

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

Marisa Casillas^{a,*}, Michael C. Frank^b

^a*Max Planck Institute for Psycholinguistics, Nijmegen*

^b*Department of Psychology, Stanford University*

Abstract

Children begin developing turn-taking skills in infancy but take several years to assimilate their growing knowledge of language into their turn-taking behavior. In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling linguistic cues to upcoming turn structure. In Experiment 1, we showed English and non-English conversations to English-speaking adults and children. In Experiment 2, we phonetically controlled lexicosyntactic and prosodic cues in English-only speech. Children's predictive looking behavior improved from ages one to six, but even one-year-olds made more anticipatory looks than would be expected by chance. In both experiments, children and adults anticipated more often after hearing questions. Like adults, prosody alone did not improve children's predictive gaze shifts. But, unlike adults, lexical cues alone were also not sufficient to improve prediction—children's performance was best overall with access to lexicosyntax and prosody together. Our findings support an account in which turn prediction emerges in infancy, but becomes fully integrated with linguistic processing only gradually.

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

¹ 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

*Corresponding author

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).

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

23 In this study, we investigate when children begin to make predictions
24 about upcoming turn structure in conversation, and how they integrate lan-
25 guage into their predictions as they grow older. In the remainder of the
26 introduction, we first give a basic review of turn-taking research and the
27 state of current knowledge about adult turn prediction. We then discuss
28 recent work on the development of turn-taking skills before turning to the
29 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 precise in its timing.
50 Across a diverse sample of conversations in 10 languages, one study found
51 a consistent average turn transition time of 0–200 msec at points of speaker
52 switch (Stivers et al., 2009). Experimental results and current models of
53 speech production suggest that it takes approximately 600 msec to produce
54 a content word, and even longer to produce a simple utterance (Griffin and
55 Bock, 2000; Levelt, 1989). So in order to achieve 200 msec turn transitions,
56 speakers must begin formulating their response before the prior turn has
57 ended (Levinson, 2013). Moreover, to formulate their response early on,
58 speakers must track and anticipate what types of response might become
59 relevant next. They also need to predict the content and form of upcoming
60 speech so that they can launch their articulation at exactly the right moment.
61 Prediction thus plays a key role in timely turn taking.

62 Adults have a lot of information at their disposal to help make accurate
63 predictions about upcoming turn content. Lexical, syntactic, and prosodic
64 information (e.g., *wh*- words, subject-auxiliary inversion, and list intonation)
65 can all inform addressees about upcoming linguistic structure (De Ruiter
66 et al., 2006; Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015).
67 Non-verbal cues (e.g., gaze, posture, and pointing) often appear at turn-
68 boundaries and can sometimes act as late indicators of an upcoming speaker
69 switch (Rossano et al., 2009; Stivers and Rossano, 2010). Additionally, the
70 sequential context of a turn can make it clear what will come next: An-
71 swers after questions, thanks or denial after compliments, et cetera (Schegloff,
72 2007).

73 Prior work suggests that adult listeners primarily use lexicosyntactic in-
74 formation to accurately predict upcoming turn structure (De Ruiter et al.,
75 2006). De Ruiter and colleagues (2006) asked participants to listen to snip-
76 plets of spontaneous conversation and to press a button whenever they antici-
77 pated that the current speaker was about to finish his or her turn. The speech
78 snippets were controlled for the amount of linguistic information present;

79 some were normal, but others had flattened pitch, low-pass filtered speech,
80 or further manipulations. With pitch-flattened speech, the timing of par-
81 ticipants' button responses was comparable to their timing with the full
82 linguistic signal. But when no lexical information was available, partici-
83 pants' responses were significantly earlier. The authors concluded that lex-
84 icosyntactic information¹ was necessary and possibly sufficient for turn-end
85 projection, while intonation was neither necessary nor sufficient. Congru-
86 ent evidence comes from studies varying the predictability of lexicosyntactic
87 and pragmatic content: Adults anticipate turn ends better when they can
88 more accurately predict the exact words that will come next (Magyari and
89 De Ruiter, 2012; see also Magyari et al., 2014). They can also identify speech
90 acts within the first word of an utterance (Gísladóttir et al., 2015), allowing
91 them to start planning their response at the first moment possible (Bögels
92 et al., 2015).

93 Despite this body of evidence, the role of prosody for adult turn predic-
94 tion is still a matter of debate. De Ruiter and colleagues' (2006) experiment
95 focused on the role of intonation, which is only a partial index of prosody.
96 And in addition, prosody is tied closely to the syntax of an utterance, so
97 the two linguistic signals are difficult to control independently (Ford and
98 Thompson, 1996). Torreira, Bögels and Levinson (2015) used a combination
99 of button-press and verbal responses to investigate the relationship between
100 lexicosyntactic and prosodic cues in turn-end prediction. Critically, their
101 stimuli were cross-spliced so that each item had full prosodic cues to accom-
102 pany the lexicosyntax. Because of the splicing, they were able to create items
103 that had syntactically-complete units with no intonational phrase boundary
104 at the end. Participants never verbally responded or pressed the "turn-end"
105 button when hearing a syntactically-complete phrase without an intonational
106 phrase boundary. And when intonational phrase boundaries were embedded
107 in multi-utterance turns, participants were tricked into pressing the "turn-
108 end" button 29% of the time. Their results suggest that listeners actually
109 do rely on prosodic cues to execute a response (see also de De Ruiter et al.
110 (2006):525). These experimental findings corroborate other corpus and ex-
perimental work promoting a combination of cues (lexicosyntactic, prosodic,

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

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

114 *1.2. Children's turn prediction*

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

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

139 While children, like adults, could use linguistic cues in the ongoing turn to
140 make predictions about upcoming turn structure, studies of early linguistic
141 development point to a possible early advantage for prosody over lexicosyn-
142 tax in children's turn-taking predictions. Infants can distinguish their native
143 language's rhythm type from others soon after birth (Mehler et al., 1988;
144 Nazzi and Ramus, 2003); they show preference for the typical stress pat-
145 terns of their native language over others by 6–9 months (e.g., iambic vs.
146 trochaic), and can use prosodic information to segment the speech stream
147 into smaller chunks from 8 months onward (Johnson and Jusczyk, 2001;
148 Morgan and Saffran, 1995). Four- to five-month-olds also prefer pauses in

149 speech to be inserted at prosodic boundaries, and by 6 months they can start
150 using prosodic markers to pick out sub-clausal syntactic units, both of which
151 are useful for extracting turn structure from ongoing speech (Jusczyk et al.,
152 1995; Soderstrom et al., 2003). In comparison, children show at best a very
153 limited lexical inventory before their first birthday (Bergelson and Swingley,
154 2013; Shi and Melancon, 2010).

155 Keitel and colleagues (2013) were one of the first to explore how children
156 use linguistic cues to predict upcoming turn structure. They asked 6-, 12-,
157 24-, 36-month-old, and adult participants to watch short videos of conver-
158 sation and tracked their eye movements at points of speaker change. They
159 showed their participants two types of conversation videos—one normal and
160 one with flattened pitch (i.e., with flattened intonation contours)—to test
161 the role of intonation in participants’ anticipatory predictions about upcom-
162 ing speech. Comparing children’s anticipatory gaze frequency to a random
163 baseline, they found that only 36-month-olds and adults made anticipatory
164 gaze switches more often than expected by chance. Among those, only 36-
165 month-olds were affected by a lack of intonation contours, leading Keitel and
166 colleagues to conclude that children’s ability to predict upcoming turn struc-
167 ture relies on their ability to comprehend the stimuli lexicosemantically. They
168 also suggested that intonation might play a secondary role in turn predic-
169 tion, but only after children acquire more sophisticated, adult-like language
170 comprehension abilities.

171 Although the Keitel et al. (2013) study constitutes a substantial advance
172 over previous work in this domain, it has some limitations. Because these
173 limitations directly inform our own study design, we review them in some
174 detail. First, their estimates of baseline gaze frequency (“random” in their
175 terminology) were not random. Instead, they used gaze switches during
176 ongoing speech as a baseline. But ongoing speech is perhaps the period in
177 which switching is least likely to occur (Hirvenkari et al., 2013)—thus, this
178 particular baseline maximizes the chance of finding a difference between gaze
179 frequency at turn transitions and baseline. A more conservative baseline
180 would be to compare participants’ looking behavior at turn transitions to
181 their looking behavior during randomly selected windows of time throughout
182 the stimulus, including turn transitions. We follow this conservative approach
183 in our work.

184 Second, the conversation stimuli Keitel et al. (2013) used were somewhat
185 unusual. The average gap between turns was 900 msec, which is much longer
186 than typical adult timing, where gaps average around 200 msec (Stivers et al.,

2009). The speakers in the videos were also asked to minimize their movements while performing a scripted and adult-directed conversation, which would have created a somewhat unnatural stimulus. Additionally, in order to produce more naturalistic conversation, it would have been ideal to localize the sound sources for the two voices in the video (i.e., to have the voices come out of separate left and right speakers). But both voices were recorded and played back on the same audio channel, which may have made it more difficult to distinguish the two talkers (again, we attempt to address these issues in our current study). Despite these minor methodological issues, the Keitel et al. (2013) study still demonstrates intriguing age-based differences in children’s ability to predict upcoming turn structure. Our current work thus takes this paradigm as a starting point.²

199 *1.3. The current study*

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

210 Because the results of our experiments are complex, we highlight three primary findings. First, although children and adults use linguistic cues to 212 make predictions about upcoming turn structure, they do so primarily in 213 predict speaker transitions after questions (a speech act effect). This speech 214 act effect, which we did not initially predict, is intriguing and suggests that 215 previous work may have neglected an important dimension of linguistic cues. 216 Second, we find that children make more predictions than expected by chance 217 starting at age two, but that this effect is small with our stimuli. Third, we 218 find no evidence of an early prosody advantage in children’s anticipations 219 and, further, no evidence that prosodic or lexical cues alone can substitute 220 for their combination in the full linguistic signal, as is proposed for adults

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

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

225 2. Experiment 1

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

238 2.1. Methods

239 2.1.1. Participants

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

249 2.1.2. Materials

250 *Video recordings.* We recorded pairs of talkers while they conversed in
251 a sound-attenuated booth (see sample frame in Figure 1). Each talker was
252 a native speaker of the language being recorded, and each talker pair was
253 male-female. Using a Marantz PMD 660 solid state field recorder, we cap-
254 tured audio from two lapel microphones, one attached to each participant,



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

255 while simultaneously recording video from the built-in camera of a MacBook
256 laptop computer. The talkers were volunteers and were acquainted with their
257 recording partner ahead of time.

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

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

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.

275 came out of separate speakers during video playback. This gave each voice in
276 the videos a localized source (from the left or right loudspeaker). We coded
277 each turn transition in the videos for language condition (English vs. non-
278 English), inter-turn gap duration (in milliseconds), and speech act (question
279 vs. non-question). The non-English stimuli were coded for speech act from
280 a monolingual English-speaker’s perspective, i.e., which turns “sound like”
281 questions, and which don’t: We asked five native American English speakers
282 to listen to the audio signal for each turn and judge whether it sounded
283 like a question. We then coded turns with at least 80% “yes” responses as
284 questions.

