

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 fluidly integrate their growing knowledge of language into their turn-taking behavior. In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling linguistic cues to turn structure. In Experiment 1, we showed English and non-English conversations to English-speaking adults and children. In Experiment 2, we phonetically controlled lexicosyntactic and prosodic cues in English-only speech. Children spontaneously made anticipatory gaze switches by age two and continued improving through age six. In both experiments, children and adults made more anticipatory switches after hearing questions. Consistent

*Corresponding author.

Address: Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

Email: marisa.casillas@mpi.nl

Telephone: +31 024 3521 566; Fax: +31 024 3521 213

with prior findings on adult turn prediction, prosodic information alone did not increase children’s anticipatory gaze shifts. But, unlike prior work with adults, lexical information alone was not sufficient either—children’s performance was best overall with lexicosyntax and prosody together. Our findings support an account in which turn tracking and turn prediction emerge in infancy and then gradually become integrated with children’s online linguistic processing.

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

1 Introduction

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

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

22 In this study, we investigate when children begin to make predictions
23 about upcoming turn structure in conversation and how online linguistic
24 processing becomes integrated into their predictions as they grow older. We
25 first give a basic review of turn-taking research and the state of current
26 knowledge about adult turn prediction. We then discuss recent work on the
27 development of turn-taking skills before presenting the details of the present
28 study.

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

40 Conversational turn taking distinguishes itself from other turn-taking be-
41 haviors by the complexity of the sequencing involved. Conversational turns
42 come grouped into semantically-contingent sequences of action. The groups
43 can span turn-by-turn exchanges (e.g., simple question–response, “How are
44 you?”–“Fine.”) or sequence-by-sequence exchanges (e.g., reciprocals, “How
45 are you?”–“Fine, and you?”–“Great!”). Compared to other turn-taking be-
46 haviors, the possible sequence and action types in everyday talk are diverse
47 and unpredictable.

48 Despite this complexity, conversational turn taking is precise in its timing.
49 Across a diverse sample of conversations in 10 languages, one study found
50 a consistent average inter-turn silence of 0–200 msec at points of speaker
51 switch (Stivers et al., 2009). Experimental results and current models of
52 speech production suggest that it takes approximately 600 msec to produce
53 a content word, and even longer to produce a simple utterance (Griffin &
54 Bock, 2000; Levelt, 1989). In order to achieve 200 msec turn transitions,
55 speakers must begin formulating their response before the prior turn has
56 ended (Levinson, 2013, 2016). Moreover, to formulate their response early
57 on, speakers must track and anticipate what types of response might become
58 relevant next. They also need to predict the content and form of upcoming
59 speech so that they can launch their articulation at exactly the right moment.
60 Prediction thus plays a key role in timely turn taking.

61 Adults have a lot of information at their disposal to help make accurate
62 predictions. Lexical, syntactic, and prosodic information (e.g., *wh*-words,
63 subject-auxiliary inversion, and list intonation) can all inform addressees
64 about upcoming linguistic structure (De Ruiter et al., 2006; Duncan, 1972;
65 Ford & Thompson, 1996; Bögels & Torreira, 2015). Non-verbal cues (e.g.,
66 gaze, posture, and pointing) often appear at turn-boundaries and can some-

67 times act as late indicators of an upcoming speaker switch (Rossano et al.,
68 2009; Stivers & Rossano, 2010). Additionally, the sequential context of a
69 turn can make the next action obvious: answers after questions, thanks or
70 denial after compliments, etc. (Schegloff, 2007).

71 Prior work suggests that adult listeners primarily use lexicosyntactic in-
72 formation to accurately predict upcoming turn structure. De Ruiter and
73 colleagues (2006) asked participants to listen to snippets of spontaneous con-
74 versation and to press a button whenever they anticipated that the current
75 speaker was about to finish his or her turn. The speech snippets were con-
76 trolled for the amount of linguistic information present; some were normal,
77 but others had flattened pitch, low-pass filtered speech, or further manip-
78 ulations. With pitch-flattened speech, the timing of participants' button
79 responses was comparable to their timing with the full linguistic signal. But
80 when no lexical information was available, participants responded signifi-
81 cantly earlier within the turn. The authors concluded that lexicosyntactic
82 information¹ was necessary and possibly sufficient for turn-end projection,

¹The “lexicosyntactic” condition only included flattened pitch and so was not exclu-
sively lexicosyntactic—the speech would still have residual prosodic structure, including
syllable duration and intensity.

83 while intonation was neither necessary nor sufficient. Congruent evidence
84 comes from studies varying the predictability of lexicosyntactic and prag-
85 matic content: adults anticipate turn ends better when they can more accu-
86 rately predict the exact words that will come next (Magyari & De Ruiter,
87 2012; see also Magyari et al., 2014). They can also identify speech acts within
88 the first word of an utterance (Gísladóttir et al., 2015), allowing them to start
89 planning their response at the first moment possible (Bögels et al., 2015).

90 Despite this body of evidence, the role of prosody for adult turn pre-
91 diction is still a matter of debate. De Ruiter and colleagues’ (2006) ex-
92 periment focused on the role of intonation, which is only a partial index
93 of prosody. Prosody is tied closely to the syntax of an utterance, so the
94 two linguistic signals are difficult to control independently (Ford & Thomp-
95 son, 1996). Bögels & Torreira (2015) used a combination of button-press
96 and verbal responses to investigate the relationship between lexicosyntac-
97 tic and prosodic cues in turn-end prediction. Critically, their stimuli were
98 cross-spliced so that each item had full prosodic cues to accompany the lex-
99 icosyntax. Because of the splicing, they were able to create items that had
100 syntactically-complete units with no intonational phrase boundary at the
101 end. Participants never verbally responded or pressed the “turn-end” button

102 when hearing a syntactically-complete phrase without an intonational phrase
103 boundary. And when intonational phrase boundaries were embedded within
104 multi-utterance turns, participants were tricked into pressing the “turn-end”
105 button 29% of the time. These findings suggest that listeners actually do
106 rely on prosodic cues to execute a response, and that their use of prosodic
107 cues interacts with their predictions about the unfolding syntactic structure
108 (see also de Ruiter et al. (2006):525). These experimental findings cor-
109 roborate other corpus and experimental work promoting a combination of
110 cues (lexicosyntactic, prosodic, and pragmatic) as key for accurate turn-end
111 prediction (Duncan, 1972; Ford & Thompson, 1996; Hirvenkari et al., 2013).

112 *Turn taking in development*

113 The majority of work on children’s early turn taking has focused on ob-
114 servations of spontaneous interaction. Children’s first turn-like structures
115 appear as early as two to three months after birth, in proto-conversation with
116 their caregivers (Bruner, 1975, 1985; Snow, 1977). During proto-conversations,
117 caregivers treat their infants as capable of making meaningful contributions:
118 they take every look, vocalization, arm flail, and burp as “utterances” in the
119 joint discourse (Bateson, 1975; Jaffe et al., 2001; Snow, 1977). Infants catch
120 onto the structure of proto-conversations quickly. By three to four months

121 they notice disturbances to the contingency of their caregivers' response and,
122 in reaction, change the rate and quality of their vocalizations (Bloom, 1988;
123 Masataka, 1993; Toda & Fogel, 1993).

124 The timing of children's responses to their caregivers' speech shows a
125 non-linear pattern. Infants' contingent vocalizations in the first few months
126 of life show very fast timing (though with a lot of vocal overlap). But by
127 nine months, their timing slows down considerably, only to gradually speed
128 up again after 12 months (Hilbrink et al., 2015). For children, taking turns
129 with brief transitions between speakers is more difficult than avoiding speaker
130 overlap; children's incidence of overlap is nearly adult-like by nine months,
131 but the timing of their non-overlapped (i.e., gapped) responses remains longer
132 than the adult 200 msec standard for the next few years (Casillas et al., 2016;
133 Garvey, 1984; Garvey & Berninger, 1981; Ervin-Tripp, 1979). This puzzling
134 pattern is likely due to children's linguistic development: taking turns on
135 time is easier when their response is a simple vocalization rather than a
136 linguistic utterance. Integrating language into the turn-taking system may
137 therefore be a major factor in children's delayed responses (Casillas et al.,
138 2016).

139 Before children manage to fully integrate linguistic processing into their

140 turn-taking behaviors (for both turn prediction and production), they can
141 rely on non-verbal interactional cues, including silence, eye gaze, body orien-
142 tation, and gesture, to identify the boundaries of social actions. For example,
143 with little to no linguistic knowledge, children are often able to infer desired
144 responses to offers and requests by taking account of their interlocutor's
145 non-verbal communicative behavior, the structure of routine events, and the
146 affordances of the current interactional context (Reddy et al., 2013; Nomikou
147 & Rohlfsing, 2011; Shatz, 1978). With respect to turn taking in particular,
148 children's spontaneous vocalizations during interaction demonstrate a sensi-
149 tivity to short inter-speaker gaps from infancy (Hilbrink et al., 2015). Thus,
150 before children can anticipate turn structure by integrating linguistic cues
151 from unfolding speech, they might react to silence as a cue to upcoming
152 speaker change. Interactional silence itself may then serve as one of chil-
153 dren's first cues to turn structure, giving them information about when to
154 respond before they can rely on language.

155 As children's language competence and speed of processing increases
156 (Kail, 1991), they become better equipped to use linguistic cues in making
157 predictions about upcoming turn structure. Studies of early linguistic devel-
158 opment point to a possible early advantage for prosody over lexicosyntax in

159 children’s turn-taking predictions. Infants can distinguish their native lan-
160 guage’s rhythm type from others soon after birth (Mehler et al., 1988; Nazzi
161 & Ramus, 2003). They also show preference for the typical stress patterns of
162 their native language over others by 6–9 months (e.g., iambic vs. trochaic),
163 and can use prosodic information to segment the speech stream into smaller
164 chunks from 8 months onward (Johnson & Jusczyk, 2001; Morgan & Saffran,
165 1995). Four- to five-month-olds also prefer pauses in speech to be inserted at
166 prosodic boundaries, and by 6 months infants can use prosodic markers to
167 pick out sub-clausal syntactic units, both of which are useful for extracting
168 turn structure from ongoing speech (Jusczyk et al., 1995; Soderstrom et al.,
169 2003). In comparison, children show at best a very limited lexical inventory
170 before their first birthday (Bergelson & Swingley, 2013; Shi & Melancon,
171 2010).

172 Keitel and colleagues (2013) were one of the first to explore how children
173 use linguistic cues to predict upcoming turn structure. They asked 6-, 12-,
174 24-, and 36-month-old infants, and adult participants to watch short videos
175 of conversation and tracked their eye movements at points of speaker change.
176 They showed their participants two types of videos—one normal and one with
177 flattened pitch—to test the role of intonation in participants’ anticipatory

178 predictions about upcoming speech. Comparing children’s anticipatory gaze
179 frequency to a random baseline, they found that only 36-month-olds and
180 adults made anticipatory gaze switches more often than expected by chance,
181 and that only 36-month-olds were affected by flattened intonation contours.
182 This finding led Keitel and colleagues to conclude that children’s ability to
183 predict upcoming turn structure relies on their ability to comprehend the
184 stimuli lexicosemantically. They also suggest that intonation might play
185 a secondary role in turn prediction, but only after children acquire more
186 sophisticated, adult-like language comprehension skills (also see Keitel &
187 Daum, 2015).

