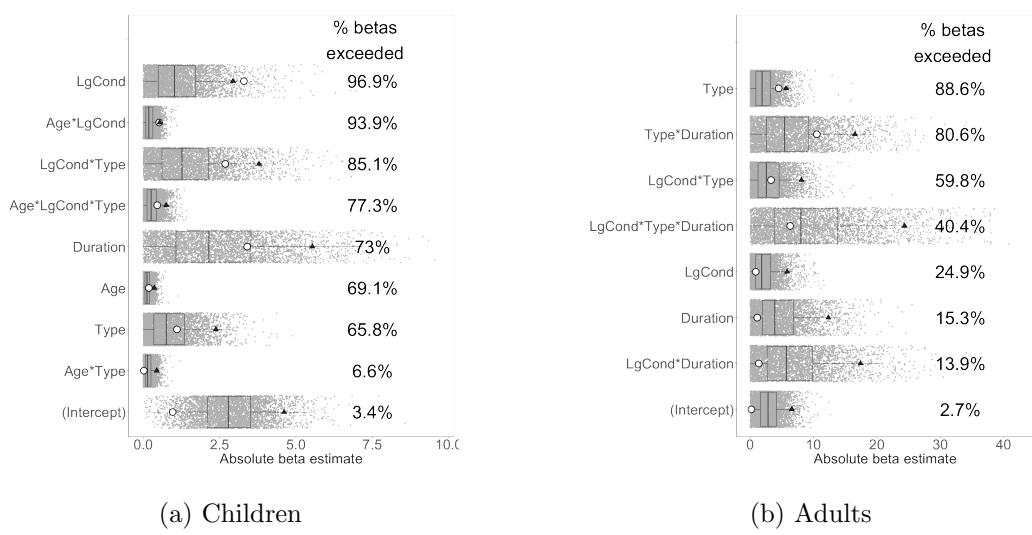


Figure A.2: Random-permutation and original $|\beta\text{-values}|$ for predictors of gaze rates in Experiment 1.

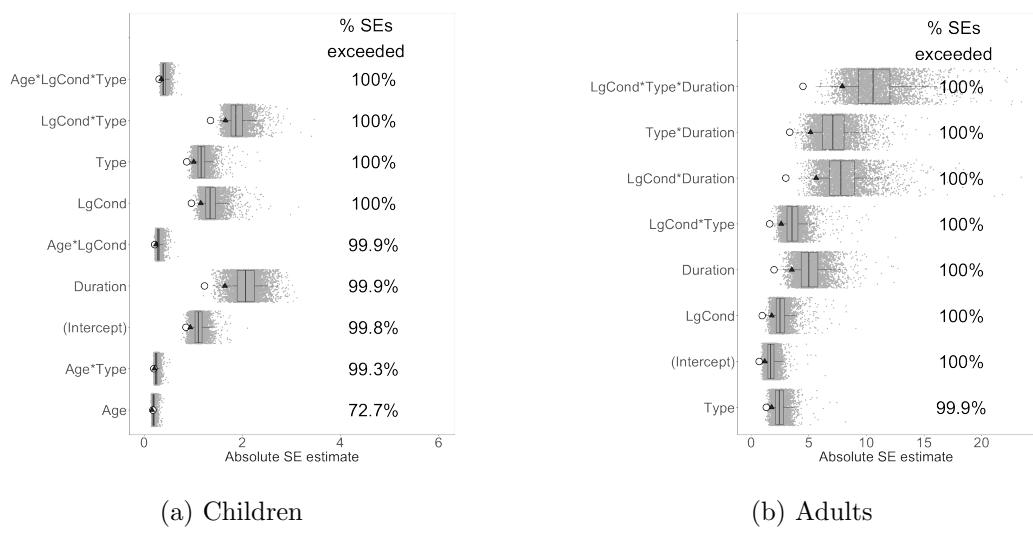


Figure A.3: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.