285 Because the conversational stimuli were recorded semi-spontaneously, the
286 duration of turn transitions and the number of speaker transitions in each
287 video was variable. We measured the duration of each turn transition from
288 the audio recording associated with each video. We excluded turn transi-
289 tions longer than 550 msec and shorter than 90 msec, including over-
290 lapped transitions, from analysis.⁴ This left approximately equal numbers
291 of turn transitions available for analysis in the English (N=20) and non-
292 English (N=16) videos. On average, the inter-turn gaps for English videos
293 (mean=318, median=302, stdev=112 msec) were slightly longer than for non-
294 English videos (mean=286, median=251, stdev=122 msec). The longer gaps
295 in the English videos could give them a slight advantage: Our definition of
296 an “anticipatory gaze shift” includes shifts that are initiated during the gap
297 between turns (Figure 2), so participants had slightly more time to make
298 anticipatory shifts in the English videos.

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

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.

306 2.1.3. *Procedure*

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

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

323 Children were less inclined to simply sit and watch videos of conversation
324 in languages they didn't speak, so we used a different procedure to keep them
325 engaged: The experimenter started each session by asking the child about
326 what languages he or she could speak, and about what other languages he
327 or she had heard of. Then the experimenter expressed her own enthusiasm
328 for learning about new languages, and invited the child to watch a video
329 about "new and different languages" together. If the child agreed to watch,
330 the experimenter and the child sat together in front of the display, with
331 the child centered in front of the tracker and the experimenter off to the
332 side. Each conversation video was preceded and followed by a 15–30 second
333 attention-getting filler video (e.g., running puppies, singing muppets, flying
334 bugs). If the child began to look bored, the experimenter would talk during
335 the fillers, either commenting on the previous conversation ("That was a neat
336 language!") or giving the language name for the next conversation ("This
337 next one is called Hebrew. Let's see what it's like.") The experimenter's
338 comments reinforced the video-watching as a joint task.

339 All participants (child and adult) completed a five-point calibration rou-
340 tine before the first video started. We used a dancing Elmo for the children's
341 calibration image. During the experiment, participants watched all six 30-
342 second conversation videos. The first and last conversations were in American

343 English and the intervening conversations were Hebrew, Japanese, German,
344 and Korean. The presentation order of the non-English videos was shuffled
345 into four lists, which participants were assigned to randomly. The entire
346 experiment, including instructions, took 10–15 minutes.

347 *2.1.4. Data preparation and coding*

348 To determine whether participants predicted upcoming turn transitions,
349 we needed to define a set of criteria for what counted as an anticipatory gaze
350 shift. Prior work using similar experimental procedures has found that adults
351 and children make anticipatory gaze shifts to upcoming talkers within a wide
352 time frame; the earliest shifts occur before the end of the prior turn, and the
353 latest occur after the onset of the response turn, with most shifts occurring
354 in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz,
355 2011). Following prior work, we measured how often our participants shifted
356 their gaze from the prior to the upcoming speaker *before* the shift in gaze
357 could have been initiated in reaction to the onset of the speaker’s response.
358 In doing so, we assumed that it takes participants 200 msec to plan an eye
359 movement, following standards from adult anticipatory processing studies
360 (e.g., Kamide et al., 2003).

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

373 As mentioned, most anticipatory switches happen in the inter-turn gap,
374 but we also allowed anticipatory gaze switches that occurred in the final
375 syllables of the prior turn. Early switches are consistent with the distribution
376 of responses in explicit turn-boundary prediction tasks. For example, in
377 a button press task, adult participants anticipate turn ends approximately
378 200 msec in advance of the turn’s end, and anticipatory responses to pitch-
379 flattened stimuli come even earlier (De Ruiter et al., 2006). We therefore

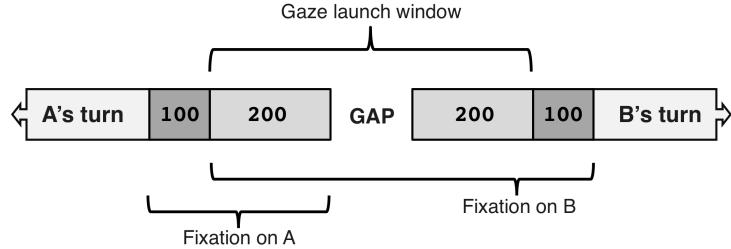


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition.

allowed switches to occur as early as 200 msec before the end of the prior turn. For very early and very late switches, our requirement for 100 msec of fixation on each speaker would sometimes extend outside of the transition window boundaries (200 msec before and after the inter-turn gap). The maximally available fixation window was 100 msec before and after the earliest and latest possible switch point (300 msec before and after the inter-turn gap). We did not count switches made during the fixation window as anticipatory. We *did* count switches made during the inter-turn gap. The period of time from the beginning of the possible fixation window on the prior speaker to the end of the possible fixation window on the responder was our total analysis window (300 msec + the inter-turn gap + 300 msec).

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

2.2. Results

Participants looked at the screen most of the time during video playback (81% and 91% on average for children and adults, respectively). They primarily kept their eyes on the person who was currently speaking in both English and non-English videos: They gazed at the current speaker between 38% and 63% of the time, looking back at the addressee between 15% and

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

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

405 20% of the time (Table 1). Even three-year-olds looked more at the current
 406 speaker than anything else, whether the videos were in a language they could
 407 understand or not. Children looked at the current speaker less than adults
 408 did during the non-English videos. Despite this, their looks to the addressee
 409 did not increase substantially in the non-English videos, indicating that their
 410 looks away were probably related to boredom rather than confusion about
 411 ongoing turn structure. Overall, participants' pattern of gaze to current
 412 speakers demonstrated that they performed basic turn tracking during the
 413 videos, regardless of language. Figure 3 shows participants' anticipatory gaze
 414 rates across age, language condition, and transition type.

415 *2.2.1. Statistical models*

416 We identified anticipatory gaze switches for all 36 usable turn transitions,
 417 based on the criteria outlined in Section 2.1.4, and analyzed them for effects
 418 of language, transition type, and age with two mixed-effects logistic regres-
 419 sions (Bates et al., 2014; R Core Team, 2014). We built one model each
 420 for children and adults. We modeled children and adults separately because
 421 effects of age are only pertinent to the children's data. The child model
 422 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

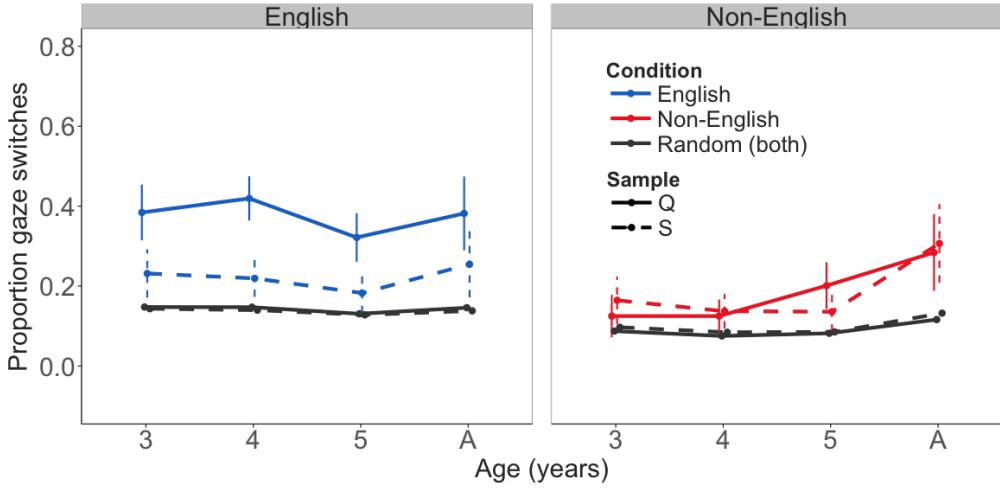


Figure 3: Anticipatory gaze rates across language condition and transition type for the real (red and blue) and randomly permuted baseline (gray). Vertical bars represent 95% confidence intervals.

non-question), age (3, 4, 5; numeric), and duration of the inter-turn gap (seconds, e.g., 0.441) as predictors, with full interactions between condition, transition type, and age. We included the duration of the inter-turn gap as a predictor since longer gaps provide more opportunities to make anticipatory switches (Figure 2). We additionally included random effects of item (turn transition) and participant, with random slopes of condition, transition type, and their interaction for participants (Barr et al., 2013).⁶ The adult model included condition, transition type, duration, and their interactions as predictors with participant and item included as random effects and random slopes of condition, transition type, and their interaction for participant.

Children’s anticipatory gaze switches showed effects of language condition

as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

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.

Children

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.96145	0.84915	-1.132	0.257531
Age	-0.18268	0.17509	-1.043	0.296764
LgCond= <i>non-English</i>	-3.29349	0.96055	-3.429	0.000606 ***
Type= <i>non-Question</i>	-1.10131	0.86520	-1.273	0.203055
Duration	3.40171	1.22878	2.768	0.005634 **
Age*LgCond= <i>non-English</i>	0.52066	0.21192	2.457	0.014015 *
Age*TypeS= <i>non-Question</i>	-0.01628	0.19442	-0.084	0.933262
LgCond= <i>non-English</i> *	2.68171	1.35045	1.986	0.047057 *
Type= <i>non-Question</i>				
Age*LgCond= <i>non-English</i> *	-0.45633	0.30168	-1.513	0.130378
Type= <i>non-Question</i>				

Adults

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1966	0.6945	-0.283	0.777062
LgCond= <i>non-English</i>	-0.8812	0.9607	-0.917	0.359028
Type= <i>non-Question</i>	-4.4953	1.3147	-3.419	0.000628 ***
Duration	-1.1227	1.9889	-0.565	0.572414
LgCond= <i>non-English</i> *	3.2972	1.6115	2.046	0.040747 *
Type= <i>non-Question</i>				
LgCond= <i>non-English</i> *	1.3625	3.0097	0.453	0.650749
Duration				
Type= <i>non-Question</i> *	10.5107	3.3482	3.139	0.001694 **
Duration				
LgCond= <i>non-English</i> *	-6.3156	4.4969	-1.404	0.160191
Type= <i>non-Question</i> *				
Duration				

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

434 ($\beta=-3.29$, $SE=0.961$, $t=-3.43$, $p<.001$) and gap duration ($\beta=3.4$, $SE=1.229$,
 435 $t=2.77$, $p<.01$) with additional effects of an age-by-language condition
 436 interaction ($\beta=0.52$, $SE=0.212$, $t=2.46$, $p<.05$) and a language condition-by-
 437 transition type interaction ($\beta=2.68$, $SE=1.35$, $t=1.99$, $p<.05$). There were
 438 no significant effects of age or transition type alone ($\beta=-0.18$, $SE=0.175$,
 439 $t=-1.04$, $p=.3$ and $\beta=-1.10$, $SE=0.865$, $t=-1.27$, $p=.2$, respectively).