188 Although the Keitel et al. (2013) study constitutes a substantial ad-
189 vance over previous work in this domain, it has some limitations. Because
190 these limitations directly inform our own study design, we review them in
191 some detail. First, their estimates of baseline gaze frequency (“random” in
192 their terminology) were not random. Instead, they used gaze switches dur-
193 ing ongoing speech as a baseline. But ongoing speech is the period in which
194 switching is least likely to occur (Hirvenkari et al., 2013)—their baseline thus
195 maximizes the chance of finding a difference in gaze frequency at turn transi-
196 tions compared to the baseline. A more conservative baseline would compare

197 participants' looking behavior at turn transitions to their looking behavior
198 during randomly selected windows of time throughout the stimulus, includ-
199 ing turn transitions. We follow this conservative approach in the current
200 study.

201 Second, the conversation stimuli Keitel et al. (2013) used were some-
202 what unusual. The average gap between turns was 900 msec, a duration
203 much longer than typical adult timing, which averages around 200 msec
204 (Stivers et al., 2009). The speakers in the videos were also asked to mini-
205 mize their movements while performing scripted, adult-directed conversation,
206 which would have created a somewhat unnatural interaction. Additionally,
207 to produce more naturalistic conversation, it would have been ideal to local-
208 ize the sound sources for the two voices in the video (i.e., to have the voices
209 come out of separate left and right speakers). But both voices were recorded
210 and played back on the same audio channel, which may have made it difficult
211 to distinguish the two talkers. Again, we attempt to address these issues in
212 our current study. Despite these minor methodological drawbacks, the Kei-
213 tel et al. (2013) study still demonstrates interesting age-based differences
214 in children's predictions about upcoming turn structure. Our current work

215 takes these findings as a starting point.²

216 *The current study*

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

228 We highlight four primary findings: first, although children and adults
229 use linguistic cues to make predictions about upcoming turn structure, they
230 do so primarily to predict speaker transitions after questions (a “speech act”
231 effect). This intriguing effect, which has not been reported previously, sug-

²But also see Casillas & Frank (2012, 2013).

gests that participants track unfolding speech for cues to upcoming speaker change, which may affect how they use linguistic cues more generally for anticipatory processing in conversation. Second, we find that children make more predictions than expected by chance starting at age two, but that this effect is small at first, and continues to improve through age six, along with children's use of linguistic cues to anticipate answers after question turns. Third, children and adults often used inter-turn silence (a non-verbal cue to turn structure) to make more predictive gaze switches to the responder, suggesting that non-verbal cues are useful for predicting turn structure early on and continue to be important in adulthood. Finally, we find no evidence for an early prosodic advantage in children's anticipations and, further, no evidence that lexical cues alone are comparable to the full linguistic signal in aiding children's predictions (as is proposed for adults; De Ruiter et al., 2006). Anticipation is strongest for stimuli with the full range of linguistic cues. Our findings support an account in which turn prediction emerges in infancy and becomes integrated with online linguistic processing gradually, possibly because of children's increased linguistic knowledge and speed of processing with development.

250 Experiment 1

251 We recorded participants’ eye movements as they watched six short videos
252 of two-person (dyadic) conversation that were interspersed with attention-
253 getting filler videos. Each conversation video featured an improvised dis-
254 course in one of five languages (English, German, Hebrew, Japanese, and
255 Korean). Participants saw two videos in English and one in every other lan-
256 guage. The participants, all native English speakers, were only expected to
257 understand the two videos in English. We showed participants non-English
258 videos to limit their access to lexical information while maintaining their
259 access to other cues to turn boundaries (e.g., non-English prosody, gaze, in-
260 breaths, phrase final lengthening). Using this method, we analyzed children
261 and adult’s anticipatory looks from the current speaker to the upcoming
262 speaker at points of turn transition in English and non-English videos.

263 *Methods*

264 *Participants*

265 We recruited 74 children ages 3;0–5;11 and 11 undergraduate adults to
266 participate in the experiment. We recruited adult participants through the
267 Stanford University Psychology participant database. Adult participants

268 were either paid or received course credit for their time. Our child sample in-
269 cluded 19 three-year-olds, 32 four-year-olds, and 23 five-year-olds, all enrolled
270 in a local nursery school and all of whom volunteered their time. All par-
271 ticipants were native English speakers. Approximately one-third (N=25) of
272 the children’s parents and teachers reported that their child regularly heard
273 a second (and sometimes third or further) language, but only one child fre-
274 quently heard a language that was used in our non-English video stimuli,
275 and we excluded his data from the analyses.³ None of the adult participants
276 reported fluency in a second language.

277 *Materials*

278 *Video recordings.* We recorded pairs of talkers while they conversed in
279 a sound-attenuated booth (Figure 1). Each talker was a native speaker of
280 the language being recorded, and each talker pair was male-female. Using
281 a Marantz PMD 660 solid state field recorder, we captured audio from two

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



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

282 lapel microphones, one attached to each participant, while simultaneously
283 recording video from the built-in camera of a MacBook laptop computer.
284 The talkers were volunteers and were acquainted with their recording partner
285 ahead of time.

286 Each recording session began with a 20-minute warm-up period of spon-
287 taneous conversation during which the pair talked for five minutes on four
288 topics (favorite foods, entertainment, hometown layout, and pets). Then we
289 asked talkers to choose a new topic—one relevant to young children (e.g.,
290 riding a bike, eating breakfast)—and to improvise a dialogue on that topic.
291 We asked them to speak as if they were on a children’s television show in
292 order to elicit child-friendly speech toward each other. We recorded until the
293 talkers achieved at least 30 seconds of uninterrupted discourse with enthu-

294 siastic, child-friendly speech. Most talker pairs took less than five minutes
295 to complete the task, usually by agreeing on a rough script at the start. We
296 encouraged talkers to ask at least a few questions to each other during the
297 improvisation. The resulting conversations were therefore not entirely spon-
298 taneous, but were as close as possible while still remaining child-oriented in
299 topic, prosodic pattern, and lexicosyntactic construction.⁴

300 After recording, we combined the audio and video recordings by hand,
301 and cropped each one to the (approximate) 30-second interval with the most
302 turn activity. Because we recorded the conversations in stereo, the male and
303 female voices came out of separate speakers during video playback. This gave
304 each voice in the videos a localized source (from the left or right loudspeaker).
305 We coded each turn transition in the videos for language condition (English
306 vs. non-English), inter-turn gap duration (in milliseconds), and transition
307 type (question vs. non-question). Each non-English turn was coded as a
308 question or non-question from a monolingual English-speaker’s perspective,
309 i.e., turns that “sound like” questions and turns that do not. We asked

⁴All 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.

310 five native American English speakers to listen to the audio recording for
311 each non-English turn and judge whether it sounded like a question. We
312 marked non-English turns as questions when at least 4 of the 5 listeners
313 (80%) said that the turn “sounded like a question”. Thus, “question” cues
314 in the non-English condition only *resembled* native English question cues,
315 and were therefore likely harder to identify than cues to questionhood in
316 the English condition. However, since participants did not speak the non-
317 English languages and would only ever treat “question-sounding” turns as
318 questions, we proceeded with these analyses to see how pervasive question
319 effects were—could they show up even without lexical access? If participants
320 primarily rely on prosodic cues to question turns, it’s possible that even non-
321 English prosody can elicit anticipatory gaze switches for question-like turns.

322 Because the conversational stimuli were recorded semi-spontaneously, the
323 duration of turn transitions and the number of speaker transitions in each
324 video was variable. We measured the duration of each turn transition from
325 the audio recording associated with each video. We excluded turn transitions
326 longer than 550 msec and shorter than 90 msec from analysis, additionally
327 excluding overlapped transitions.⁵ This left approximately equal numbers

⁵Overlap occurs when a responder begins a new turn before the current turn is finished.

328 of turn transitions available for analysis in the English (N=20) and non-
329 English (N=16) videos. On average, the inter-turn gaps for English videos
330 (mean=318, median=302, stdev=112 msec) were slightly longer than for non-
331 English videos (mean=286, median=251, stdev=122 msec).

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

339 *Procedure*

340 Participants sat in front of an SMI 120Hz corneal reflection eye-tracker
341 mounted beneath a large flatscreen display. The display and eye-tracker were

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). They would therefore still be fixated on the prior speaker when the overlap started, and would have to switch their gaze *reactively* to the responder.

342 secured to a table with an ergonomic arm that allowed the experimenter to
343 position the whole apparatus at a comfortable height and approximately 60
344 cm from the viewer. We placed stereo speakers on the table, to the left and
345 right of the display.

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

356 Children were less inclined to simply sit and watch videos of conversation
357 in languages they didn't speak, so we used a different procedure to keep them
358 engaged: the experimenter started each session by asking the child about
359 what languages he or she could speak, and about what other languages he
360 or she had heard of. Then the experimenter expressed her own enthusiasm

361 for learning about new languages, and invited the child to watch a video
362 about “new and different languages” together. If the child agreed to watch,
363 the experimenter and the child sat together in front of the display, with
364 the child centered in front of the tracker and the experimenter off to the
365 side. Each conversation video was preceded and followed by a 15–30 second
366 attention-getting filler video (e.g., running puppies, singing muppets, flying
367 bugs). If the child began to look bored, the experimenter would talk during
368 the fillers, either commenting on the previous conversation (“That was a neat
369 language!”) or giving the language name for the next conversation (“This
370 next one is called Hebrew. Let’s see what it’s like.”) The experimenter’s
371 comments reinforced the video-watching as a joint task.

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

380 *Data preparation and coding*

381 To determine whether participants predicted upcoming turn transitions,
382 we needed to define a set of criteria for what counted as an anticipatory gaze
383 shift. Prior work using similar experimental procedures has found that adults
384 and children make anticipatory gaze shifts to upcoming talkers within a wide
385 time frame; the earliest shifts occur before the end of the prior turn, and the
386 latest occur after the onset of the response turn, with most shifts occurring
387 in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz,
388 2011). Following prior work, we measured how often our participants shifted
389 their gaze from the prior to the upcoming speaker *before* the shift in gaze
390 could have been initiated in reaction to the onset of the speaker’s response.
391 In doing so, we assumed that it takes participants 200 msec to plan an eye
392 movement, following standards from adult anticipatory processing studies
393 (e.g., Kamide et al., 2003).

394 We checked each participant’s gaze at each turn transition for three char-
395 acteristics (Figure 2): (1) that the participant fixated on the prior speaker
396 for at least 100 msec at the end of the prior turn, (2) that immediately
397 thereafter the participant switched to fixate on the upcoming speaker for at
398 least 100 ms, and (3) that the switch in gaze was initiated within the first

399 200 msec of the response turn, or earlier. These criteria guarantee that we
400 only counted gaze shifts when: (1) participants were tracking the previous
401 speaker, (2) switched their gaze to track the upcoming speaker, and (3) did
402 so before they could have simply reacted to the onset of speech in the re-
403 sponse. Under the assumption that it takes at least 200 msec to plan an eye
404 movement, gaze shifts initiated within the first 200 msec of the response (or
405 earlier) were planned *before* participants could react to the onset of speech
406 itself.