1346 Appendix A.2. Experiment 2

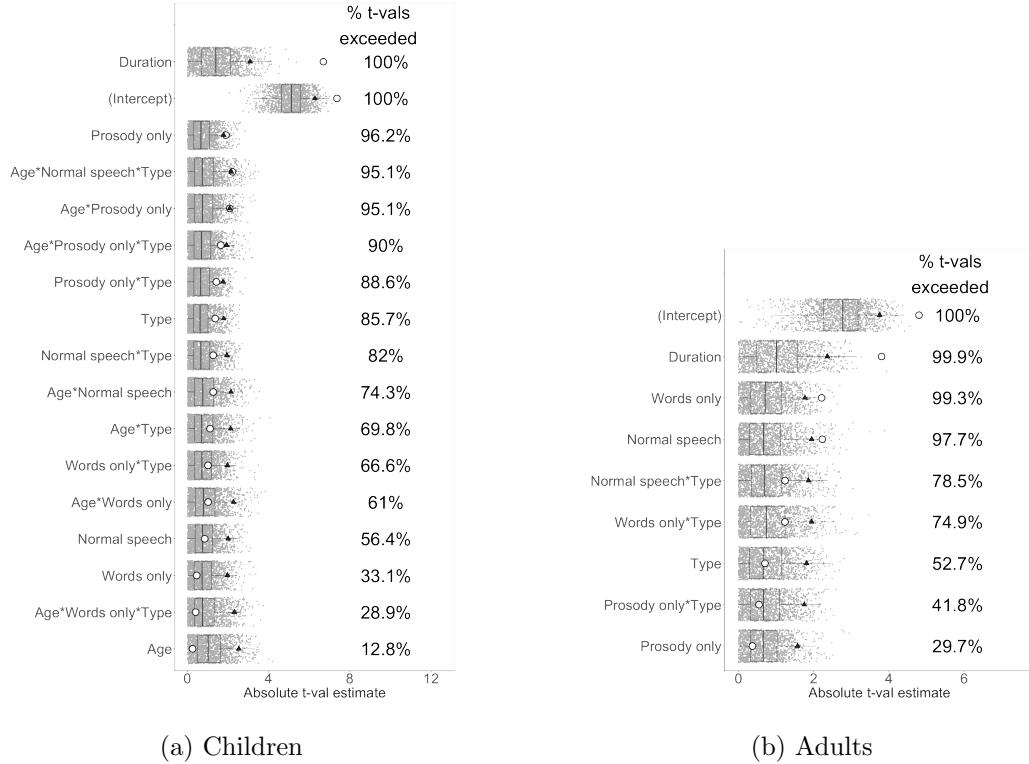


Figure A.4: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

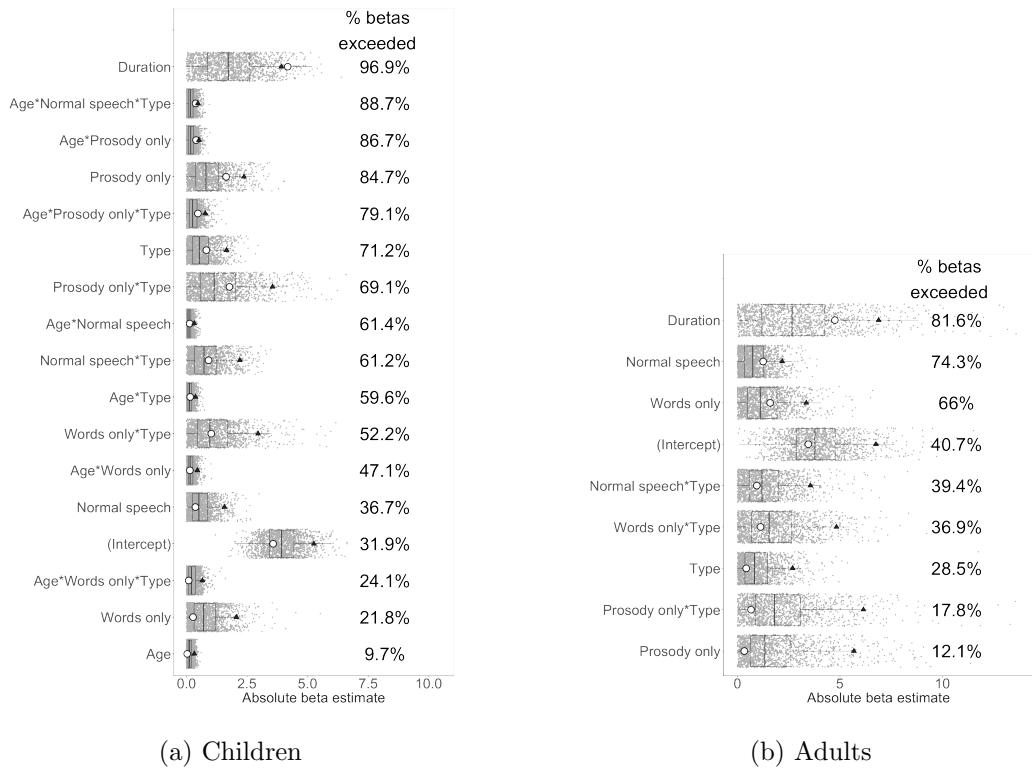


Figure A.5: Random-permutation and original $|\beta\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