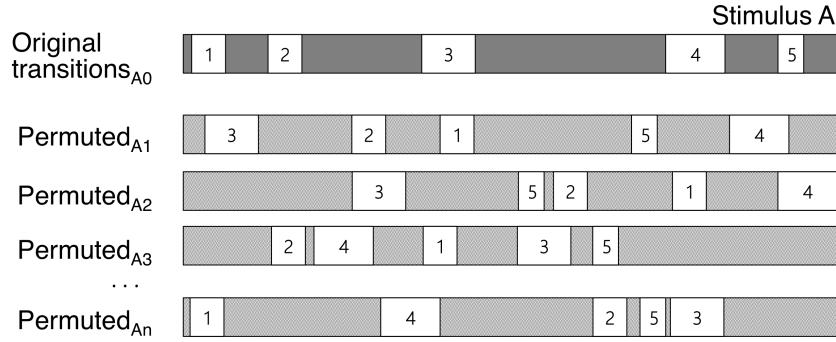


Figure 4: Example of analysis window permutations for a stimulus with five turn transitions. The windows were ± 300 msec around the inter-turn gap.

440 Adults' anticipatory gaze switches shows an effect of transition type ($\beta=$
 441 4.5 , $SE=1.314$, $t=-3.42$, $p<.001$) and significant interactions between lan-
 442 guage condition and transition type ($\beta=3.3$, $SE=1.61$, $t=2.05$, $p<.05$) and
 443 transition type and gap duration ($\beta=10.51$, $SE=3.346$, $t=3.141$, $p<.01$).

444 *2.2.2. Random baseline comparison*

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

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

463 related (original) and non-turn-related (randomly permuted) looking behav-
464 ior using the same eye movements. The resulting anticipatory gazes from the
465 randomly permuted analysis windows represent an average anticipatory gaze
466 rate over all possible starting points: a random baseline. By running the real
467 and randomly permuted data sets through identical statistical models, we
468 can also estimate how likely it is that predictor effects in the original data
469 (e.g., the effect of language condition; Table 2) arose from random looking.
470 Because these analyses are complex, we report their full details in Appendix
471 A.

472 Our baseline analyses revealed that none of the significant predictors
473 from models of the original, turn-related data can be explained by random
474 looking. For the children’s data, the original t -values for language condi-
475 tion, gap duration, the age-language condition interaction, and the language
476 condition-transition type interaction were all greater than 95% of t -values for
477 the randomly permuted data (99.9%, 95.5%, 99.4%, and 96%, respectively).
478 Similarly, the adults’ data showed significant differentiation from the ran-
479 domly permuted data for two of the three originally significant predictors—
480 transition type and the transition type-gap duration interaction (greater than
481 99.9% and 99.7% of random t -values, respectively)—with marginal differen-
482 tiation for the interaction of language condition and transition type (greater
483 than 94.6% of random t -values).

484 *2.2.3. Developmental effects*

485 The models reported above revealed a significant interaction of age and
486 language condition (Table 2) that was unlikely be due to random looking
487 (Figure 3). To further explore this effect, we compared the average effect
488 of language condition for each age group: Using the permutation analyses
489 above, we extracted the average difference score for the two language con-
490 ditions (English minus non-English) for each subject, computing an overall
491 average for each random permutation of the data. For each random permu-
492 tation, we then made pairwise comparisons of the average difference scores
493 across participant age groups. Details are given in Appendix B.

494 These analyses showed that, while 3- and 4-year olds showed similarly
495 large effects of language condition, 5-year-olds showed a significantly smaller
496 effect of language condition, compared to both younger age groups. In other
497 words, the difference in the effect of language condition for 5-year-olds com-
498 pared to younger children was larger than would be expected by chance in
499 99.52% of the randomly permuted data sets for 3-year-olds and 99.96% of

500 the data sets for 4-year-olds—differences of $p < .01$ and $p < .001$, respectively
501 (see Figure B.1 for difference score distributions).

502 When does spontaneous turn prediction emerge developmentally during
503 natural speech? To test whether the youngest age group (3-year-olds) already
504 exceeded chance in their anticipatory gaze switches, we used two-tailed t -
505 tests to compare their real gaze rates to the random baseline in the English
506 condition. Although the overall effect was small, we found that three-year-
507 olds made anticipatory gaze switches significantly above chance, when all
508 transitions were considered ($t(22.824) = -4.147, p < .001$) as well as for question
509 transitions alone ($t(21.677) = -5.268, p < .001$).

510 *2.3. Discussion*

511 Children and adults spontaneously tracked the turn structure of the con-
512 versations, making anticipatory gaze switches at an above-chance rate across
513 all ages and conditions. Children’s anticipatory gaze rates were affected by
514 language condition, transition type, age, and gap duration (Table 2), none
515 of which could be explained by a baseline of random gaze switching (Figure
516 A.1a). These data show a number of important features that bear on our
517 questions of interest.

518 First, both adults’ and children’s anticipations were strongly affected by
519 transition type. Both groups made more anticipatory switches after hearing
520 questions, compared to non-questions. Even in the English videos, when
521 participants had full access to linguistic cues, their rates of anticipation were
522 relatively low—in fact, comparable to the non-English videos—unless the
523 turn was a question. Prior work using online, metalinguistic tasks has shown
524 that participants can use linguistic cues to accurately predict upcoming turn
525 ends (Torreira et al., 2015; Magyari and De Ruiter, 2012; De Ruiter et al.,
526 2006). The current results add a new dimension to our understanding of
527 how listeners make predictions about turn ends: Both children and adults
528 spontaneously monitor the linguistic structure of unfolding turns for cues to
529 upcoming responses.

530 Second, we saw developmental effects such that older children anticipated
531 more reliably, but only in the non-English videos. In the English videos all
532 children anticipated, especially foro questions. But the language condition
533 interaction suggests that the 5-year-olds were able to leverage anticipatory
534 cues in the non-English videos in a way that 3- and 4-year-olds could not,
535 possibly by shifting more attention to the non-native prosodic or non-verbal
536 cues. Prior work on children’s turn-structure anticipation proposed that

537 children's turn-end predictions rely primarily on lexicosyntactic structure
538 (and not, e.g., prosody) as they get older (Keitel et al., 2013). The current
539 results suggest more flexibility in children's predictions; when they do not
540 have access to lexical information, older children and adults are likely to find
541 alternative cues to turn taking behavior.

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

552 3. Experiment 2

553 Experiment 2 used native-language stimuli, controlled for lexical and
554 prosodic information, eliminating non-verbal cues, and tested children from a
555 wider age range. To tease apart the role of lexical and prosodic information,
556 we phonetically manipulated the speech signal for pitch, syllable duration,
557 and lexical access. By testing one- to six-year-olds we hoped to find the de-
558 velopmental onset of turn-predictive gaze. We also hoped to measure changes
559 in the relative roles of prosody and lexicosyntax across development.

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

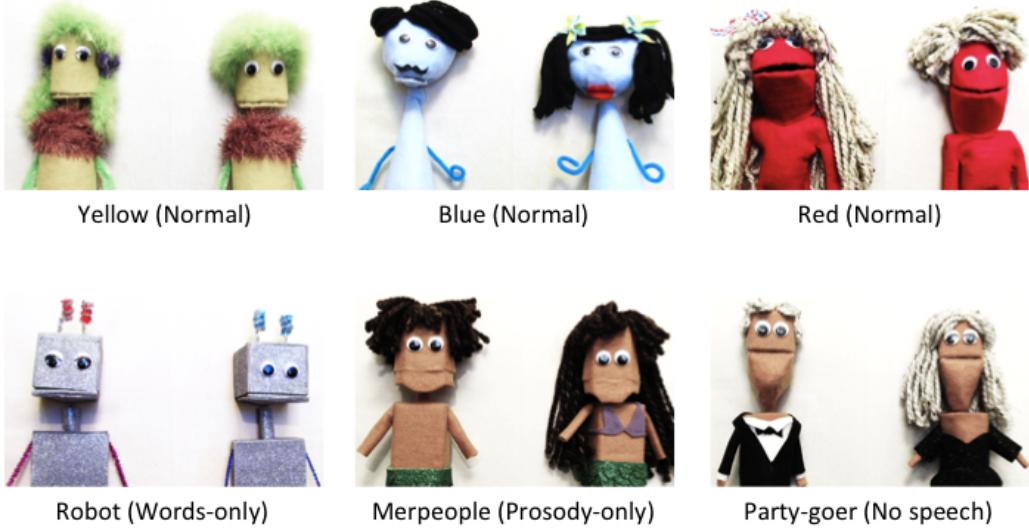


Figure 5: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

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

576 *3.1. Methods*

577 *3.1.1. Participants*

578 We recruited 27 undergraduate adults and 129 children between ages 1;0–
 579 6;11 to participate in our experiment. We recruited our child participants
 580 from Children's Discovery Museum of San Jose, California, targeting approx-
 581 imately 20 children for each of the six one-year age groups (range: 20–23).
 582 All participants were native English speakers, though some parents (N=27)
 583 reported that their child heard a second (and sometimes third) language at
 584 home. None of the adult participants reported fluency in a second language.
 585 We ran Experiment 2 at a local children's museum because it gave us access
 586 to children with a more diverse range of ages.

587 3.1.2. *Materials*

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

593 *Audio recordings.* The recording session was set up in the same way as
594 the first experiment, but with a shorter warm up period (5–10 minutes) and
595 a pre-determined topic for the child-friendly improvisation ('riding bikes',
596 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the
597 talkers were native English speakers, and were recorded in male-female pairs.
598 As before, we asked talkers to speak "as if they were on a children's television
599 show" and to ask at least a few questions during the improvisation. We cut
600 each audio recording down to the 20-second interval with the most turn
601 activity. The 20-second clips were then phonetically manipulated and used
602 in the final video stimuli.

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

608 The *words only* clips were manipulated to have robot-like speech: We
609 flattened the intonation contours to each talker's average pitch (F0) and
610 we reset the duration of every nucleus and coda to each talker's average
611 nucleus and coda duration.⁷ We made duration and pitch manipulations
612 using PSOLA resynthesis in Praat (Boersma and Weenink, 2012). Thus,
613 the *words only* versions of the audio clips had no pitch or durational cues
614 to upcoming turn boundaries, but did have intact lexicosyntactic cues (and
615 residual phonetic correlates of prosody, e.g., intensity).

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

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

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

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

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

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

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

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

659 spontaneously. We measured turn transitions from the audio recording of
660 the *normal*, *words only*, and *prosody only* conditions. There was no audio
661 from the original conversation in the *no speech* condition videos, so we mea-
662 sured turn transitions from the video recording, using ELAN video editing
663 software (Wittenburg et al., 2006).