407 As mentioned, most anticipatory switches happen in the inter-turn gap,
408 but we also allowed anticipatory gaze switches that occurred in the final syl-
409 lables of the prior turn. Early switches are consistent with the distribution
410 of responses in explicit turn-boundary prediction tasks. For example, in a
411 button press task, adult participants anticipated turn ends approximately
412 200 msec in advance of the turn's end, and anticipatory responses to pitch-
413 flattened stimuli came even earlier (De Ruiter et al., 2006). We therefore
414 allowed switches to occur as early as 200 msec before the end of the prior
415 turn. Again, because it takes 200 msec to plan an eye movement, we counted
416 anticipatory switches, at the latest, 200 msec after the onset of speech. There-
417 fore, for very early and very late switches, our requirement of 100 msec of

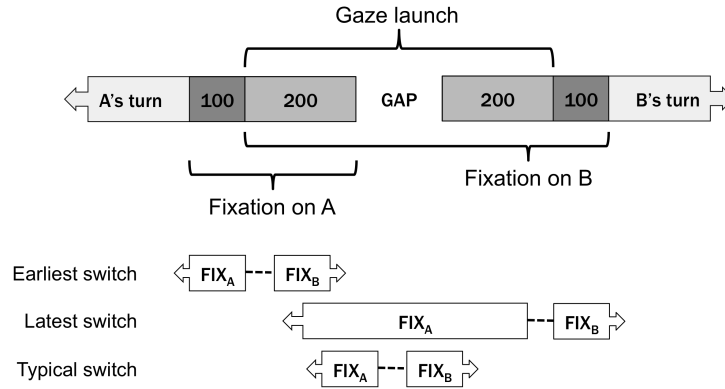


Figure 2: Schematic summary of the criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition. FIX = hypothetical fixation on speaker A or speaker B; dashed lines = hypothetical saccadic time.

418 fixation on each speaker would sometimes extend outside of the gaze launch
 419 window boundaries (200 msec before and after the inter-turn gap; dark gray
 420 boxes Figure 2). The maximally available fixation window was therefore 100
 421 msec before and after the earliest and latest possible switch point (300 msec
 422 before and after the inter-turn gap). We did not count switches made during
 423 the fixation window as anticipatory. We *did* count switches made during the
 424 inter-turn gap. The period of time from the beginning of the possible fixation
 425 window on the prior speaker to the end of the possible fixation window on
 426 the responder was our total analysis window (300 msec + the inter-turn gap
 427 + 300 msec).

428 *Predictions.* We expected participants to show greater anticipation in the
429 English videos than in the non-English videos because of their increased
430 access to linguistic information in English. We also predicted that anticipa-
431 tion would be greater following questions compared to non-questions; ques-
432 tions have early cues to upcoming turn transition (e.g., *wh*-words, subject-
433 auxiliary inversion) and also make a next response immediately relevant.
434 Our third prediction was that anticipatory looks would increase with devel-
435 opment, along with children’s increased linguistic competence and speed of
436 processing. Finally, we predicted that transitions with longer inter-turn gaps
437 would show greater anticipation because longer gaps provide (a) more time
438 to make a gaze switch and (b) are themselves a cue to possible upcoming
439 speaker switch.

440 *Results*

441 Participants looked at the screen most of the time during video playback
442 (81% and 91% on average for children and adults, respectively). They pri-
443 marily kept their eyes on the person who was currently speaking in both
444 English and non-English videos: they gazed at the current speaker between
445 38% and 63% of the time, looking back at the addressee between 15% and
446 20% of the time (Table 1). Even three-year-olds looked more at the current

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 across ages in Experiment 1.

447 speaker than anything else, whether or not the videos were in a language they
 448 could understand. Children looked at the current speaker less than adults
 449 did during the non-English videos. Despite this, their looks to the addressee
 450 did not increase substantially in the non-English videos, indicating that their
 451 looks away were probably related to boredom rather than confusion about
 452 ongoing turn structure. Overall, participants' pattern of gaze to current
 453 speakers demonstrated that they performed basic turn tracking during the
 454 videos, regardless of language. Figure 3 shows participants' anticipatory gaze
 455 rates across age, language condition, and transition type.

456 *Statistical models*

457 We identified anticipatory gaze switches for all 36 usable turn transitions,
 458 based on the criteria outlined above, and analyzed them for effects of lan-

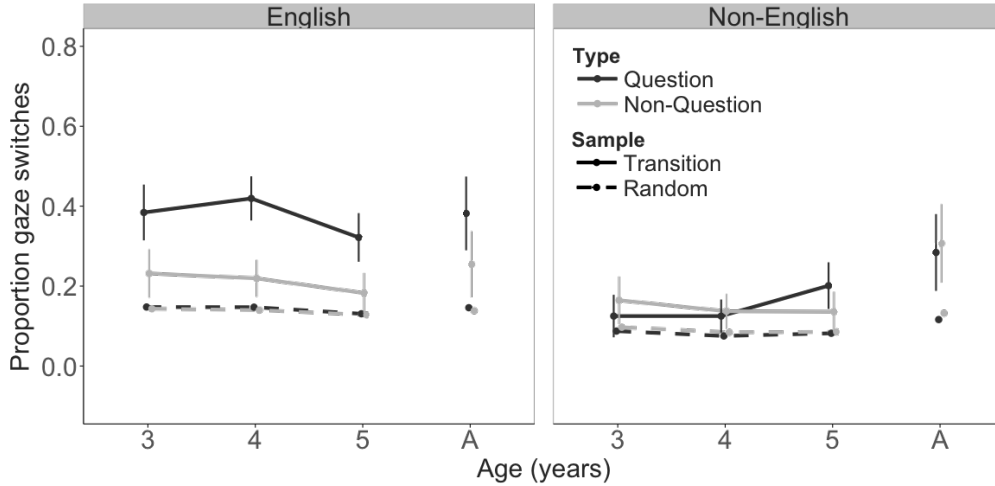


Figure 3: Anticipatory gaze rates across language condition and transition type for the real and randomly permuted datasets. Vertical bars represent 95% confidence intervals.

459 guage, transition type, and age with two mixed-effects logistic regressions
 460 (Bates et al., 2014; R Core Team, 2014). We built one model each for chil-
 461 dren and adults. We modeled children and adults separately because effects
 462 of age are only pertinent to the children’s data.

463 The child model included condition (English vs. non-English)⁶, transition
 464 type (question vs. non-question), age (3, 4, 5; numeric; intercept as age=0),

⁶Because each non-English language was represented by a single stimulus, we cannot treat individual languages as factors. Gaze behavior might be best for non-native languages that have the most structural overlap with participants’ native language: English speakers can make predictions about the strength of upcoming Swedish prosodic boundaries nearly as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

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

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

465 and duration of the inter-turn gap (seconds, e.g., 0.441) as predictors, with
466 two-way interactions between gap duration and the other simple fixed effects
467 (language condition, transition type, and age) and a three-way interaction
468 between language condition, transition type, and age. We included the two-

way interactions with gap duration in case the effect of inter-turn silence changes with age or linguistic cueing (e.g., if children older children rely less on silence as a cue).⁷ We also included random effects of item (turn transition) and participant, with maximal random slopes of condition, transition type, and their interaction for participants (Barr et al., 2013).⁸

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

Children’s anticipatory gaze switches showed effects of language condition ($\beta=-3.65$, $SE=1.16$, $z=-3.15$, $p<.01$) and transition type ($\beta=-2.95$, $SE=1.13$, $z=-2.61$, $p<.01$) with additional effects of an age-by-language condition interaction ($\beta=0.5$, $SE=0.212$, $z=2.35$, $p<.05$), a language condition-

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

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

484 by-transition type interaction ($\beta=2.69$, $SE=1.35$, $z=1.99$, $p<.05$), and a
 485 transition type-gap duration interaction ($\beta=5.52$, $SE=2.28$, $z=2.42$, $p<.05$).
 486 There were no significant effects of age or gap duration alone ($\beta=-0.002$,
 487 $SE=0.26$, $z=-0.009$, $p=.99$ and $\beta=2.25$, $SE=3.19$, $z=0.7$, $p=.48$, respec-
 488 tively).

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

496 *Random baseline comparison*

497 Our primary analysis (above) makes the assumption that participants'
 498 eye movements generally follow the turn structure of the stimulus, i.e., that
 499 participants track the current speaker and switch their gaze to the upcoming
 500 speaker near turn transitions. As just described, based on this assumption,
 501 we used linear mixed effects regressions to see how anticipatory looking is
 502 affected by aspects of participant group (e.g., age) and stimulus (e.g., transi-

tion type, language condition). But what if the assumption that participants generally track turn structure were wrong? Could these results have emerged if participants' eye movements were *not* linked to turn structure? For example, if participants were randomly looking back and forth between the two speakers, we might still find some anticipatory switching by chance. To test whether our primary results (the regression output above) could have arisen from random switching we conducted a secondary analysis comparing participants' anticipatory gaze at real and randomly shuffled points of turn transition.

We conducted this analysis by running the same regression models on participants' eye-tracking data, only this time calculating their anticipatory gaze switches with respect to randomly permuted turn transition windows. This process involved: (1) randomizing the order and temporal placement of the analysis windows within each stimulus (Figure 4; "analysis window" is as shown in Figure 2) to randomly redistribute the analysis windows across the eye-tracking signal, (2) re-running each participant's eye tracking data through switch identification (described above) on each of the randomly permuted analysis windows, and (3) modeling the anticipatory switches from the randomly permuted data (our random baseline dataset) with the same statis-

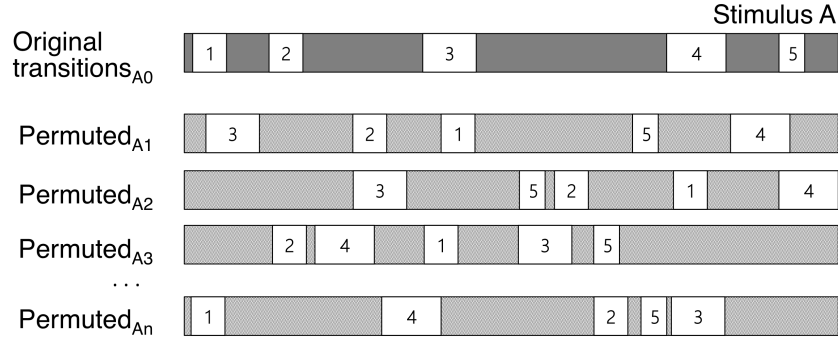


Figure 4: Example of analysis window permutations for a stimulus with five turn transitions. The windows included ± 300 msec around the inter-turn gap.

522 tical models we used for the original dataset (Table 2). Importantly, although
 523 the onset time of each transition was shuffled within the eye-tracking signal,
 524 the other intrinsic properties of each turn transition (e.g., prior speaker iden-
 525 tity, transition type, gap duration, language condition, etc.) stayed constant
 526 across each permutation.

527 The random shuffling procedure de-links participants' gaze data from the
 528 turn structure in the original stimulus, thereby allowing us to compare turn-
 529 related (original) and non-turn-related (randomly permuted) looking behav-
 530 ior using the same eye movement data. We created 5,000 permutations of the
 531 original turn transitions, thereby creating 5,000 anticipatory gaze datasets
 532 with randomly de-linked gaze data. Because the randomly shuffled turn tran-
 533 sitions could occur anywhere in the stimulus (so long as they didn't overlap

each other within a single iteration), the resulting turn-transition windows collectively covered the entire stimulus—during speech and silence, during speaker change and speaker continuation, and during all turn transitions in the stimulus, even those excluded in the original analyses (e.g., because they were overlapped). This technique crucially differs from that used by Keitel and colleagues (2013, 2015), which tests anticipatory gaze at turn transitions against anticipatory gaze during speech. Pooled together, our 5,000 anticipatory gaze datasets yielded an average anticipatory switch rate for each participant over all possible starting points in the stimuli: a random baseline. Using this technique we compared participants’ anticipatory switches at turn transition windows to their anticipatory switches over the stimulus as a whole. If participants looked randomly back and forth between the speakers, we would have seen similar patterns in both cases.