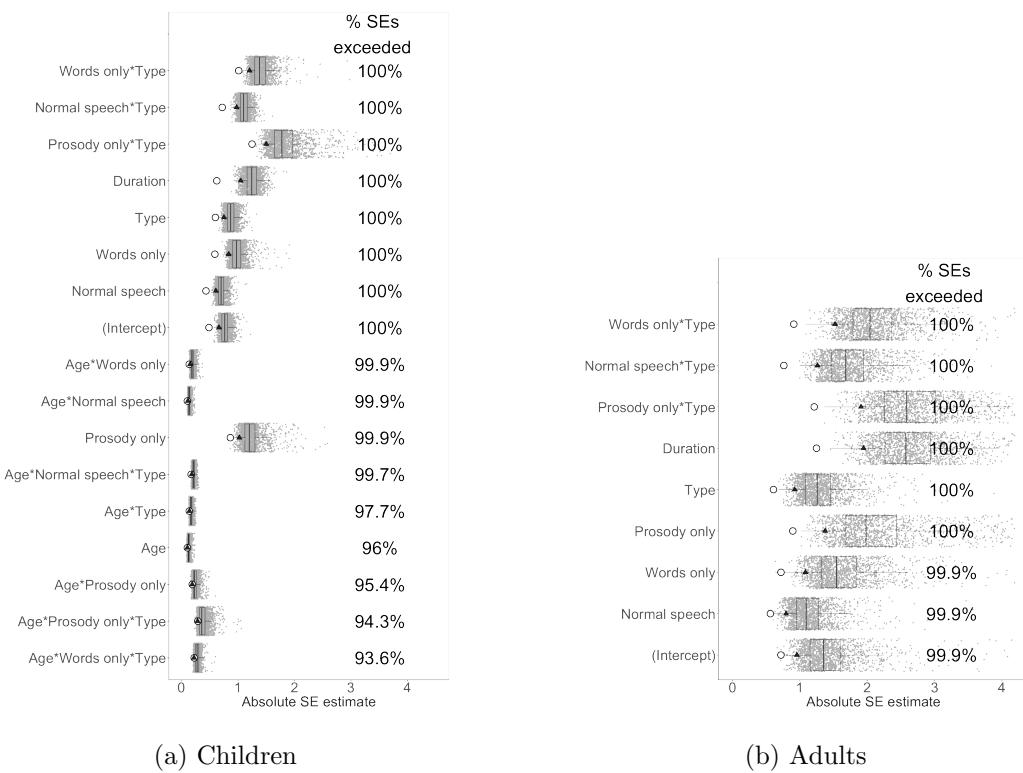


Figure A.6: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

1347 **Appendix B. Pairwise developmental tests: Real vs. randomly
1348 permuted effects**

1349 In each of the plots below, the dot represents the original data value for
1350 the effect and the 5,000 randomly permuted data effect sizes are shown in the
1351 distribution. The percentage shown is the percentage of random permutation
1352 values exceeded by the original data value (taking the absolute value of all
1353 data points for a two-tailed test.)

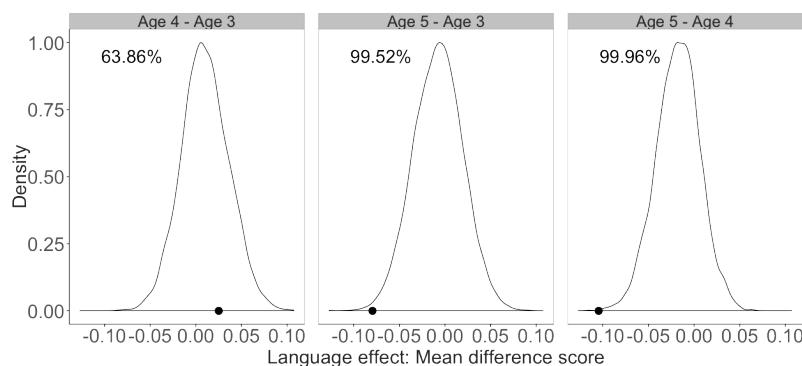


Figure B.1: Pairwise comparisons of the language condition effect across ages
in Experiment 1.