664 There were 85 turn transitions for analysis after excluding transitions
665 longer than 550 msec and shorter than 90 msec. The remaining turn tran-
666 sitions had slightly more questions than non-question ($N=50$ and $N=35$, re-
667 spectively), with transitions distributed somewhat evenly across conditions
668 (keeping in mind that there were three *normal* videos and only one lim-
669 ited cue video for each experiment version): *normal* ($N=36$), *words only*
670 ($N=13$), *prosody only* ($N=17$), and *no speech* ($N=19$). Inter-turn gaps for
671 questions (mean=365, median=427) were longer than those for non-questions
672 (mean=302, median=323) on average, but gap duration was overall com-
673 parable across conditions: *normal* (mean=334, median=321), *words only*
674 (mean=347, median=369), *prosody only* (mean=365, median=369), and *no*
675 *words* (mean=319, median=329). The longer gaps for question transitions
676 could give them an advantage because our anticipatory measure includes
677 shifts initiated during the gap between turns (Figure 2).

678 3.2. Procedure

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

687 3.2.1. Data preparation and coding

688 We coded each turn transition for its linguistic condition (*normal*, *words*
689 *only*, *prosody only*, and *no speech*) and transition type (question/non-question)⁸
690 and identified anticipatory gaze switches to the upcoming speaker using the
691 methods from Experiment 1.

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

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

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

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

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

692 3.3. Results

693 Participants' pattern of gaze indicated that they performed basic turn
 694 tracking across all ages and in all conditions. Participants again looked at
 695 the screen most of the time during video playback (82% and 86% average
 696 for children and adults, respectively), primarily looking at the person who
 697 was currently speaking (Table 2). They tracked the current speaker in every
 698 condition—even one-year-olds looked more at the current speaker than at
 699 anything else in the three limited cue conditions (40% for *words only*, 43%
 700 for *prosody only*, and 39% for *no speech*). There was a steady overall increase
 701 in looks to the current speaker with age and added linguistic information
 702 (Tables 3 and 4). Looks to the addressee also decreased with age, but the
 703 change was minimal. Figure 6 shows participants' anticipatory gaze rates
 704 across age, the four language conditions, and transition type.

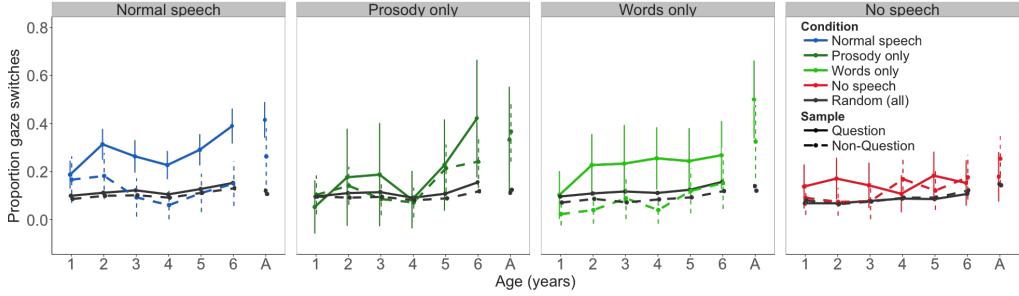


Figure 6: Anticipatory gaze rates across language condition and transition type for the real (blue, dark green, light green, and red) and randomly permuted baseline (gray). Vertical bars represent 95% confidence intervals.

705 3.3.1. Statistical models

706 We identified anticipatory gaze switches for all 85 usable turn transitions,
 707 and analyzed them for effects of language condition, transition type,
 708 and age with two mixed-effects logistic regressions. We again built separate
 709 models for children and adults because effects of age were only pertinent to
 710 the children’s data. The child model included condition (normal/prosody
 711 only/words only/no speech; with no speech as the reference level), transition
 712 type (question vs. non-question), age (1, 2, 3, 4, 5, 6; numeric), and duration
 713 of the inter-turn gap (in seconds) as predictors, with full interactions between
 714 language condition, transition type, and age. We again included the duration
 715 of the inter-turn gap as a control predictor and added random effects of
 716 item (turn transition) and participant, with random slopes of transition type
 717 for participants. The adult model included condition, transition type, their
 718 interactions, and duration as a control predictor, with participant and item
 719 included as random effects and random slopes of condition and transition
 720 type.

Children

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.49403	0.48454	-7.211	5.55e-13 ***
Age	0.02436	0.10249	0.238	0.8121
Type= <i>non-Question</i>	-0.88900	0.61192	-1.453	0.1463
Duration	3.94743	0.61668	6.401	1.54e-10 ***
Age*Type= <i>non-Question</i>	0.15359	0.13996	1.097	0.2725
Condition= <i>normal</i>	0.37337	0.43421	0.860	0.3899

Age*Condition= <i>normal</i>	0.12950	0.10217	1.267	0.2050
Condition= <i>normal</i> *	0.91074	0.72581	1.255	0.2096
Type= <i>non-Question</i>				
Age*Condition= <i>normal</i> *	-0.37965	0.17031	-2.229	0.0258 *
Type= <i>non-Question</i>				
Condition= <i>prosody</i>	-1.60734	0.86680	-1.854	0.0637 .
Age*Condition= <i>prosody</i>	0.39271	0.18905	2.077	0.0378 *
Condition= <i>prosody</i> *	1.68552	1.05414	1.599	0.1098
Type= <i>non-Question</i>				
Age*Condition= <i>prosody</i> *	-0.32360	0.23229	-1.393	0.1636
Type= <i>non-Question</i>				
Condition= <i>words</i>	-0.26996	0.59313	-0.455	0.6490
Age*Condition= <i>words</i>	0.14044	0.13565	1.035	0.3005
Condition= <i>words</i> *	-1.03066	1.01610	-1.014	0.3104
Type= <i>non-Question</i>				
Age*Condition= <i>words</i> *	0.08829	0.22387	0.394	0.6933
Type= <i>non-Question</i>				
Adults				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.3811	0.6884	-4.912	9.03e-07 ***
Type= <i>non-Question</i>	0.4375	0.5854	0.747	0.4549
Duration	4.6961	1.1804	3.978	6.94e-05 ***
Condition= <i>normal</i>	1.2033	0.5359	2.245	0.0247 *
Condition= <i>normal</i> *	-0.9627	0.7358	-1.308	0.1907
Type= <i>non-Question</i>				
Condition= <i>prosody</i>	0.2407	0.8011	0.301	0.7638
Condition= <i>prosody</i> *	0.5525	0.9374	0.589	0.5556
Type= <i>non-Question</i>				
Condition= <i>words</i>	1.5613	0.7087	2.203	0.0276 *
Condition= <i>words</i> *	-1.1557	0.8854	-1.305	0.1918
Type= <i>non-Question</i>				

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

721 Children's anticipatory gaze switches showed an effect of gap duration
 722 ($\beta=4.18$, $SE=0.624$, $t=6.689$, $p<.001$), a two-way interaction of age and
 723 language condition (for *prosody only* speech compared to the *no speech* refer-
 724 ence level; $\beta=0.393$, $SE=0.189$, $t=2.08$, $p<.05$), and a three-way interaction
 725 of age, transition type, and language condition (for *normal* speech compared
 726 to the *no speech* reference level; $\beta=-0.375$, $SE=0.17$, $t=-2.213$, $p<.05$). There
 727 were no significant effects of age or transition type alone (Table 3.3.1), with
 728 only a marginal effect of language condition (for *prosody only* compared to

729 the *no speech* reference level; $\beta=-1.634$, $SE=0.864$, $t=-1.89$, $p=.06$)
730 Adults' anticipatory gaze switches showed effects of gap duration ($\beta=4.75$,
731 $SE=1.248$, $t=3.806$, $p<.001$) and language condition (for *normal* speech
732 $\beta=1.256$, $SE=0.563$, $t=2.229$, $p<.05$ and *words only* speech $\beta=1.594$, $SE=0.721$,
733 $t=2.211$, $p<.05$ compared to the *no speech* reference level). There were no
734 effects of transition type ($\beta=0.429$, $SE=0.609$, $t=0.705$, $p=.48$).

735 *3.3.2. Random baseline comparison*

736 Using the same technique described in Experiment 1 (Section 2.2.2), we
737 created and modeled random permutations of participants' anticipatory gaze.
738 These analyses revealed that none of the significant predictors from models of
739 the original, turn-related data could be explained by random looking. In the
740 children's data, the original model's *t*-values for language condition (*prosody*
741 *only*), gap duration, the two-way interaction of age and language condition
742 (*prosody only*) and the three-way interaction of age, transition type, and
743 language condition (*normal* speech) were all greater than 95% of the ran-
744 domly permuted *t*-values (96.2%, 100%, 95.1%, and 95.1%, respectively).
745 Similarly, the adults' data showed significant differentiation from the ran-
746 domly permuted data for all originally significant predictors: gap duration
747 and language condition for *normal* speech and words-only speech (greater
748 than 100%, 96.8%, and 98.7% of random *t*-values, respectively).

749 *3.3.3. Developmental effects*

750 Our main goal in extending the age range to 1- and 2-year-olds in Ex-
751 periment 2 was to find the age of emergence for spontaneous turn structure
752 predictions in our paradigm. As in Experiment 1, we used two-tailed *t*-tests
753 to compare children's real gaze rates to the random baseline in the *normal*
754 speech condition, in which the speech stimulus is most like what children hear
755 every day. We tested real gaze rates against baseline for three age groups:
756 ages one, two, and three. Two- and three-year-old children made anticipa-
757 tory gaze switches significantly above chance both when all transitions were
758 considered (2-year-olds: $t(26.193)=-4.137$, $p<.001$; 3-year-olds: $t(22.757)=-$
759 2.662 , $p<.05$) and for question transitions alone (2-year-olds: $t(25.345)=$
760 4.269 , $p<.001$; 3-year-olds: $t(21.555)=-3.03$, $p<.01$). One-year-olds, how-
761 ever, made anticipatory gaze shifts that were not statistically significant for
762 turn transitions overall and for question turns alone (overall: $t(24.784)=$
763 2.049 , $p=.051$; questions: $t(25.009)=-2.03$, $p=.053$).

764 Regression models for the children’s data also revealed two significant
765 interactions with age. The first was a significant interaction of age and
766 language condition (for *prosody only* compared to the *no speech* reference
767 level), suggesting a different age effect between the two linguistic conditions.
768 As in Experiment 1, we further explored each age interaction by extracting an
769 average difference score over subjects for the effect of language condition (*no*
770 *speech* vs. *prosody only*) within each random permutation of the data, making
771 pairwise comparisons between the six age groups. These tests revealed that
772 children’s anticipation in the *prosody only* condition significantly improved
773 at ages five and six (difference scores greater than 95% of the random data
774 scores ($p < .05$)).

775 The second age-based interaction was a three-way interaction of age, transi-
776 tion type, and language condition (for *normal* speech compared to the *no*
777 *speech* baseline). We again created pairwise comparisons of the average dif-
778 ference scores for the transition type-language condition interaction across
779 age groups in each random permutation of the data, finding that the effect
780 of transition type in the *normal* speech condition became larger with age,
781 with significant improvements by age 4 over ages 1 and 2 (99.9% and 98.86%,
782 respectively), by age 5 over age 4 (97.54%), and by age 6 over ages 1, 2, and 5
783 (99.5%, 97.36%, and 95.04%), all significantly different from chance ($p < .05$).