Rather than simply comparing participants’ overall anticipatory switch rates with real and random transition windows, we estimated the likelihood that each of the predictor effects in the original data (e.g., the effect of language condition; Table 2) could have arisen with random gaze switching: we ran identical statistical models on the real and randomly permuted data sets. This tells us not only whether participants’ switches were above chance,

553 but whether the specific underlying effects of their anticipatory gaze patterns
554 (e.g., the effect of language condition) were above that expected by chance.
555 Because these analyses are complex and secondary to the main results, we
556 report their full details in Supplementary material A.

557 Our baseline analyses revealed that none of the significant predictors from
558 models of the original, turn-related data can be explained by random looking.
559 For the children’s data, the original z -values for language condition, transi-
560 tion type, the age-language condition interaction, the transition type-gap
561 duration interaction, and the language condition-transition type interaction
562 were all greater than 95% of z -values from models of the randomly permuted
563 data (99.3%, 99.1%, 98.9%, 97%, and 96%, respectively, all $p < .05$). Simi-
564 larly, the adults’ data showed significant differentiation from the randomly
565 permuted data for all three significant predictors from the real transition
566 dataset. Transition type, the interaction between transition type and gap
567 duration, and the interaction between language condition and transition type
568 showed z -values that exceeded 100%, 99.8%, and 95% of random z -values,
569 respectively (all $p \leq .05$). See Supplementary material A for more information
570 on each predictor’s random permutation distribution.⁹

⁹This baseline analysis tests “random looking” against “turn-driven looking”, but it

571 *Developmental effects*

572 The models reported above revealed a significant interaction of age and
573 language condition (Table 2) that was unlikely be due to random gaze switch-
574 ing (Figure 3). To further explore this effect, we compared the effect of lan-
575 guage condition across age groups: using the permuted datasets described
576 above, we extracted the average difference score for the two language condi-
577 tions (English minus non-English) for each participant, computing an overall
578 average for each random permutation of the data. Then, within each per-
579 mutation, we made pairwise comparisons of the average difference scores
580 across participant age groups. This process yielded a distribution of random
581 permutation-based difference scores that we could then compare to the dif-
582 ference score in the actual data. Details are given in Supplementary material
583 B.

584 These analyses revealed that, while 3- and 4-year olds showed similarly-
585 sized effects of language condition, 5-year-olds had a significantly smaller
586 effect of language condition, compared to both younger age groups. The

does not test subtypes of turn-driven looking. For example, children might switch their gaze from the current speaker to the addressee out of boredom with the ongoing speech rather than from active anticipation of an upcoming response. We address this hypothesis about “boredom” gaze switches vs. “turn-transition” gaze switches in Supplementary material C.

587 difference in the language condition effect between 5-year-olds and 3-year-
588 olds was greater than would be expected by chance (99.52% of the randomly
589 permuted data sets; $p < .01$). Similarly, the difference in the language con-
590 dition effect between 5-year-olds and 4-year-olds was greater than would be
591 expected by chance (99.96% of the data sets; $p < .001$). See Supplementary
592 material Figure B.1 for each difference score distribution.

593 When does spontaneous turn prediction emerge developmentally? We
594 tested whether the youngest age group (3-year-olds) already exceeded chance
595 in their anticipatory gaze switches by comparing children's real gaze rates
596 to the random baseline in the English condition with two-tailed t -tests.
597 We used the English condition because we are most interested in finding
598 out when children begin to make spontaneous turn predictions for natu-
599 ral speech. We found that three-year-olds made anticipatory gaze switches
600 significantly above chance, when all transitions were considered ($t(22.824) =$
601 4.147 , $p < .001$) as well as for question transitions alone ($t(21.677) =$
602 -5.268 , $p < .001$).

603 *Discussion*

604 Children and adults spontaneously tracked the turn structure of the con-
605 versations, making anticipatory gaze switches at an above-chance rate across

all ages and conditions. Children’s anticipatory gaze rates were affected by language condition, transition type, age, and gap duration (Table 2), none of which could be explained by a baseline of random gaze switching (see Supplementary material Figure A.1). These data show a number of important features that bear on our questions of interest.

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

Second, children made more anticipatory switches overall in English videos, compared to non-English videos. This effect suggests that linguistic access is

625 important for children’s ability to anticipate upcoming turn structure, con-
626 sistent with prior work on turn-end prediction in adults (De Ruiter et al.,
627 2006; Magyari & De Ruiter, 2012) and children (Keitel et al., 2013).

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

641 Finally, children and adults made more anticipatory switches in tran-
642 sitions with longer inter-turn gaps, though this effect was limited to non-
643 question turns (Table 2). This finding suggests that gap duration indeed

644 serves as a cue to upcoming turn structure; while short gaps may be perceived
645 as within-turn pauses (Männel & Friederici, 2009), long gaps could instead
646 be indicative of between-turn pauses (where speaker transition occurs). Par-
647 ticipants might use long silences to retroactively assign turn boundaries and
648 anticipate speaker switches that were otherwise not anticipated (in this case,
649 because the preceding turn was not a question). An alternative explanation
650 for effects of gap duration is that longer inter-turn gaps result in longer anal-
651 ysis windows, which gives participants more time to make an anticipatory
652 gaze. However, if participants are generally more likely to make a switch at
653 question transitions (as our results suggest), and if question-driven switches
654 aren't already at ceiling when gaps are short, we would expect that longer
655 gaps would benefit questions more than non-questions—the opposite pattern
656 from what the data show here. We take this as evidence that inter-turn
657 silence may be most useful when participants have limited ability to make
658 predictions about upcoming speaker transitions.

659 In Experiment 2, we followed up on these findings, improving on two
660 aspects of the design: first, our language manipulation in this first experi-
661 ment was too coarse to provide data regarding specific linguistic information
662 channels (e.g., the effect of prosodic information alone). In Experiment 2, we

663 compared lexicosyntactic and prosodic cues with phonetically altered speech
664 and used puppets to eliminate non-verbal cues to turn taking. Second, we
665 were not able to pinpoint the emergence of anticipatory switching because
666 the youngest age group in our sample was already able to make anticipa-
667 tory switches at above-chance rates. In Experiment 2, we explored a wider
668 developmental range.

669 **Experiment 2**

670 Experiment 2 used English-only stimuli, controlled for lexical and prosodic
671 information, eliminated non-verbal cues, and tested children from a wider age
672 range. To tease apart the role of lexical and prosodic information, we phonet-
673 ically manipulated the speech signal for pitch, syllable duration, and lexical
674 access. By testing 1- to 6-year-olds we hoped to find the developmental onset
675 of turn-predictive gaze. We also hoped to measure changes in the relative
676 roles of prosody and lexicosyntax across development.

677 Non-verbal gestural cues in Experiment 1 could have helped partici-
678 pants make predictions about upcoming turn structure (Rossano et al., 2009;
679 Stivers & Rossano, 2010). Since our focus here is on linguistic cues, we
680 eliminated all gaze and gestural signals in Experiment 2 by replacing the
681 videos of human actors with videos of puppets. Puppets are less realis-

682 tic and expressive than human actors, but they create a natural context for
683 having somewhat motionless talkers in the videos. Additionally, the prosody-
684 controlled condition (described below) included small but global changes to
685 syllable duration that would have required complex video manipulation or
686 precise re-enactment with human talkers, neither of which was feasible. For
687 these reasons, we decided to use puppet videos rather than human videos in
688 the final stimuli. As in the first experiment, we recorded participants' eye
689 movements as they watched six short videos of dyadic conversation, and then
690 analyzed their anticipatory glances from the current speaker to the upcoming
691 speaker at points of turn transition.

692 *Methods*

693 *Participants*

694 We recruited 27 undergraduate adults and 129 children ages 1;0–6;11 to
695 participate in our experiment. Adult participants were recruited again via
696 the Stanford University Psychology participant database and were either paid
697 or received course credit for their time. We recruited our child participants
698 from the Children's Discovery Museum in San Jose, California¹⁰, targeting
699 approximately 20 children for each of the six one-year age groups (range:

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



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.

20–23). All participants were native English speakers, though some parents
 (N=27) reported that their child heard a second (and sometimes third) lan-
 guage at home. None of the adult participants reported fluency in a second
 language.

Materials

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

710 video to accompany the manipulated audio, featuring puppets as talkers.

711 *Audio recordings.* The recording session was set up in the same way as
712 the first experiment, but with a shorter warm up period (5–10 minutes) and
713 a pre-determined topic for the child-friendly improvisation (‘riding bikes’,
714 ‘pets’, ‘breakfast’, ‘birthday cake’, ‘rainy days’, or ‘the library’). All of the
715 talkers were native English speakers, and were recorded in male-female pairs.
716 As before, we asked talkers to speak “as if they were on a children’s television
717 show” and to ask at least a few questions during the improvisation. We cut
718 each audio recording down to the (approximate) 20-second interval with the
719 most turn activity. The 20-second clips were then phonetically manipulated
720 and used in the final video stimuli.

721 *Audio Manipulation.* We created four versions of each audio conversa-
722 tion: *normal*, *words only*, *prosody only*, and *no speech*. That is, one version
723 with a full linguistic signal (*normal*), and three with incomplete linguistic
724 information (hereafter “partial cue” conditions). The *normal* conversations
725 were the unmanipulated, original audio clips.

726 The *words only* conversations were manipulated to have robot-like speech:
727 we flattened the intonation contours to each talker’s average pitch (F_0) and
728 we reset the duration of every nucleus and coda to each talker’s average

nucleus and coda duration.¹¹ We made duration and pitch manipulations using PSOLA resynthesis in Praat (Boersma & Weenink, 2012). Thus, the *words only* versions of the conversations had no pitch or durational cues to upcoming turn boundaries, but did have intact lexicosyntactic cues (and some residual phonetic correlates of prosody, e.g., intensity).

We created the *prosody only* conversations by low-pass filtering the original recording at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006). This manipulation creates a “muffled speech” effect because low-pass filtering removes most of the phonetic information used to distinguish between phonemes. The *prosody only* versions of the conversations lacked lexical information, but retained their intonational and rhythmic cues to upcoming turn boundaries.

The *no speech* condition served as a non-linguistic baseline. For this condition, we replaced the original audio clip for the conversation with multi-talker babble: we overlaid multiple child-oriented conversations (excluding the original one), and then cropped the result to the duration of the original conversation clip. Thus, the *no speech* conversation lacked any linguistic information to upcoming turn boundaries—the only cue to turn taking was

¹¹We excluded hyper-lengthened words like [wau:] ‘wooooow!’.

747 the opening and closing of the puppets’ mouths.

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

754 *Video recordings.* We created puppet video recordings to match the ma-
755 nipulated 20-second audio clips. The puppets were minimally expressive; the
756 puppeteer could only control the opening and closing of their mouths, and
757 the puppets’ heads, eyes, arms, and bodies stayed still. Puppets were posi-
758 tioned side-by-side, looking in the same direction to eliminate shared gaze as
759 a cue to turn structure (Thorgrímsson et al., 2015). We took care to match
760 the puppets’ mouth movements to the syllable onsets as closely as possible,
761 specifically avoiding mouth movement before the onset of a turn. We then
762 added the manipulated audio clips to the puppet video recordings by hand
763 with video editing software.