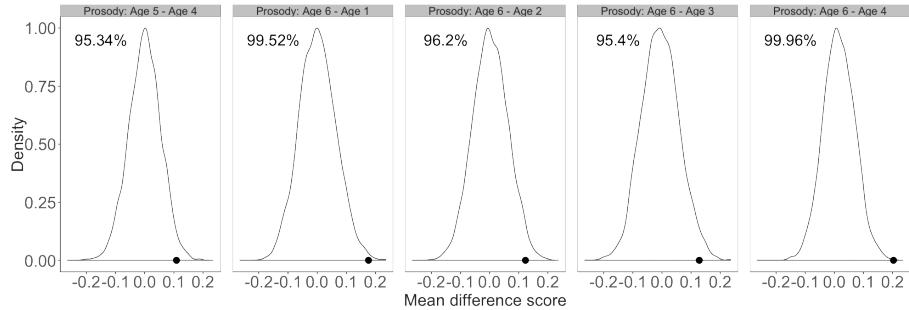


Figure B.2: Significant pairwise comparisons of the *prosody only-no speech* linguistic condition effect, across ages in Experiment 2

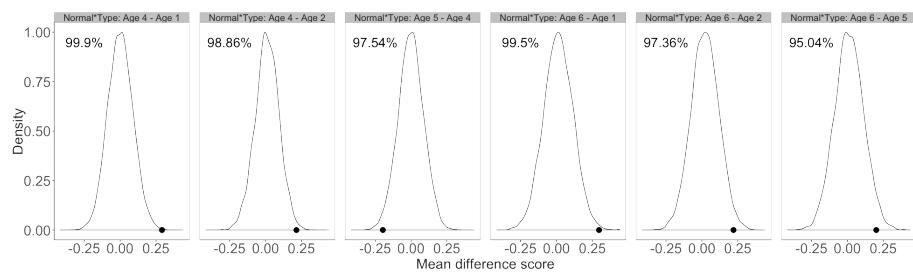


Figure B.3: Significant pairwise comparisons of the *normal speech-no speech* language condition effect for transition type, across ages, in Experiment 2.

1354 **Appendix C. Non-convergent models**

1355 Non-convergent models made up 22–24% of the 5,000 models of randomly
 1356 permuted data in Experiment 1 and 69–70% of the 5,000 for Experiment 2.
 1357 We excluded these non-convergent models because they displayed erratic

1358 β and SE estimates, as summarized below in the table of t -values from
1359 convergent and non-convergent models in Experiment 1. The non-convergent
1360 models from Experiment 2 showed similar patterns. The high frequency
1361 of problematic models persisted even when we changed optimizers and we
1362 suspect the issue derives from data sparsity in some of the random runs.

Variable	Mean	Median	SD	Minimum	Maximum
<i>Children</i>					
(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)
<i>Adults</i>					
(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 the adult and child models from Experiment 1. Converging ('C') and non-converging ('non-C') values are shown as: 'C (non-C)'.

1363 **Appendix D. Miscellaneous**

1364 One alternative hypothesis for children's anticipatory gazes is that they
1365 simply grow bored and start looking away at a constant rate after a turn
1366 begins. This data plotted here show a hypothetical group of participants
1367 who begin to lose interest (at a linear rate) after one second of a turn (gray
1368 dots) compared to participants' real data from Experiment 2 (black dots).
1369 This pattern suggests that, though children do look away with time, their
1370 looks away are not simply driven by boredom.

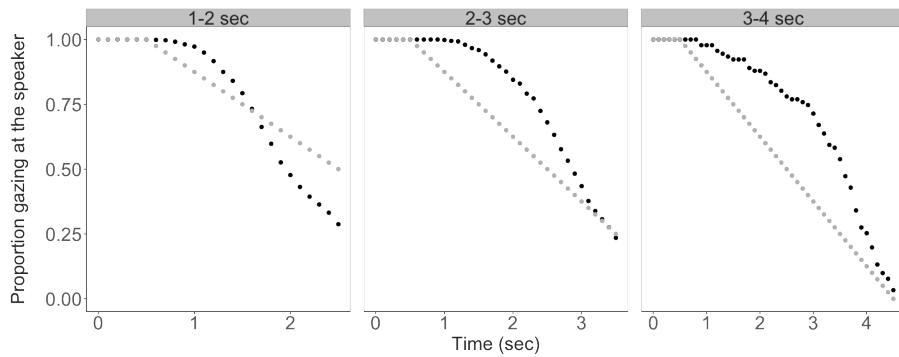


Figure D.1: Proportion participants looking at the current speaker: hypothetical boredom-driven data (gray dots) versus real data from Experiment 2 (black dots).