784 3.4. Discussion

785 The core aims of Experiment 2 were to gain better traction on the indi-
786 vidual roles of prosody and lexicosyntax in children’s turn predictions, and to
787 find the age of emergence for spontaneous turn anticipation. Taken together,
788 our results replicate the findings from Experiment 1: Participants make more
789 anticipatory switches when they have access to lexical information and, when
790 they do, tend to make more anticipatory switches for questions compared to
791 non-questions.

792 As in Experiment 1, with normal speech, children and adults sponta-
793 neously tracked the turn structure of the conversations, making anticipatory
794 gaze switches at an above-chance rate across all ages. And in addition, they
795 made far more anticipations for questions than for non-question turns—at
796 least for two-year-olds and older. But these effects were different for the
797 two comparison conditions, *prosody-only* and *words-only*. In the *prosody-*
798 *only* condition, performance was low for younger children and increased sub-
799 stantially for older children (especially for questions). In the *words-only*
800 condition, anticipation performance was relatively robust for questions for

801 two-year-olds and older (much like in normal speech), but never rose above
802 chance for the children for non-question turns. Thus, these findings do not
803 support an early role for prosody in children’s spontaneous turn structure
804 predictions. On the contrary, children’s predictions appeared to stem more
805 from lexical information, probably primarily from question words.

806 **4. General Discussion**

807 Children begin to develop conversational turn-taking skills long before
808 their first words (Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001;
809 Snow, 1977). As they acquire language, they also acquire the information
810 needed to make accurate predictions about upcoming turn structure. Until
811 recently, we have had very little data on how children weave language into
812 their already-existing turn-taking behaviors. In two experiments investigating
813 children’s anticipatory gaze to upcoming speakers, we found evidence that
814 turn prediction develops early in childhood and that spontaneous predictions
815 are primarily driven by participants’ expectation of an immediate response
816 in the next turn (e.g., for questions). In making predictions about upcom-
817 ing turn structure, children used a combination of lexical and prosodic cues;
818 neither lexical nor prosodic cues alone were sufficient to support increased
819 anticipatory gaze. We also found no early advantage for prosody over lexi-
820 cosyntax, and instead found that children were unable to make above-chance
821 anticipatory gazes in the *prosody only* condition until age five. We discuss
822 these findings with respect to the role of linguistic cues in predictions about
823 upcoming turn structure, the importance of questions in spontaneous predic-
824 tions about conversation, and children’s developing competence as conversa-
825 tionalists.

826 *4.1. Predicting turn structure with linguistic cues*

827 Prior work with adults has found a consistent and critical role for lex-
828 icosyntax in predicting upcoming turn structure (De Ruiter et al., 2006;
829 Magyari and De Ruiter, 2012), with the role of prosody still under debate
830 (Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015). Knowing
831 that children comprehend more about prosody than lexisyntax early on
832 (see introduction; also see Speer and Ito, 2009 for a review), we thought it
833 possible that young children would instead show an advantage for prosody in
834 their predictions about turn structure in conversation. Our results suggest

835 that, on the contrary, when presented with *only* prosodic information, chil-
836 dren's spontaneous predictions about upcoming turn structure are limited
837 until age five.

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

855 *4.2. The question effect*

856 In both experiments, anticipatory looking was primarily driven by ques-
857 tion transitions, a pattern that had not been previously reported in other an-
858 ticipatory gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari,
859 2013; Tice and Henetz, 2011). Questions make an upcoming speaker switch
860 immediately relevant, helping the listener to predict with high certainty what
861 will happen next (i.e., an answer from the addressee), and are often easily
862 identifiable by overt prosodic and lexicosyntactic cues.

863 Compared to prosodic cues (e.g., final rising intonation), lexicosyntactic
864 cues (e.g., *wh*-words, *do*-insertion, and subject-auxiliary inversion) were fre-
865 quent, categorical, and early-occurring in the utterance. Children may have
866 therefore had an easier time picking out and interpreting lexical cues to ques-
867 tionhood. The question effect showed its first significant gains between ages
868 three and four in the *normal* speech condition of Experiment 2, by which
869 time children frequently hear and use a variety of polar *wh*-questions (Clark,
870 2009). Furthermore, while lexicosyntactic question cues were available on
871 every instance of *wh*- and *yes/no* questions in our stimuli, prosodic question

872 cues were only salient on *yes/no* questions and, even then, the mapping of
873 prosodic contour to speech act (e.g., high final rises for polar questions) is
874 far from one-to-one.

875 Prior work on children’s acquisition of questions indicates that they may
876 already have some understanding about question-answer sequences by the
877 time they begin to speak: Questions make up approximately one third of
878 the utterances children hear, before and after the onset of speech, and even
879 into their preschool years, even though the types and complexity of questions
880 change throughout development (Casillas et al., In press; Fitneva, 2012; Hen-
881 ning et al., 2005; Shatz, 1979).⁹ For the first few years, many of the ques-
882 tions directed to children are “test” questions—questions that the caregiver
883 already has the answer to (e.g., “What does a cat say?”), but this changes
884 as children get older. Questions help caregivers to get their young children’s
885 attention and to ensure that information is in common ground, even if the
886 responses are non-verbal or infelicitous (Bruner, 1985; Fitneva, 2012; Snow,
887 1977). So, in addition to having a special interactive status (for adults and
888 children alike), questions are a core characteristic of many caregiver-child
889 interactions, motivating a general benefit for questions in turn structure an-
890 ticipation.

891 Two important questions for future work are then: (1) How does chil-
892 dren’s ability to monitor for questions in conversation relate to their prior
893 experience with questions? and (2) what is it about questions that makes
894 children and adults more likely to anticipatorily switch their gaze to ad-
895 dressees? Other request formats, such as imperatives, compliments, and
896 complaints make a response from the addressee highly likely in the next turn
897 (Schegloff, 2007). Rhetorical and tag questions, on the other hand, take a
898 similar form to prototypical polar questions, but often do not require an an-
899 swer. So, though it is clear that adults and children anticipated responses
900 more often for questions than non-questions, we do not yet know whether
901 their predictive action is limited to turns formatted as questions or is gener-
902 ally applicable to turn structures that project an immediate response from
903 the addressee.

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

904 4.3. Early competence for turn taking?

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

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

927 4.4. Limitations and future work

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

937 The prediction measure used in these studies is based on an observer’s
938 view of third-party conversation but, because participants’ role in the inter-
939 action could affect their online predictions about turn taking, an ideal exper-
940 imental measure would capture first-person behavior. First-person measures

941 of spontaneous turn prediction will be key to revealing how participants dis-
942 tribute their attention over linguistic and non-verbal cues while taking part
943 in everyday interaction, the implications of which relate to theories of online
944 language processing for both language learning and everyday talk.

945 *4.5. Conclusions*

946 Conversation plays a central role in children’s language learning. It is
947 the driving force behind what children say and what they hear. Adults use
948 linguistic information to accurately predict turn structure in conversation,
949 which facilitates their online comprehension and allows them to respond rel-
950 evantly and on time. The present study offers new findings regarding the
951 role of speech acts and linguistic processing in online turn prediction, and
952 has given evidence that turn prediction emerges by age two is not integrated
953 with linguistic cues until much later. Using language to make predictions
954 about upcoming interactive content takes time and, for both children and
955 adults, is primarily driven by participants’ orientation to what will happen
956 beyond the end of the current turn.

957 **Acknowledgements**

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

969 **References**

- 970 Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure
971 for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory*
972 and *Language* 68, 255–278.

- 973 Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4:
974 Linear mixed-effects models using Eigen and S4. URL:
975 <https://github.com/lme4/lme4><http://lme4.r-forge.r-project.org/>.
976 [Computer program] R package version 1.1-7.
- 977 Bateson, M.C., 1975. Mother-infant exchanges: The epigenesis of conver-
978 sational interaction. Annals of the New York Academy of Sciences 263,
979 101–113.
- 980 Bergelson, E., Swingley, D., 2013. The acquisition of abstract words by young
981 infants. Cognition 127, 391–397.
- 982 Bloom, K., 1988. Quality of adult vocalizations affects the quality of infant
983 vocalizations. Journal of Child Language 15, 469–480.
- 984 Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer. URL:
985 <http://www.praat.org>. [Computer program] Version 5.3.16.
- 986 Bögels, S., Magyari, L., Levinson, S.C., 2015. Neural signatures of response
987 planning occur midway through an incoming question in conversation. Sci-
988 entific Reports 5.
- 989 Bruner, J., 1985. Child's talk: Learning to use language. Child Language
990 Teaching and Therapy 1, 111–114.
- 991 Bruner, J.S., 1975. The ontogenesis of speech acts. Journal of Child Language
992 2, 1–19.
- 993 Carlson, R., Hirschberg, J., Swerts, M., 2005. Cues to upcoming swedish
994 prosodic boundaries: Subjective judgment studies and acoustic correlates.
995 Speech Communication 46, 326–333.
- 996 Casillas, M., Bobb, S.C., Clark, E.V., In press. Turn taking, timing, and
997 planning in early language acquisition. Journal of Child Language .
- 998 Casillas, M., Frank, M.C., 2012. Cues to turn boundary prediction in adults
999 and preschoolers, in: Proceedings of SemDial.
- 1000 Casillas, M., Frank, M.C., 2013. The development of predictive processes
1001 in children's discourse understanding, in: Proceedings of the 35th Annual
1002 Meeting of the Cognitive Science Society.

- 1003 Clark, E.V., 2009. First language acquisition. Cambridge University Press.
- 1004 De Ruiter, J.P., Mitterer, H., Enfield, N.J., 2006. Projecting the end of
1005 a speaker's turn: A cognitive cornerstone of conversation. *Language* 82,
1006 515–535.
- 1007 De Vos, C., Torreira, F., Levinson, S.C., 2015. Turn-timing in signed con-
1008 versations: coordinating stroke-to-stroke turn boundaries. *Frontiers in*
1009 *Psychology* 6.
- 1010 Dingemanse, M., Torreira, F., Enfield, N., 2013. Is “Huh?” a universal word?
1011 Conversational infrastructure and the convergent evolution of linguistic
1012 items. *PloS one* 8, e78273.
- 1013 Duncan, S., 1972. Some signals and rules for taking speaking turns in con-
1014 versations. *Journal of Personality and Social Psychology* 23, 283.
- 1015 Ervin-Tripp, S., 1979. Children's verbal turn-taking, in: Ochs, E., Schieffelin,
1016 B.B. (Eds.), *Developmental Pragmatics*. Academic Press, New York, pp.
1017 391–414.
- 1018 Fitneva, S., 2012. Beyond answers: questions and children's learning, in:
1019 De Ruiter, J.P. (Ed.), *Questions: Formal, Functional, and Interactional*
1020 *Perspectives*. Cambridge University Press, Cambridge, UK, pp. 165–178.
- 1021 Ford, C.E., Thompson, S.A., 1996. Interactional units in conversation: Syn-
1022 tactic, intonational, and pragmatic resources for the management of turns.
1023 *Studies in Interactional Sociolinguistics* 13, 134–184.
- 1024 Garvey, C., 1984. Children's Talk. volume 21. Harvard University Press.
- 1025 Gísladóttir, R., Chwilla, D., Levinson, S.C., 2015. Conversation electrified:
1026 ERP correlates of speech act recognition in underspecified utterances. *PloS*
1027 *one* 10, e0120068.
- 1028 Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. *Psycho-*
1029 *logical science* 11, 274–279.
- 1030 Hart, B., Risley, T.R., 1992. American parenting of language-learning chil-
1031 dren: Persisting differences in family-child interactions observed in natural
1032 home environments. *Developmental Psychology* 28, 1096.