764 We used three pairs of puppets for the *normal* condition—‘red’, ‘blue’
765 and ‘yellow’—and one pair of puppets for each partial cue condition: ‘robots’,

766 ‘merpeople’, and ‘party-goers’ (Figure 5). We randomly assigned half of the
767 conversation topics (‘birthday cake’, ‘pets’, and ‘breakfast’) to the *normal*
768 condition, and half to the partial cue conditions (‘riding bikes’, ‘rainy days’,
769 and ‘the library’). We then created three versions of the experiment, so that
770 each of the six puppet pairs was associated with three different conversation
771 topics across the different versions of the experiment (18 videos in total; 6
772 videos per experiment version). We ensured that the position of the talkers
773 (left and right) was counterbalanced in each version by flipping the video and
774 audio channels as needed.

775 As before, the duration of turn transitions and the number of speaker
776 changes across videos was variable because the conversations were recorded
777 semi-spontaneously. We measured turn transitions from the audio signal of
778 the *normal*, *words only*, and *prosody only* conditions. There was no audio
779 from the original conversation in the *no speech* condition videos, so we mea-
780 sured turn transitions from puppets’ mouth movements in the video signal,
781 using ELAN video annotation software (Wittenburg et al., 2006).

782 There were 85 turn transitions for analysis after excluding transitions
783 longer than 550 msec and shorter than 90 msec. The remaining turn tran-
784 sitions had more questions than non-questions (N=47 and N=38, respec-

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

Procedure

We used the same experimental apparatus and procedure as in the first experiment. Each participant watched six puppet videos in random order, with 15–30 second filler videos placed in-between (e.g., running puppies, moving balls, flying bugs). Three of the puppet videos had *normal* audio while the other three had *words only*, *prosody only*, and *no speech* audio. As before, the experimenter immediately began each session with calibration and then stimulus presentation. Participants were given no instruction about how to watch the videos or what their purpose was, they were simply encouraged to

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 in Experiment 2.

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

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

804 watch the “(fun/nice) puppet videos”. The entire experiment took less than
805 five minutes.

806 *Data preparation and coding*

807 We coded each turn transition for its linguistic condition (*normal*, *words*
808 *only*, *prosody only*, and *no speech*) and transition type (question/non-question)¹²,
809 and identified anticipatory gaze switches to the upcoming speaker using the

¹²We coded *wh*-questions as “non-questions” for the *prosody only* videos. Polar questions often have a final rising intonational contour, but *wh*-questions do not (Hedberg et al., 2010).

810 methods from Experiment 1.

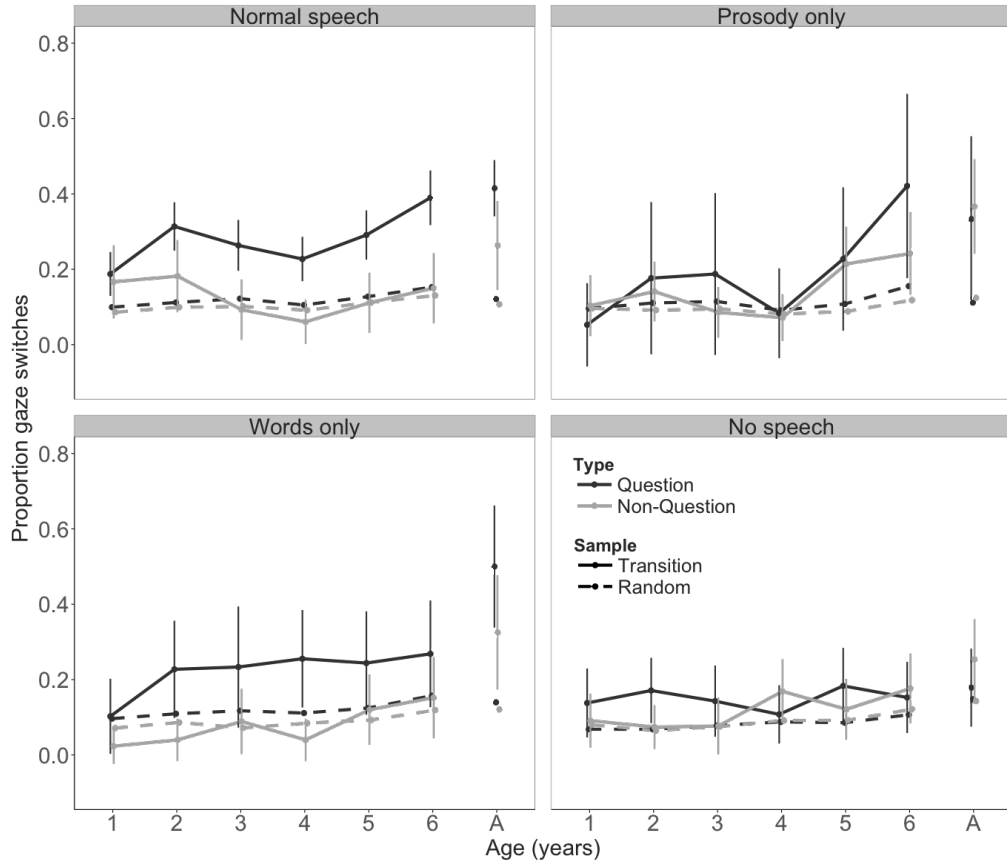


Figure 6: Anticipatory gaze rates across language condition and transition type for the real and randomly permuted datasets. Vertical bars represent 95% confidence intervals.

811 Results

812 Participants' pattern of gaze indicated that they performed basic turn
 813 tracking across all ages and in all conditions. Participants looked at the
 814 screen most of the time during video playback (82% and 86% average for

815 children and adults, respectively), primarily looking at the person who was
816 currently speaking (Tables 3 and 4). They tracked the current speaker in
817 every condition—even one-year-olds looked more at the current speaker than
818 at anything else in the three partial cue conditions (40% for *words only*, 43%
819 for *prosody only*, and 39% for *no speech*). There was a steady overall increase
820 in looks to the current speaker with age and added linguistic information
821 (Tables 3 and 4). Looks to the addressee also decreased with age, but the
822 change was minimal. Figure 6 shows participants’ anticipatory gaze rates
823 across age, the four language conditions, and transition type.

824 *Statistical models*

825 We identified anticipatory gaze switches for all 85 usable turn transi-
826 tions, and analyzed them for effects of language condition, transition type,
827 and age with two mixed-effects logistic regressions. We again built separate
828 models for children and adults because effects of age were only pertinent to
829 the children’s data. The child model included condition (*normal/prosody*
830 *only/words only/no speech*; with *no speech* as the reference level), transition
831 type (question vs. non-question), age (1, 2, 3, 4, 5, 6; numeric, intercept as
832 age=0), and duration of the inter-turn gap (in seconds) as predictors, with
833 full interactions between language condition, transition type, and age and

two-way interactions between gap duration and the other basic fixed effects (age, linguistic condition, and transition type). We also included random effects of participant and item (turn transition), with maximal random slopes of transition type for participant. The adult model included condition, transition type, their interactions, gap duration, and two-way interactions between gap duration and condition and transition type, with participant and item as random effects and maximal random slopes of condition and transition type for participant.

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

853 Adults' anticipatory gaze switches showed a significant effect of language
854 condition (for *words only* speech compared to the *no speech* reference level;
855 $\beta=3.79$, $SE=1.62$, $z=2.34$, $p<.05$) and a marginal two-way interaction be-
856 tween language condition and transition type (for *words only* speech com-
857 pared to the *no speech* reference level; $\beta=-1.68$, $SE=0.89$, $z=-1.89$, $p=.06$).
858 There was no significant effect of transition type alone (Table 6; $\beta=-0.02$,
859 $SE=1.44$, $z=-0.02$, $p=.99$).

860 *Random baseline comparison*

861 Using the same technique described in Experiment 1, we created and
862 modeled random permutations of participants' anticipatory gaze switches.
863 These analyses revealed that the significant predictors from models of the
864 original, turn-related data were unlikely to be explained by random looking.
865 In the children's data, the original model's z -values for gap duration, the
866 two-way interaction of age and language condition (*prosody only*) and the
867 three-way interaction of age, transition type, and language condition (*nor-*
868 *mal* speech) were all greater than 93% of the randomly permuted z -values
869 (95.6%, 94%, and 93.3%, respectively, $p=.04$, .06, and .07). Similarly, the
870 adults' data showed significant differentiation from the randomly permuted
871 data for the effect of language condition (*words only* speech; greater than

98.3% of random z -values, $p < .02$). See Supplementary material A for more information on each predictor's random permutation distribution.

<i>Children</i>				
	Estimate	Std. Error	z value	$\Pr(> z)$
(Intercept)	-3.452	0.76	-4.543	5.55e-06 ***
Age	-0.054	0.143	-0.379	0.705
TType= <i>non-Question</i>	-1.217	0.958	-1.27	0.204
GapDuration	3.852	1.735	2.221	0.026 *
Age*TType= <i>non-Question</i>	0.152	0.141	1.081	0.28
Age*GapDuration	0.214	0.266	0.805	0.421
TType= <i>non-Question</i> * GapDuration	0.995	2.134	0.466	0.641
Condition= <i>normal</i>	0.54	0.742	0.728	0.467
Age*Condition= <i>normal</i>	0.125	0.103	1.221	0.222
Condition= <i>normal</i> * TType= <i>non-Question</i>	0.908	0.748	1.215	0.224
Age*Condition= <i>normal</i> * TType= <i>non-Question</i>	-0.355	0.173	-2.051	0.04 *
Condition= <i>normal</i> * GapDuration	-0.431	1.67	-0.258	0.797
Condition= <i>prosody</i>	0.549	1.452	0.378	0.705
Age*Condition= <i>prosody</i>	0.375	0.191	1.967	0.049 *
Condition= <i>prosody</i> * TType= <i>non-Question</i>	1.076	1.105	0.974	0.33
Age*Condition= <i>prosody</i> * TType= <i>non-Question</i>	-0.296	0.235	-1.257	0.209
Condition= <i>prosody</i> * GapDuration	-4.767	2.625	-1.816	0.069 (.)
Condition= <i>words</i>	0.684	1.06	0.645	0.519
Age*Condition= <i>words</i>	0.127	0.136	0.934	0.350
Condition= <i>words</i> * TType= <i>non-Question</i>	-1.244	1.031	-1.207	0.228
Age*Condition= <i>words</i> * TType= <i>non-Question</i>	0.111	0.225	0.495	0.621
Condition= <i>words</i> * GapDuration	-2.285	2.232	-1.024	0.306

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

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

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

876 *Developmental effects*

877 Our main goal in extending the age range to 1- and 2-year-olds in Exper-
878 iment 2 was to find the age of emergence for spontaneous predictions about
879 upcoming turn structure. As in Experiment 1, we used two-tailed *t*-tests
880 to compare children's real gaze rates to the random baseline rates in the
881 *normal* speech condition (in which the speech stimulus is most like what
882 children hear every day). We tested real gaze rates against baseline rates
883 for three age groups: one-, two-, and three-year-olds. Two- and three-year-

884 old children made anticipatory gaze switches significantly above chance both
 885 when all transitions were considered (2-year-olds: $t(26.193)=-4.137$, $p<.001$;
 886 3-year-olds: $t(22.757)=-2.662$, $p<.05$) and for question transitions alone (2-
 887 year-olds: $t(25.345)=-4.269$, $p<.001$; 3-year-olds: $t(21.555)=-3.03$, $p<.01$).
 888 One-year-olds, however, only made anticipatory gaze shifts marginally above
 889 chance for turn transitions overall and for question turns alone (overall:
 890 $t(24.784)=-2.049$, $p=.051$; questions: $t(25.009)=-2.03$, $p=.053$).