1371 The design for Experiment 2 does not fully cross puppet pair (e.g., robots,

1372 blue puppets) with linguistic condition (e.g., “words only” and “no speech”).
1373 Even though each puppet pair is associated with different conversation clips
1374 across children (e.g. robots talking about kitties, birthday parties, or pan-
1375 cakes), robots were only associated with “words only” speech, merpeople
1376 were only associated with “prosody only” speech, and the puppets with fancy
1377 clothes were only associated with the “no speech” condition. We did this to
1378 increase the pragmatic felicity of the experiments for the older children (i.e.,
1379 robots make robot sounds, merpeople’s voices are muffled under water, the
1380 fancy-clothed puppets are in a room with main other voices). It is therefore
1381 fair to point out a possible confound between linguistic condition and puppet
1382 pair. Thankfully, we also ran a short follow-up study at the museum with
1383 3–5-year-olds in which each child only saw one video—the normal speech
1384 conversation about birthday parties—with a randomly assigned puppet pair
1385 performing the conversation. Five children watched each puppet pair, for a
1386 total of 30 children across the six pairs. This experiment holds all things
1387 constant except for the appearance of the puppets. We then used a mixed
1388 effects logistic regression of children’s anticipatory switches (yes or no at each
1389 transition), with puppet pair (robots/merpeople/fancy dress/other-3; Figure
1390 1) as a fixed effect and participant and turn transition as random effects.

1391 In four versions of this model we systematically varied the reference level to
1392 check for differences between every puppet pair, finding no significant affects
1393 of puppet type on switching rate. We take this as evidence that, although
1394 we did not fully cross puppet pairs and linguistic conditions in Experiment
1395 2, it was unlikely to have had strong effects on children's looking rates above
1396 and beyond the intended effects of linguistic condition.

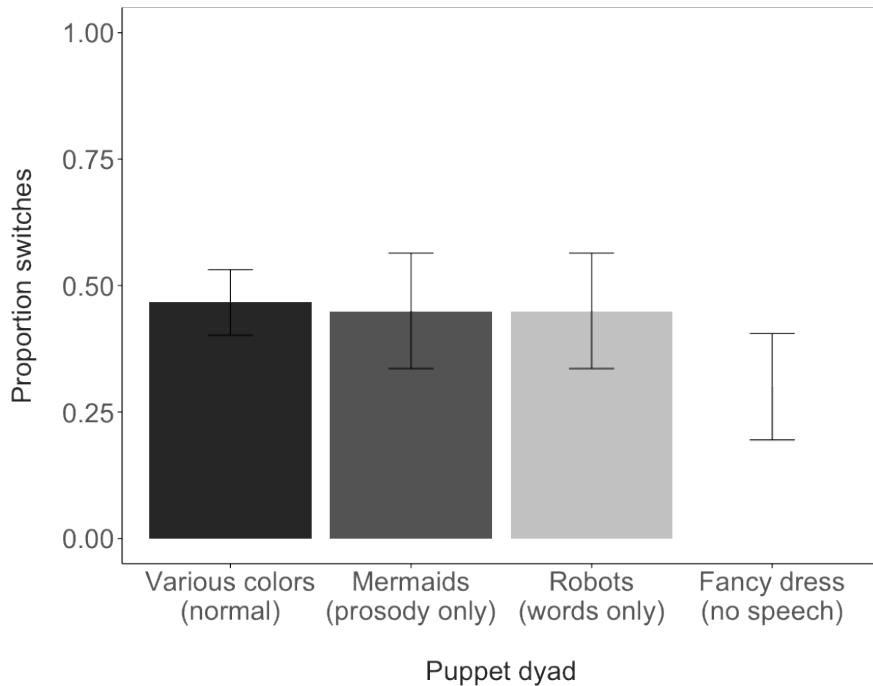


Figure D.2: Proportion gaze switches across puppet pairs when linguistic condition and conversation are held constant.