- 1033 Hedberg, N., Sosa, J.M., Görgülü, E., Mameni, M., 2010. The prosody and
1034 meaning of Wh-questions in American English, in: Speech Prosody 2010–
1035 Fifth International Conference.
- 1036 Henning, A., Striano, T., Lieven, E.V., 2005. Maternal speech to infants at
1037 1 and 3 months of age. *Infant Behavior and Development* 28, 519–536.
- 1038 Hilbrink, E., Gattis, M., Levinson, S.C., 2015. Early developmental changes
1039 in the timing of turn-taking: A longitudinal study of mother-infant inter-
1040 action. *Frontiers in Psychology* 6.
- 1041 Hirvenkari, L., Ruusuvuori, J., Saarinen, V.M., Kivioja, M., Peräkylä, A.,
1042 Hari, R., 2013. Influence of turn-taking in a two-person conversation on
1043 the gaze of a viewer. *PLoS one* 8, e71569.
- 1044 Jaffé, J., Beebe, B., Feldstein, S., Crown, C.L., Jasnow, M.D., Rochat, P.,
1045 Stern, D.N., 2001. Rhythms of dialogue in infancy: Coordinated timing in
1046 development. *Monographs of the Society for Research in Child Develop-
1047 ment*. JSTOR.
- 1048 Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds:
1049 When speech cues count more than statistics. *Journal of Memory and
1050 Language* 44, 548–567.
- 1051 Jusczyk, P.W., 2000. *The Discovery of Spoken Language*. MIT press.
- 1052 Jusczyk, P.W., Hohne, E., Mandel, D., Strange, W., 1995. Picking up reg-
1053 ularities in the sound structure of the native language. *Speech perception*
1054 and linguistic experience: Theoretical and methodological issues in cross-
1055 language speech research , 91–119.
- 1056 Kamide, Y., Altmann, G., Haywood, S.L., 2003. The time-course of predic-
1057 tion in incremental sentence processing: Evidence from anticipatory eye
1058 movements. *Journal of Memory and Language* 49, 133–156.
- 1059 Keitel, A., Prinz, W., Friederici, A.D., Hofsten, C.v., Daum, M.M., 2013.
1060 Perception of conversations: The importance of semantics and intonation
1061 in childrens development. *Journal of Experimental Child Psychology* 116,
1062 264–277.

- 1063 Lemasson, A., Glas, L., Barbu, S., Lacroix, A., Guilloux, M., Remeuf, K.,
1064 Koda, H., 2011. Youngsters do not pay attention to conversational rules:
1065 is this so for nonhuman primates? *Nature Scientific Reports* 1.
- 1066 Levelt, W.J., 1989. Speaking: From intention to articulation. MIT press.
- 1067 Levinson, S.C., 2006. On the human “interaction engine”, in: Enfield, N.,
1068 Levinson, S. (Eds.), *Roots of human sociality: Culture, cognition and*
1069 *interaction*. Oxford: Berg, pp. 39–69.
- 1070 Levinson, S.C., 2013. Action formation and ascriptions, in: Stivers, T., Sid-
1071 nell, J. (Eds.), *The Handbook of Conversation Analysis*. Wiley-Blackwell,
1072 Malden, MA, pp. 103–130.
- 1073 Magyari, L., Bastiaansen, M.C.M., De Ruiter, J.P., Levinson, S.C., 2014.
1074 Early anticipation lies behind the speed of response in conversation. *Jour-
1075 nal of Cognitive Neuroscience* 26, 2530–2539.
- 1076 Magyari, L., De Ruiter, J.P., 2012. Prediction of turn-ends based on antici-
1077 pation of upcoming words. *Frontiers in Psychology* 3:376, 1–9.
- 1078 Masataka, N., 1993. Effects of contingent and noncontingent maternal stimu-
1079 lation on the vocal behaviour of three-to four-month-old Japanese infants.
1080 *Journal of Child Language* 20, 303–312.
- 1081 Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoni, J., Amiel-
1082 Tison, C., 1988. A precursor of language acquisition in young infants.
1083 *Cognition* 29, 143–178.
- 1084 Morgan, J.L., Saffran, J.R., 1995. Emerging integration of sequential and
1085 suprasegmental information in preverbal speech segmentation. *Child De-
1086 velopment* 66, 911–936.
- 1087 Nazzi, T., Ramus, F., 2003. Perception and acquisition of linguistic rhythm
1088 by infants. *Speech Communication* 41, 233–243.
- 1089 R Core Team, 2014. R: A Language and Environment for Statistical Com-
1090 puting. R Foundation for Statistical Computing. Vienna, Austria. URL:
1091 <http://www.R-project.org>. [Computer program] Version 3.1.1.
- 1092 Ratner, N., Bruner, J., 1978. Games, social exchange and the acquisition of
1093 language. *Journal of Child Language* 5, 391–401.

- 1094 Ross, H.S., Lollis, S.P., 1987. Communication within infant social games.
1095 *Developmental Psychology* 23, 241.
- 1096 Rossano, F., Brown, P., Levinson, S.C., 2009. Gaze, questioning and culture,
1097 in: Sidnell, J. (Ed.), *Conversation Analysis: Comparative Perspectives*.
1098 Cambridge University Press, Cambridge, pp. 187–249.
- 1099 Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the
1100 organization of turn-taking for conversation. *Language* 50, 696–735.
- 1101 Schegloff, E.A., 2007. Sequence organization in interaction: Volume 1: A
1102 primer in conversation analysis. Cambridge University Press.
- 1103 Shatz, M., 1979. How to do things by asking: Form-function pairings in
1104 mothers' questions and their relation to children's responses. *Child Develop-*
1105 *ment* 50, 1093–1099.
- 1106 Shi, R., Melancon, A., 2010. Syntactic categorization in French-learning
1107 infants. *Infancy* 15, 517–533.
- 1108 Snow, C.E., 1977. The development of conversation between mothers and
1109 babies. *Journal of Child Language* 4, 1–22.
- 1110 Soderstrom, M., Seidl, A., Kemler Nelson, D.G., Jusczyk, P.W., 2003. The
1111 prosodic bootstrapping of phrases: Evidence from prelinguistic infants.
1112 *Journal of Memory and Language* 49, 249–267.
- 1113 Speer, S.R., Ito, K., 2009. Prosody in first language acquisition—Acquiring
1114 intonation as a tool to organize information in conversation. *Language and*
1115 *Linguistics Compass* 3, 90–110.
- 1116 Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann,
1117 T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., et al., 2009.
1118 Universals and cultural variation in turn-taking in conversation. *Proceed-
1119 ings of the National Academy of Sciences* 106, 10587–10592.
- 1120 Stivers, T., Rossano, F., 2010. Mobilizing response. *Research on Language*
1121 and Social Interaction 43, 3–31.
- 1122 Takahashi, D.Y., Narayanan, D.Z., Ghazanfar, A.A., 2013. Coupled oscillator
1123 dynamics of vocal turn-taking in monkeys. *Current Biology* 23, 2162–2168.

- 1124 Thorgrímsson, G., Fawcett, C., Liszkowski, U., 2015. 1- and 2-year-olds'
 1125 expectations about third-party communicative actions. *Infant Behavior*
 1126 and *Development* 39, 53–66.
- 1127 Tice (Casillas), M., Henetz, T., 2011. Turn-boundary projection: Looking
 1128 ahead, in: *Proceedings of the 33rd Annual Meeting of the Cognitive Science*
 1129 *Society*.
- 1130 Torreira, F., Bögels, S., Levinson, S.C., 2015. Intonational phrasing is neces-
 1131 sary for turn-taking in spoken interaction. *Journal of Phonetics* 52, 46–57.
- 1132 Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006.
 1133 Elan: a professional framework for multimodality research, in: *Proceedings*
 1134 of LREC.

1135 **Appendix A. Permutation Analyses**

1136 We completed this baseline procedure on 5,000 random permutations of
 1137 the original turn transition analysis windows and compared the *t*-values from
 1138 each predictor in the original models (Table 2) to the distribution of *t*-values
 1139 for each predictor in the 5,000 models of the randomly permuted datasets.¹⁰
 1140 We could then test whether significant effects from the original statistical
 1141 models differed from the random baseline by calculating the proportion of
 1142 random data *t*-values exceeded by the original *t*-value for each predictor,
 1143 using the absolute value of all *t*-values for a two-tailed test. For example,
 1144 children's original "language condition" *t*-value was |3.429|, which is greater
 1145 than 99.9% of all |*t*-value| estimates from the randomly-permuted data mod-
 1146 els (i.e., *p*=.001). This leads us to conclude that that the effect of language
 1147 condition in the original model was highly unlikely to be the result of random
 1148 gaze shifting.

1149 E1: We excluded the output of random-permutation models that gave
 1150 convergence warnings in order to remove unreliable estimates from our anal-
 1151 yses (non-converging models were 22.4% and 24.4% of all models for children

¹⁰We report *t*-values rather than beta estimates because the standard errors in the ran-
 domly 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 Supple-
 mentary Materials.

1152 and adults, respectively; see the Supplementary Materials for more information
 1153 on model exclusion).

1154 E2: As before, we excluded the output of random-permutation models
 1155 that resulted in convergence warnings in order to remove unreliable model
 1156 estimates from our analyses (non-converging models made up 69% and 70% of
 1157 models for children and adults, respectively; see the Supplementary Materials
 1158 for more information on model exclusion).

1159 In all of the following plots, the gray dots represent the randomly per-
 1160 muted data's model estimates for the value listed (beta or standard error),
 1161 the white dots represent the model estimates from the original data, and the
 1162 triangles represent the 95th percentile for each distribution being shown.