891 We also tested the two baseline linguistic conditions against each other—
 892 *no speech* and *normal speech*—to find out when linguistic information made
 893 a difference in children’s anticipations. Because, as we have seen, children
 894 primarily show linguistic effects in question-answer turn transitions, we in-
 895 vestigated the use of linguistic cues across age by testing anticipation sep-
 896 arately for question and non-question turns. Compared to the *no speech*
 897 condition, children made significantly more anticipatory switches in the *nor-*
 898 *mal* speech condition for questions at ages 6, 4, and 3, and also marginally at
 899 age 2 (6-year-olds: $t(36.919)=3.8019$, $p<.001$; 4-year-olds: $t(41.449)=2.9777$,
 900 $p<.01$; 3-year-olds: $t(35.724)=2.4286$, $p<.05$; 2-year-olds: $t(41.078)=1.8018$,
 901 $p=.079$). Children’s anticipatory switches for questions did not significantly
 902 differ in the *no speech* and *normal* speech conditions at ages 5 or 1 (5-year-

903 olds: $t(29.406)=1.2783$, $p=.211$; 1-year-olds: $t(35.907)=0.4961$, $p=.623$). In
904 contrast, children's anticipatory switch rates for non-question turns were not
905 significantly different between the *no speech* and *normal* speech conditions at
906 any age (all $p>0.09$). Thus, consistent with the regression results, children
907 were more likely to show an effect of linguistic content as they got older, but
908 only for question transitions.

909 The regression models for the children's data also revealed two signifi-
910 cant interactions with age. The first was a significant interaction of age and
911 language condition (for *prosody only* compared to the *no speech* reference
912 level), suggesting a different age effect between the two linguistic conditions.
913 As in Experiment 1, we explored each age interaction by extracting an av-
914 erage difference score over participants for the effect of language condition
915 (*no speech* vs. *prosody only*) within each random permutation of the data,
916 making pairwise comparisons between the six age groups. These tests re-
917 vealed that children's anticipation in the *prosody only* condition significantly
918 improved at ages five and six compared to the *no speech* baseline (with dif-
919 ference scores greater than 95% of the random data scores; $p<.05$). See
920 Supplementary material Figure B.2 for these *prosody only* difference score
921 distributions.

922 The second age-based interaction was a three-way interaction of age, tran-
923 sition type, and language condition (for *normal* speech compared to the *no*
924 *speech* baseline). We again created pairwise comparisons of the average dif-
925 ference scores for the transition type-language condition interaction across
926 age groups in each random permutation of the data, finding that the effect
927 of transition type in the *normal* speech condition became larger with age,
928 with significant improvements by age 4 over ages 1 and 2 (99.9% and 98.86%,
929 respectively), by age 5 over age 4 (97.54%), and by age 6 over ages 1, 2, and 5
930 (99.5%, 97.36%, and 95.04%), all significantly different from chance ($p < .05$).
931 See Supplementary material Figure B.3 for these *normal* speech difference
932 score distributions.

933 *Discussion*

934 The core aims of Experiment 2 were to gain better traction on the indi-
935 vidual roles of prosody and lexicosyntax in children's turn predictions, and
936 to find the age of emergence for spontaneous turn anticipation. Many of our
937 results replicate the findings from Experiment 1: participants often made
938 more anticipatory switches when they had access to linguistic information
939 and, when they did, tended to make more anticipatory switches for ques-
940 tions compared to non-questions.

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

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

960 the developing relationship between the online use of linguistic cues, inter-
961 turn silence, and speech act in spontaneous predictions about upcoming turn
962 structure.

963 **General Discussion**

964 Children begin to develop conversational turn-taking skills long before
965 their first words emerge (Bateson, 1975; Hilbrink et al., 2015; Jaffe et al.,
966 2001; Snow, 1977). As they become fast and knowledgeable language users,
967 they also become able to make accurate predictions about upcoming turn
968 structure. Until recently, we have had very little data on how children weave
969 language into their already-existing turn-taking behaviors. In two experi-
970 ments investigating children’s anticipatory gaze to upcoming speakers, we
971 found evidence that turn prediction develops early in childhood and that,
972 when spontaneous predictions begin, they are primarily driven by partici-
973 pants’ expectation of an immediate response in the next turn (e.g., after
974 questions). In making predictions about upcoming turn structure, children
975 used a combination of lexical and prosodic cues; neither signal alone was
976 sufficient to support increased anticipatory gaze. We also found no early
977 advantage for prosody over lexicosyntax; children’s anticipatory switch rates
978 in the *prosody only* condition were initially low, but showed significant gains

979 by age five. We discuss these findings with respect to the role of linguis-
980 tic processing and inter-turn silence for predicting upcoming turn structure,
981 the importance of questions in predictions about conversation, and children's
982 developing competence as conversationalists.

983 *Predicting upcoming turn structure*

984 Prior work with adults has found a consistent role for lexicosyntax in pre-
985 dicting upcoming turn structure (De Ruiter et al., 2006; Magyari & De Ruiter,
986 2012), whereas the role of prosody is still under debate (Duncan, 1972; Ford
987 & Thompson, 1996; Bögels & Torreira, 2015). Knowing that children compre-
988 hend more about prosody than lexicosyntax early on (see Speer & Ito, 2009
989 for a review), we thought it possible that young children would instead show
990 an advantage for prosody in their predictions about turn structure in con-
991 versation. Our results suggest that, on the contrary, exclusively presenting
992 prosodic information to children limits their spontaneous predictions about
993 upcoming turn structure until age five.

994 Thus, using prosody alone to accurately predict turn boundaries in con-
995 versation appears to be difficult for adults and children. Prosodic information
996 is continuous, multidimensional, and can index multiple meanings at once—
997 it encodes syntactic structure, speech act, and extralinguistic information

without clear one-to-one mappings between form and meaning (Cutler et al., 1997; Shriberg et al., 1998; Lammertink et al., 2015). For these reasons, prosodic information alone may not be enough for young children to easily make precise temporal predictions about turn structure, and identify question turns in unfolding speech. Therefore, although children show early facility with prosodic discrimination (Nazzi & Ramus, 2003; Soderstrom et al., 2003; Johnson & Jusczyk, 2001; Jusczyk et al., 1995; Morgan & Saffran, 1995; Mehler et al., 1988), using prosodic knowledge for turn prediction may be difficult without additional information from lexical or syntactic cues.

Our findings suggest that there is one prosodic cue that is an exception to this rule: inter-turn silence. Generally speaking, participants showed a greater anticipatory switches for longer inter-turn gaps, but the effect of inter-turn gap duration is strongest in our data when upcoming responses are less predictable, whether due to the asymmetrical response expectations for questions vs. non-questions (Experiment 1) or the lack of non-verbal cues and any linguistic information (Experiment 2). Notably, there were no significant interactions of gap duration with participant age. This pattern of results suggests that, when predictive information about upcoming responses is absent, long silences may increase participants' expectation for

1017 a speaker change and promote more anticipatory gaze switches. Pauses are
1018 detected and related to phrasal structure from early on; 5-month-old infants
1019 use pauses to parse intonational phrases (Männel & Friederici, 2009). The
1020 lack of interactions between age and gap duration suggests that the use of
1021 inter-turn silence remains important for older speakers and the interactions
1022 between transition type and gap duration (Experiment 1) and condition and
1023 gap duration (Experiment 2; marginal), suggest that this effect is not simply
1024 the result of having more time to make a gaze switch. These findings thus
1025 suggest that silence is an early and lasting cue for identifying turn structure
1026 online when other predictive information is not adequate.

1027 Notably, many other non-linguistic cues encode information about tran-
1028 sition type, including gaze and gesture. We did not systematically test those
1029 cues here but, like inter-turn silence, they may play a critical role in parsing
1030 and making predictions about turn structure when other linguistic informa-
1031 tion is not sufficient to make accurate predictions.

1032 Perhaps surprisingly, we found no evidence that lexical information alone
1033 is equivalent to the full linguistic signal in driving children’s predictions, as
1034 has been shown previously for adults (Magyari & De Ruiter, 2012; De Ruiter
1035 et al., 2006) and as is replicated with adult participants in the current study.

1036 Unlike prosodic cues, lexicosyntactic cues are discreet and have much clearer
1037 form-to-meaning mappings, with clear lexicosyntactic cues to questionhood
1038 that occur early within turns (e.g., *wh*-words, *do*-insertion, and subject-
1039 auxiliary inversion). That said, children’s lexical and syntactic knowledge is
1040 limited for quite some time (Tomasello & Brooks, 1999, but see also Bergel-
1041 son & Swingley, 2013; Shi & Melancon, 2010). Although our stimuli were
1042 made in a child-friendly style, they are still other-directed and fairly complex,
1043 with 20–30 seconds of continuous conversational speech.

1044 It is perhaps for this reason that children’s performance was always best
1045 with the full signal, where lexicosyntactic information was supported by
1046 prosodic information and vice versa. Even in adults, Bögels and Torreira
1047 (2015) showed that the trade-off in informativity between lexical and prosodic
1048 cues is more subtle in semi-natural speech. The present findings are the first
1049 to show evidence of a similar effect developmentally.

1050 *The question effect*

1051 In both experiments, anticipatory looking was primarily driven by ques-
1052 tion transitions, a pattern that has not been previously reported in other an-
1053 ticipatory gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari,
1054 2013; Tice and Henetz, 2011). Questions make an upcoming speaker switch

1055 immediately relevant, helping the listener to predict with high certainty what
1056 will happen next (i.e., an answer from the addressee), and are often easily
1057 identifiable by overt prosodic and lexicosyntactic cues.

1058 Prior work on children’s acquisition of questions indicates that they may
1059 already have some knowledge of question-answer sequences by the time they
1060 begin to speak: questions make up approximately one third of the utter-
1061 ances children hear, before and after the onset of speech, and even into
1062 their preschool years, though the type and complexity of questions changes
1063 throughout development (Casillas et al., 2016; Fitneva, 2012; Henning et al.,
1064 2005; Shatz, 1979).¹³ For the first few years, many of the questions directed
1065 to children are “test” questions—questions that the caregiver already has the
1066 answer to (e.g., “What does a cat say?”), but this changes as children get
1067 older. Questions help caregivers to get their young children’s attention and to
1068 ensure that information is in common ground, even if the responses are non-
1069 verbal or infelicitous (Bruner, 1985; Fitneva, 2012; Snow, 1977). Moreover,
1070 because of their high frequency and relatively limited number of formats,
1071 questions, especially *wh*-questions, may be more identifiable and predictable

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

1072 compared to other types of speech acts. So, in addition to having a special
1073 interactive status, questions are a frequent, predictable, and core character-
1074 istic of many caregiver-child interactions, motivating a general benefit for
1075 questions in turn structure anticipation.