1163 *Appendix A.1. Experiment 1*

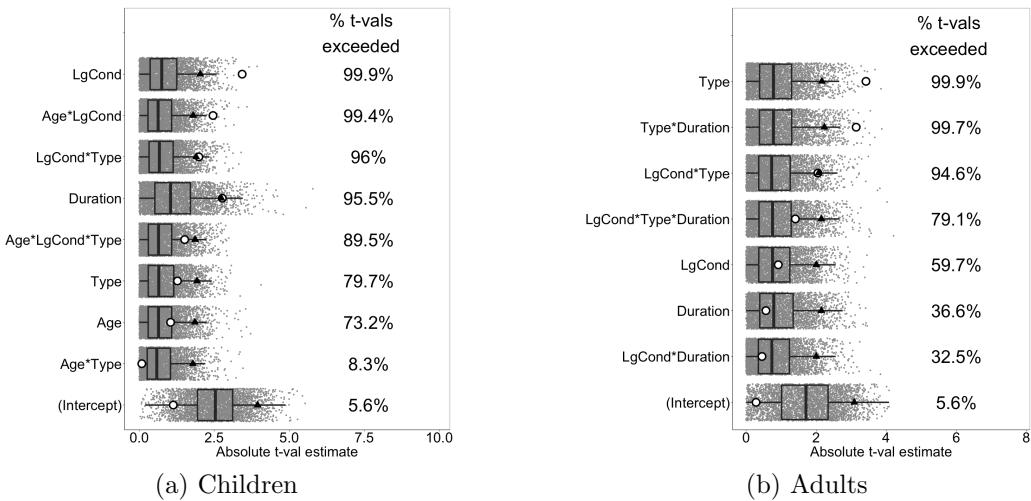


Figure A.1: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.

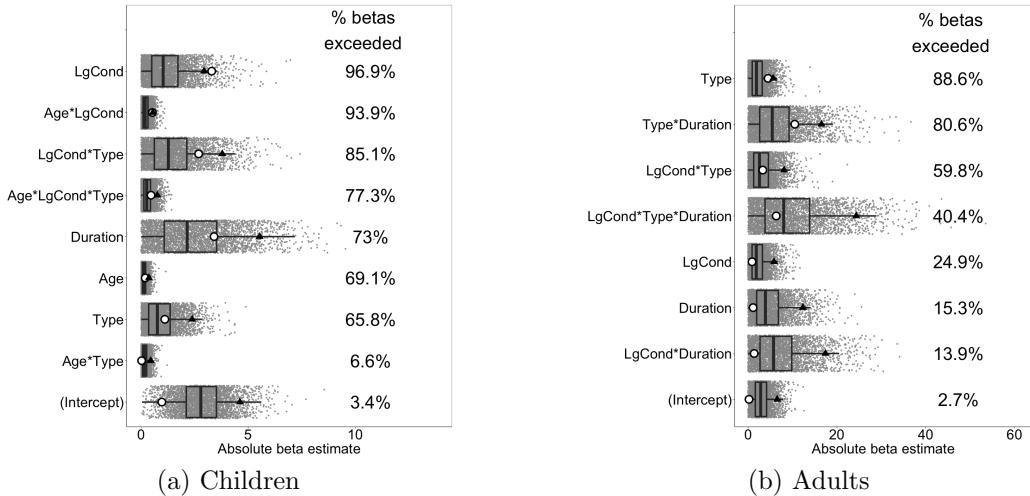


Figure A.2: Random-permutation and original $|\beta\text{-values}|$ for predictors of gaze rates in Experiment 1.

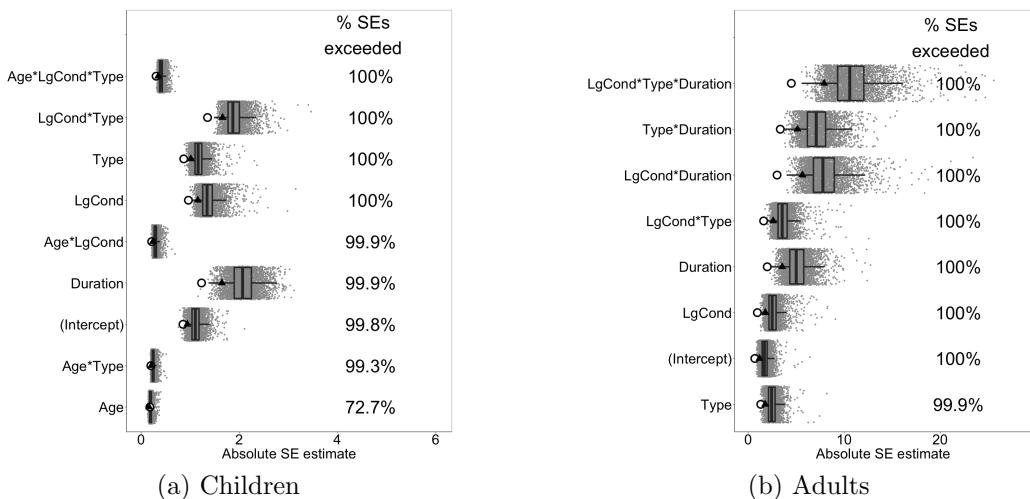


Figure A.3: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 1.

1164 *Appendix A.2. Experiment 2*

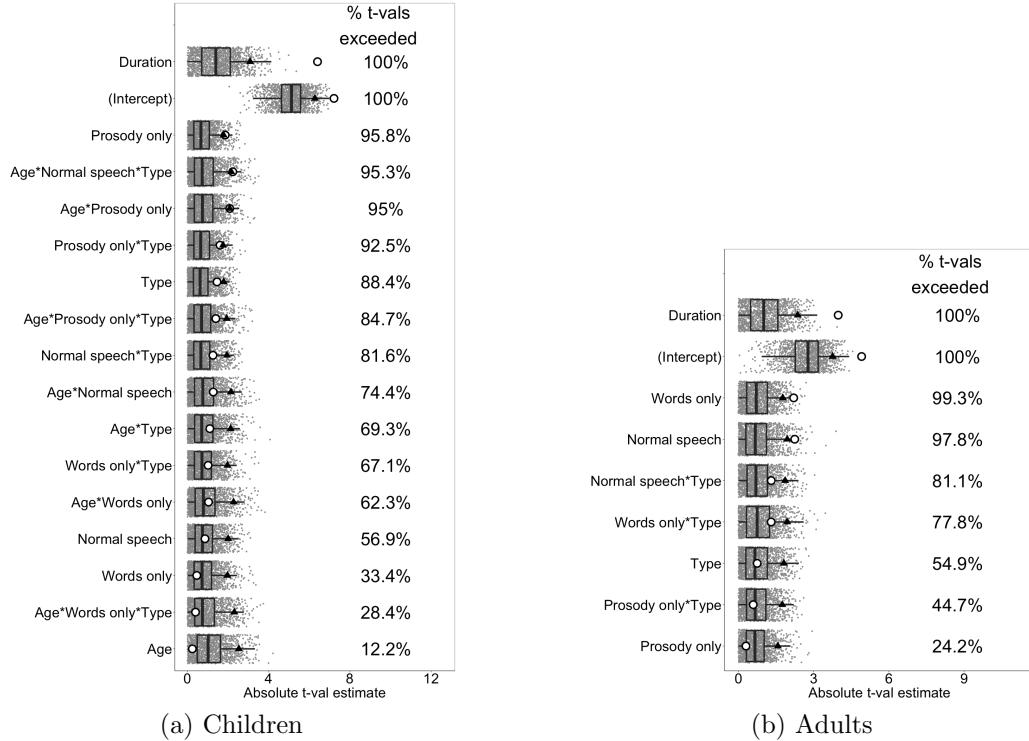


Figure A.4: Random-permutation and original $|t\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

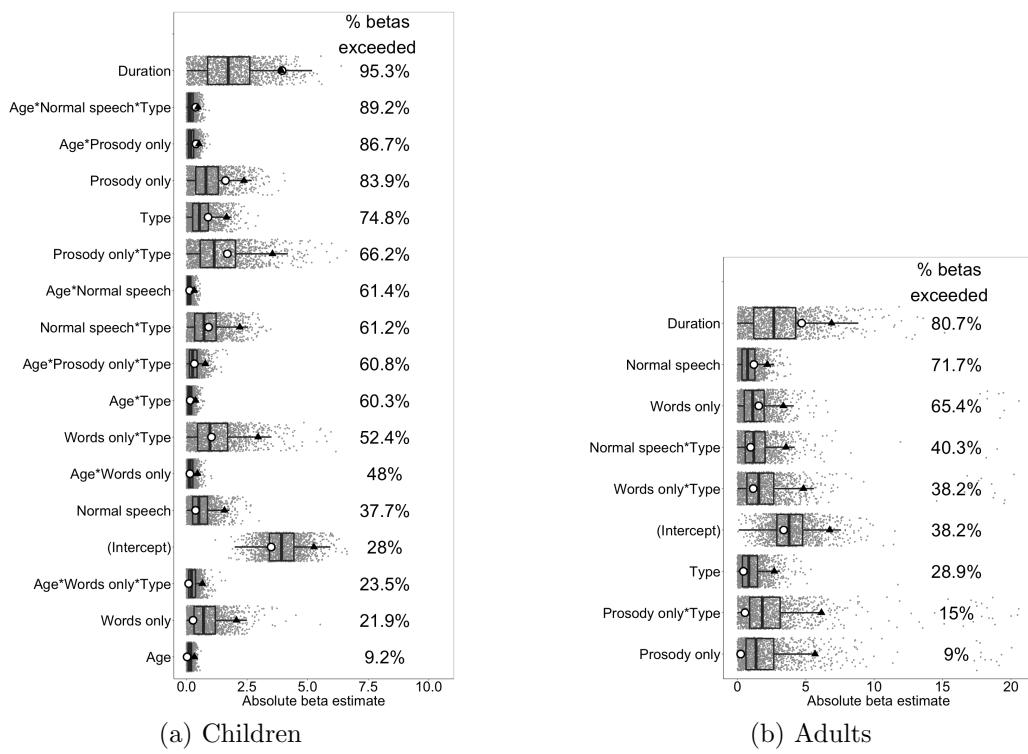


Figure A.5: Random-permutation and original $|\beta\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

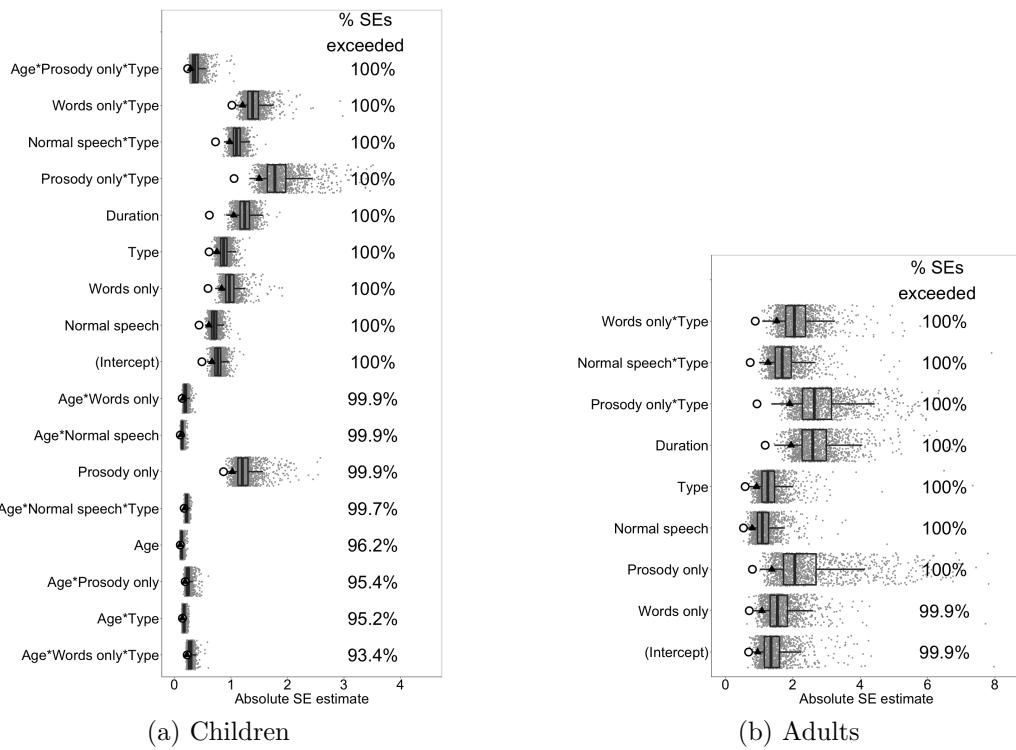


Figure A.6: Random-permutation and original $|SE\text{-values}|$ for predictors of anticipatory gaze rates in Experiment 2.