1076 Two important routes for future work are then: (1) how does children’s
1077 ability to monitor for questions in conversation relate to their prior experi-
1078 ence with questions? and (2) what is it about questions that makes children
1079 and adults more likely to anticipatorily switch their gaze to addressees? If
1080 this “question” effect exists for all turns that require an immediate response
1081 (“adjacency pairs”; Schegloff, 2007), other turn types, such as imperatives,
1082 compliments, and complaints should show similar patterns. If the effect is
1083 instead about overall predictability of the syntactic frame, children would
1084 instead show similar patterns for other frequent frames from child-directed
1085 speech (e.g., “Look at the X”; Mintz, 2003). The recognizability and pre-
1086 dictability of syntactic frames is likely to play a role in turn prediction as
1087 children become more sophisticated language users, even if the effect is truly
1088 about adjacency pairs; for example, rhetorical and tag questions take a very
1089 similar form to prototypical polar questions, but usually do not require an
1090 answer. So, though it is clear that adults and children anticipate responses

1091 more often for questions than non-questions, we do not yet know whether
1092 their predictive action is limited to turns formatted as questions, turns with
1093 high recognizability and predictability, or turns that project an immediate
1094 response from the addressee.

1095 A question effect suggests that participants' spontaneous predictions may
1096 be driven by what lies *beyond* the end of the current turn—not just by the
1097 upcoming end of the turn itself, as has been focused on in prior work (Bögels
1098 & Torreira, 2015; Keitel et al., 2013; Magyari & De Ruiter, 2012; De Ruiter
1099 et al., 2006). In future work, it will be crucial to measure prediction from
1100 a first-person perspective to find out what kinds of predictions are most
1101 relevant to addressees in conversation.

1102 One possible scenario is that listeners in spontaneous, first-person con-
1103 versation use multiple strategies to make predictions about upcoming turn
1104 structure: they could semi-passively attend to incoming speech for cues to
1105 upcoming speaker transition (e.g., questions and other adjacency pairs) and,
1106 when possible upcoming transition is detected, switch into a more precise
1107 turn-end prediction mode (à la De Ruiter et al., 2006). A flexible prediction
1108 system like this one allows listeners to continuously monitor ongoing conver-
1109 sation for turn-related cues at a low cost while still managing to plan their

1110 responses and come in quickly when needed.

1111 To test this hypothesis, we would need to look at prediction from a first-
1112 person perspective, which very little work so far has accomplished (present
1113 work included). Although third-party measures enable us to measure partic-
1114 ipants' predictions without any interference from language production, they
1115 also limit our knowledge about how the need to give a response might it-
1116 self play an important role in addressees' prediction strategies. Recent work
1117 has shown that shifts in addressee gaze similar to those measured here in-
1118 deed occur in spontaneous conversation (Holler & Kendrick, 2015), but much
1119 more work is needed to determine how participants make predictions about
1120 turn structure in first-person contexts and whether those mechanisms shift
1121 at points of imminent speaker change.

1122 *Early competence for turn taking?*

1123 One of the core aims of our study was to test whether children show an
1124 early competence for turn taking, as is proposed by studies of spontaneous
1125 mother-infant proto-conversation and theories about the mechanisms under-
1126 lying human interaction in general (Hilbrink et al., 2015; Levinson, 2006).
1127 We found evidence that young children make spontaneous predictions about
1128 upcoming turn structure: definitely by age two and marginally by age one.

1129 These results contrast with Keitel and colleagues' (2013) finding that chil-
1130 dren cannot anticipate upcoming turn structure at above-chance rates until
1131 age three. The current study used an appreciably more conservative random
1132 baseline than the one used in Keitel and colleagues' study. Therefore, this
1133 difference in age of emergence more likely stems from our use of a more en-
1134 gaging speech style, stereo speech playback, and more typical turn transition
1135 durations. The child-friendly style of speech in particular may have helped in
1136 two ways: keeping children more engaged with the stimuli and using less syn-
1137 tactically complex and more prosodically exaggerated speech (Fernald et al.,
1138 1989; Werker & McLeod, 1989; Snow, 1977) compared to what they would
1139 get with adult-adult conversation.

1140 To be clear, young children's "above chance" performance was often still
1141 far from adult-like predictive behavior—turn prediction (and the concurrent
1142 use of linguistic cues from unfolding speech) increased only gradually with
1143 age. Children at ages one and two were still very close to chance in their
1144 anticipations and, even at age six, children were not fully adult-like in their
1145 predictions. This indicates that young children may at first rely primarily
1146 on non-verbal cues, like inter-turn silence, to anticipate turn transitions but
1147 that, by adulthood, listeners use both verbal and non-verbal cues to make

1148 predictions. Relatedly, adult listeners may be more expert in flexibly adapt-
1149 ing to the turn-relevant cues present at any moment, e.g., responding to
1150 non-English prosodic cues in Experiment 1.

1151 Taken together, our data suggest that turn-taking skills do begin to
1152 emerge in infancy, but that children cannot consistently make effective pre-
1153 dictions until they can identify question turns in unfolding speech and react
1154 to them quickly. This finding leads us to wonder how participant role (first-
1155 instead of third-person) and differences in early interactional experience (e.g.,
1156 frequent vs. infrequent question-asking from caregivers) feed into this early
1157 predictive skill. It also bridges prior work showing a predisposition for turn
1158 taking in infancy (e.g., Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001;
1159 Snow, 1977) with children’s apparently *late* acquisition of adult-like com-
1160 petence for turn taking in spontaneous conversation (Casillas et al., 2016;
1161 Garvey, 1984; Garvey & Berninger, 1981; Ervin-Tripp, 1979). It also rein-
1162 forces the idea that it takes children several years to fully integrate linguistic
1163 information into their turn-taking systems (Casillas et al., 2016; Garvey &
1164 Berninger, 1981).

1165 What makes the integration of linguistic information so gradual? We
1166 suspect that two slow-developing processes—children’s linguistic knowledge

1167 (e.g., *wh*-words, subject-auxiliary inversion) and their speed of processing for
1168 linguistic information (e.g., parsing and retrieval)—both contribute to their
1169 ability to make predictions about turn structure in unfolding speech. Chil-
1170 dren may be able to integrate predictive cues for turn taking from the start,
1171 but their knowledge of these cues and their speed in parsing and recognizing
1172 them may be too slow at first for use in online prediction. This account
1173 falls in line with the early and continued use of non-verbal cues found in the
1174 current study, but more work is needed to tease these developmental threads
1175 apart.

1176 *Limitations and future work*

1177 There are at least two major limitations to our work: speech naturalness
1178 and participant role. Following prior work (De Ruiter et al., 2006; Keitel
1179 et al., 2013), we used phonetically manipulated speech in Experiment 2.
1180 This decision resulted in speech sounds that children don't usually hear in
1181 their natural environment. Many prior studies have used phonetically-altered
1182 speech with infants and young children (cf. Jusczyk, 2000), but few of them
1183 have done so in a conversational context. Future work could instead carefully
1184 script speech or cross-splice sub-parts of turns to control for the presence of
1185 linguistic cues for turn transition (see, e.g., Bögels & Torreira, 2015).

1186 The prediction measure used in our studies is based on an observer’s view
1187 of conversation but, because participants’ role in the interaction could affect
1188 their online predictions about turn taking, an ideal measure would instead
1189 capture first-person predictions. If conversational participants’ predictions
1190 are partly shaped by their need to respond, first-person measures of spon-
1191 taneous turn prediction will be key to revealing how participants distribute
1192 their attention over verbal and non-verbal cues while taking part in everyday
1193 interaction, the implications of which relate to theories of online language
1194 processing for both language learning and everyday talk.

1195 That said, the third-person paradigm used in the present study still has
1196 much to tell us about turn prediction. The task is natural and intuitive
1197 in that no instruction is required, which means that it captures spontaneous
1198 predictive behavior and can be used with participants of all ages. Frequencies
1199 of anticipatory gaze switching appear to be stable across language commu-
1200 nities where similar tasks have been tested (Keitel et al., 2013; Keitel &
1201 Daum, 2015; Holler & Kendrick, 2015; Hirvenkari et al., 2013)—even from a
1202 first-person perspective—so the task is one that measures robust predictive
1203 behavior relevant to conversational processing across languages. It also lends
1204 itself to many possibilities for controlling the presence of individual verbal

1205 and non-verbal cues and has a clear method for assessing random switch-
1206 ing baselines across the entire stimulus. Also, if it is the case that response
1207 preparation interferes with our ability to see prediction at the ends of incom-
1208 ing turns (Levinson, 2016), third-person paradigms are one of the only ways
1209 to measure prediction processes in isolation.

1210 The current findings also make predictions about what we would see in
1211 first-person paradigms. For example, a focus on possible upcoming speaker
1212 transitions is even more important when the participants themselves may
1213 need to respond; we would thus expect question-like effects to occur in first-
1214 person paradigms, and perhaps even be amplified compared to third-person
1215 paradigms. If so, participants' use of linguistic information would still sub-
1216 serve this goal, with prediction at a premium. Regarding development, the
1217 same facts about the complexity of prosody-based prediction and children's
1218 initial limited lexical inventories would still hold, as would the use of silence
1219 and non-verbal cues to assess and predict turn structure in the absence of
1220 clear predictive linguistic information. The paradigm presented here thus has
1221 important contributions to make in our understanding of how participants
1222 attend to and make predictions about conversational interaction.

1223 *Conclusions*

1224 Conversation plays a central role in children’s language learning. It is
1225 the driving force behind what children say and what they hear. Adults use
1226 linguistic information to accurately predict turn structure in conversation,
1227 which facilitates their online comprehension and allows them to respond rel-
1228 evantly and on time. The present study offers new findings regarding the
1229 role of speech acts and linguistic processing in online turn prediction, and
1230 has given evidence that turn prediction emerges by age two, increases with
1231 age, and is driven by the ability to identify and react to question turns in un-
1232 folding speech. However, children’s successful integration of online linguistic
1233 processing and online predictions about upcoming turn structure develops
1234 gradually. When participants can’t use predictive linguistic cues (because
1235 they are absent, unfamiliar, or are processed too late), children and adults
1236 alike rely on retroactive cues such as inter-turn silence to predict upcoming
1237 speaker change. Using language to make predictions about upcoming inter-
1238 active content takes time to develop and, for participants of all ages appears
1239 to be primarily driven by participants’ expectations about what will happen
1240 next, beyond the end of the current turn.

Acknowledgements

We gratefully acknowledge the parents and children at Bing Nursery School and the Children’s Discovery Museum of San Jose. This work was supported by an ERC Advanced Grant to Stephen C. Levinson (269484-INTERACT), an NSF Graduate Research Fellowship and NSF Dissertation Improvement Grant to MC, and a Merck Foundation fellowship to MCF. Earlier versions of these data and analyses were presented to conference audiences (Casillas & Frank, 2012, 2013). We also thank Tania Henetz, Francisco Torreira, Stephen C. Levinson, and Eve V. Clark for their feedback on earlier versions of this work. The analysis code for this project can be found on GitHub at https://github.com/langcog/turn_taking/.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at XX.