1165 **Appendix B. Pairwise developmental tests: Real vs. randomly
1166 permuted effects**

1167 In each of the plots below, the dot represents the original data value for
1168 the effect and the 5,000 randomly permuted data effect sizes are shown in the
1169 distribution. The percentage shown is the percentage of random permutation
1170 values exceeded by the original data value (taking the absolute value of all
1171 data points for a two-tailed test.)

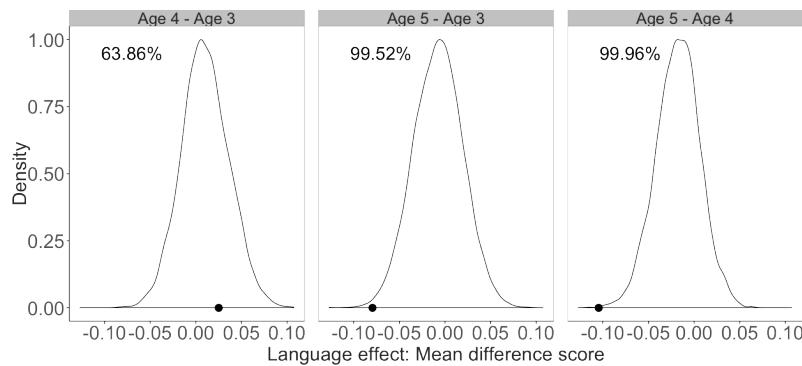


Figure B.1: Pairwise comparisons of the language condition effect across ages in Experiment 1.

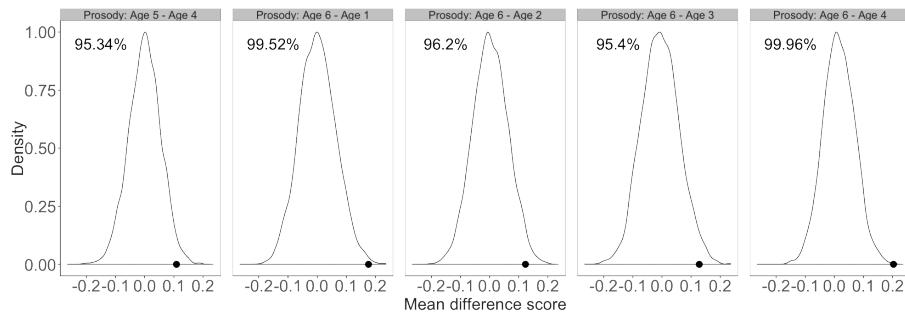


Figure B.2: Significant pairwise comparisons of the *prosody only-no speech* linguistic condition effect, across ages in Experiment 2

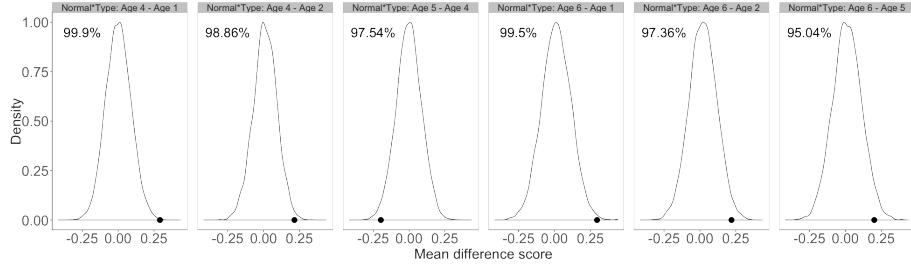


Figure B.3: Significant pairwise comparisons of the *normal speech-no speech* language condition effect for transition type, across ages, in Experiment 2.

1172 **Appendix C. Non-convergent models**

1173 Non-convergent models made up 22–24% of the 5,000 models of randomly
1174 permuted data in Experiment 1 and 69–70% of the 5,000 for Experiment 2.
1175 We excluded these non-convergent models because they displayed erratic
1176 β and SE estimates, as summarized below in the table of t -values from
1177 convergent and non-convergent models in Experiment 1. The non-convergent
1178 models from Experiment 2 showed similar patterns. The high frequency
1179 of problematic models persisted even when we changed optimizers and we
1180 suspect the issue derives from data sparsity in some of the random runs.

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

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

1181 **Appendix D. Miscellaneous**

1182 One alternative hypothesis for children’s anticipatory gazes is that they
1183 simply grow bored and start looking away at a constant rate after a turn
1184 begins. This data plotted here show a hypothetical group of participants
1185 who begin to lose interest (at a linear rate) after one second of a turn (gray
1186 dots) compared to participants’ real data from Experiment 2 (black dots).
1187 This pattern suggests that, though children do look away with time, their
1188 looks away are not simply driven by boredom.

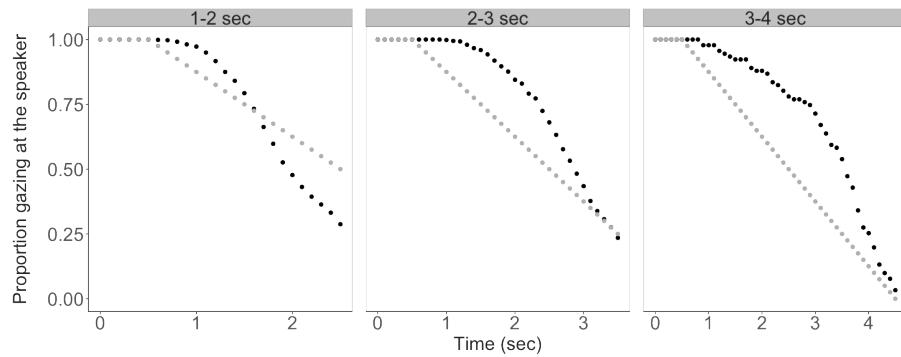


Figure D.1: Proportion participants looking at the current speaker: hypothetical boredom-driven data (gray dots) versus real data from Experiment 2 (black dots).

1189 The design for Experiment 2 does not fully cross puppet pair (e.g., robots,
1190 blue puppets) with linguistic condition (e.g., “words only” and “no speech”).
1191 Even though each puppet pair is associated with different conversation clips
1192 across children (e.g. robots talking about kitties, birthday parties, or pan-
1193 cakes), robots were only associated with “words only” speech, merpeople
1194 were only associated with “prosody only” speech, and the puppets with fancy
1195 clothes were only associated with the “no speech” condition. We did this to
1196 increase the pragmatic felicity of the experiments for the older children (i.e.,
1197 robots make robot sounds, merpeople’s voices are muffled under water, the
1198 fancy-clothed puppets are in a room with main other voices). It is therefore
1199 fair to point out a possible confound between linguistic condition and puppet
1200 pair. Thankfully, we also ran a short follow-up study at the museum with

1201 3–5-year-olds in which each child only saw one video—the normal speech
 1202 conversation about birthday parties—with a randomly assigned puppet pair
 1203 performing the conversation. Five children watched each puppet pair, for a
 1204 total of 30 children across the six pairs. This experiment holds all things
 1205 constant except for the appearance of the puppets. We then used a mixed
 1206 effects logistic regression of children’s anticipatory switches (yes or no at each
 1207 transition), with puppet pair (robots/merpeople/fancy dress/other-3; Figure
 1208 1) as a fixed effect and participant and turn transition as random effects.
 1209 In four versions of this model we systematically varied the reference level to
 1210 check for differences between every puppet pair, finding no significant affects
 1211 of puppet type on switching rate. We take this as evidence that, although
 1212 we did not fully cross puppet pairs and linguistic conditions in Experiment
 1213 2, it was unlikely to have had strong effects on children’s looking rates above
 1214 and beyond the intended effects of linguistic condition.

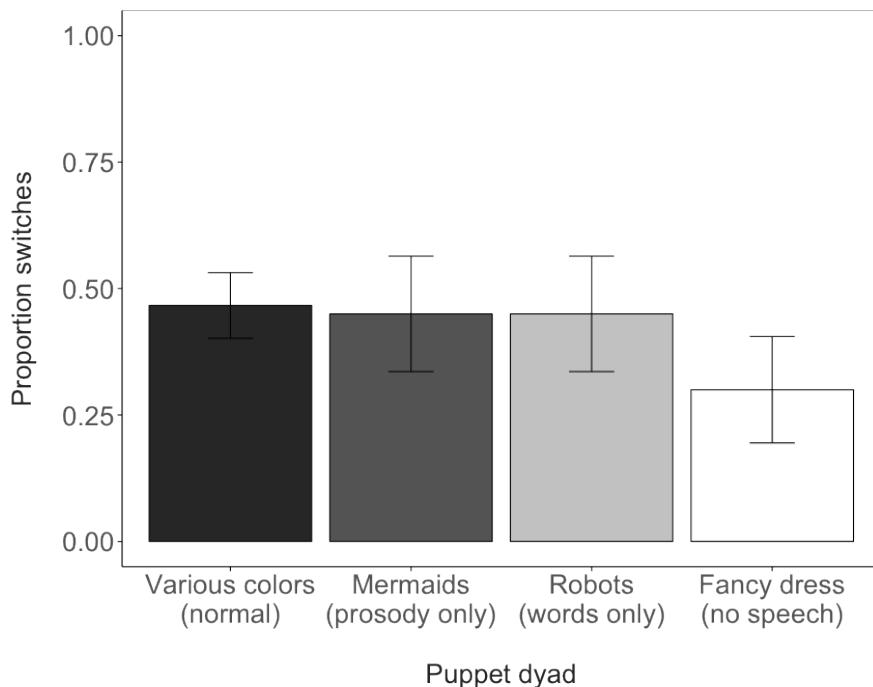


Figure D.2: Proportion gaze switches across puppet pairs when linguistic condition and conversation are held constant.