References

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

- 1259 Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014).
 1260 *lme4: Linear mixed-effects models using Eigen and S4*. URL:
 1261 <https://github.com/lme4/lme4/http://lme4.r-forge.r-project.org/>
 1262 [Computer program] R package version 1.1-7.
- 1263 Bateson, M. C. (1975). Mother-infant exchanges: The epigenesis of conver-
 1264 sational interaction. *Annals of the New York Academy of Sciences*, 263,
 1265 101–113.
- 1266 Bergelson, E., & Swingle, D. (2013). The acquisition of abstract words by
 1267 young infants. *Cognition*, 127, 391–397.
- 1268 Bloom, K. (1988). Quality of adult vocalizations affects the quality of infant
 1269 vocalizations. *Journal of Child Language*, 15, 469–480.
- 1270 Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer*.
 1271 URL: <http://www.praat.org> [Computer program] Version 5.3.16.
- 1272 Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of
 1273 response planning occur midway through an incoming question in conver-
 1274 sation. *Scientific Reports*, 5, Article number: 12881.
- 1275 Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase bound-

- aries to project turn ends in spoken interaction. *Journal of Phonetics*, 52,
46–57.
- Bruner, J. (1985). Child’s talk: Learning to use language. *Child Language
Teaching and Therapy*, 1, 111–114.
- Bruner, J. S. (1975). The ontogenesis of speech acts. *Journal of Child Lan-
guage*, 2, 1–19.
- Carlson, R., Hirschberg, J., & Swerts, M. (2005). Cues to upcoming Swedish
prosodic boundaries: Subjective judgment studies and acoustic correlates.
Speech Communication, 46, 326–333.
- Casillas, M., Bobb, S. C., & Clark, E. V. (2016). Turn taking, timing, and
planning in early language acquisition. *Journal of Child Language*, (pp.
1–28).
- Casillas, M., & Frank, M. C. (2012). Cues to turn boundary prediction in
adults and preschoolers. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson
(Eds.), *Proceedings of SemDial (SeineDial): The 16th Workshop on the
Semantics and Pragmatics of Dialogue* (pp. 61–69).
- Casillas, M., & Frank, M. C. (2013). The development of predictive processes

- 1293 in children’s discourse understanding. In M. Knauff, M. Pauen, N. Sebanz,
1294 & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the*
1295 *Cognitive Science Society* (pp. 299–304).
- 1296 Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the com-
1297 prehension of spoken language: A literature review. *Language and Speech*,
1298 *40*, 141–201.
- 1299 De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of
1300 a speaker’s turn: A cognitive cornerstone of conversation. *Language*, *82*,
1301 515–535.
- 1302 De Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed
1303 conversations: coordinating stroke-to-stroke turn boundaries. *Frontiers in*
1304 *Psychology*, *6*.
- 1305 Dingemanse, M., Torreira, F., & Enfield, N. (2013). Is “Huh?” a univer-
1306 sal word? Conversational infrastructure and the convergent evolution of
1307 linguistic items. *PloS one*, *8*, e78273.
- 1308 Duncan, S. (1972). Some signals and rules for taking speaking turns in
1309 conversations. *Journal of Personality and Social Psychology*, *23*, 283–292.

- 1310 Ervin-Tripp, S. (1979). Children's verbal turn-taking. In E. Ochs, & B. B.
1311 Schieffelin (Eds.), *Developmental Pragmatics* (pp. 391–414). Academic
1312 Press, New York.
- 1313 Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies,
1314 B., & Fukui, I. (1989). A cross-language study of prosodic modifications
1315 in mothers' and fathers' speech to preverbal infants. *Journal of Child*
1316 *Language*, 16, 477–501.
- 1317 Fitneva, S. (2012). Beyond answers: questions and children's learning. In
1318 J.-P. De Ruiter (Ed.), *Questions: Formal, Functional, and Interactional*
1319 *Perspectives* (pp. 165–178). Cambridge University Press, Cambridge, UK.
- 1320 Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation:
1321 Syntactic, intonational, and pragmatic resources for the management of
1322 turns. *Studies in Interactional Sociolinguistics*, 13, 134–184.
- 1323 Garvey, C. (1984). *Children's Talk: Volume 21*. Harvard University Press.
- 1324 Garvey, C., & Berninger, G. (1981). Timing and turn taking in children's
1325 conversations. *Discourse Processes*, 4, 27–57.
- 1326 Gísladóttir, R., Chwilla, D., & Levinson, S. C. (2015). Conversation electri-

- 1327 fied: ERP correlates of speech act recognition in underspecified utterances.
1328 *PloS one*, *10*, e0120068.
- 1329 Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psy-*
1330 *chological Science*, *11*, 274–279.
- 1331 Hart, B., & Risley, T. R. (1992). American parenting of language-learning
1332 children: Persisting differences in family-child interactions observed in nat-
1333 ural home environments. *Developmental Psychology*, *28*, 1096–1105.
- 1334 Hedberg, N., Sosa, J. M., Görgülü, E., & Mameni, M. (2010). The prosody
1335 and meaning of Wh-questions in American English. In *The Proceedings of*
1336 *Speech Prosody* (pp. 100045:1–4).
- 1337 Henning, A., Striano, T., & Lieven, E. V. (2005). Maternal speech to infants
1338 at 1 and 3 months of age. *Infant Behavior and Development*, *28*, 519–536.
- 1339 Hilbrink, E., Gattis, M., & Levinson, S. C. (2015). Early developmental
1340 changes in the timing of turn-taking: A longitudinal study of mother-
1341 infant interaction. *Frontiers in Psychology*, *6*.
- 1342 Hirvenkari, L., Ruusuvuori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A.,

- 1343 & Hari, R. (2013). Influence of turn-taking in a two-person conversation
1344 on the gaze of a viewer. *PloS one*, 8, e71569.
- 1345 Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in
1346 multi-person interaction. *Frontiers in Psychology*, 6.
- 1347 Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat,
1348 P., & Stern, D. N. (2001). *Rhythms of dialogue in infancy: Coordinated*
1349 *timing in development*. Monographs of the Society for Research in Child
1350 Development. JSTOR.
- 1351 Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-
1352 olds: When speech cues count more than statistics. *Journal of Memory*
1353 *and Language*, 44, 548–567.
- 1354 Jusczyk, P. W. (2000). *The Discovery of Spoken Language*. MIT press.
- 1355 Jusczyk, P. W., Hohne, E., Mandel, D., & Strange, W. (1995). Picking up
1356 regularities in the sound structure of the native language. *Speech perception*
1357 *and linguistic experience: Theoretical and methodological issues in cross-*
1358 *language speech research*, (pp. 91–119).

- 1359 Kail, R. (1991). Developmental change in speed of processing during child-
1360 hood and adolescence. *Psychological Bulletin*, 109, 490.
- 1361 Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of
1362 prediction in incremental sentence processing: Evidence from anticipatory
1363 eye movements. *Journal of Memory and Language*, 49, 133–156.
- 1364 Keitel, A., & Daum, M. M. (2015). The use of intonation for turn anticipation
1365 in observed conversations without visual signals as source of information.
1366 *Frontiers in Psychology*, 6.
- 1367 Keitel, A., Prinz, W., Friederici, A. D., Hofsten, C. v., & Daum, M. M.
1368 (2013). Perception of conversations: The importance of semantics and
1369 intonation in childrens development. *Journal of Experimental Child Psy-*
1370 *chology*, 116, 264–277.
- 1371 Lammertink, I., Casillas, M., Benders, T., Post, B., & Fikkert, P. (2015).
1372 Dutch and english toddlers' use of linguistic cues in predicting upcoming
1373 turn transitions. *Frontiers in Psychology*, 6.
- 1374 Lemasson, A., Glas, L., Barbu, S., Lacroix, A., Guilloux, M., Remeuf, K., &
1375 Koda, H. (2011). Youngsters do not pay attention to conversational rules:
1376 is this so for nonhuman primates? *Nature Scientific Reports*, 1.

- 1377 Levelt, W. J. (1989). *Speaking: From intention to articulation*. MIT press.
- 1378 Levinson, S. C. (2006). On the human “interaction engine”. In N. Enfield,
1379 & S. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and*
1380 *Interaction* (pp. 39–69). Oxford: Berg.
- 1381 Levinson, S. C. (2013). Action formation and ascriptions. In T. Stivers, &
1382 J. Sidnell (Eds.), *The Handbook of Conversation Analysis* (pp. 103–130).
1383 Wiley-Blackwell, Malden, MA.
- 1384 Levinson, S. C. (2016). Turn-taking in Human Communication – Origins
1385 and Implications for Language Processing. *Trends in Cognitive Sciences*,
1386 *20*, 6–14.
- 1387 Magyari, L., Bastiaansen, M. C. M., De Ruiter, J. P., & Levinson, S. C.
1388 (2014). Early anticipation lies behind the speed of response in conversation.
1389 *Journal of Cognitive Neuroscience*, *26*, 2530–2539.
- 1390 Magyari, L., & De Ruiter, J. P. (2012). Prediction of turn-ends based on
1391 anticipation of upcoming words. *Frontiers in Psychology*, *3*:376, 1–9.
- 1392 Männel, C., & Friederici, A. D. (2009). Pauses and intonational phrasing:

- 1393 ERP studies in 5-month-old German infants and adults. *Journal of Cog-*
1394 *nitive Neuroscience*, *21*, 1988–2006.
- 1395 Masataka, N. (1993). Effects of contingent and noncontingent maternal stim-
1396 ulation on the vocal behaviour of three-to four-month-old Japanese infants.
1397 *Journal of Child Language*, *20*, 303–312.
- 1398 Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-
1399 Tison, C. (1988). A precursor of language acquisition in young infants.
1400 *Cognition*, *29*, 143–178.
- 1401 Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in
1402 child directed speech. *Cognition*, *90*, 91–117.
- 1403 Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential
1404 and suprasegmental information in preverbal speech segmentation. *Child*
1405 *Development*, *66*, 911–936.
- 1406 Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic
1407 rhythm by infants. *Speech Communication*, *41*, 233–243.
- 1408 Nomikou, I., & Rohlfing, K. J. (2011). Language does something: Body

- 1409 action and language in maternal input to three-month-olds. *IEEE Trans-*
1410 *actions on Autonomous Mental Development*, 3, 113–128.
- 1411 R Core Team (2014). *R: A Language and Environment for Statistical Com-*
1412 *puting*. R Foundation for Statistical Computing Vienna, Austria. URL:
1413 <http://www.R-project.org> [Computer program] Version 3.1.1.
- 1414 Ratner, N., & Bruner, J. (1978). Games, social exchange and the acquisition
1415 of language. *Journal of Child Language*, 5, 391–401.
- 1416 Reddy, V., Markova, G., & Wallot, S. (2013). Anticipatory adjustments to
1417 being picked up in infancy. *PloS one*, 8, e65289.
- 1418 Ross, H. S., & Lollis, S. P. (1987). Communication within infant social games.
1419 *Developmental Psychology*, 23, 241–248.
- 1420 Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and cul-
1421 ture. In J. Sidnell (Ed.), *Conversation Analysis: Comparative Perspectives*
1422 (pp. 187–249). Cambridge University Press, Cambridge.
- 1423 Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for
1424 the organization of turn-taking for conversation. *Language*, 50, 696–735.

- 1425 Schegloff, E. A. (2007). *Sequence organization in interaction: Volume 1: A*
1426 *primer in conversation analysis*. Cambridge University Press.
- 1427 Shatz, M. (1978). On the development of communicative understandings:
1428 An early strategy for interpreting and responding to messages. *Cognitive*
1429 *Psychology*, 10, 271–301.
- 1430 Shatz, M. (1979). How to do things by asking: Form-function pairings in
1431 mothers' questions and their relation to children's responses. *Child Devel-*
1432 *opment*, 50, 1093–1099.
- 1433 Shi, R., & Melancon, A. (2010). Syntactic categorization in French-learning
1434 infants. *Infancy*, 15, 517–533.
- 1435 Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R.,
1436 Taylor, P., Ries, K., Martin, R., & Van Ess-Dykema, C. (1998). Can
1437 prosody aid the automatic classification of dialog acts in conversational
1438 speech? *Language and Speech*, 41, 443–492.
- 1439 Snow, C. E. (1977). The development of conversation between mothers and
1440 babies. *Journal of Child Language*, 4, 1–22.
- 1441 Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003).

- 1442 The prosodic bootstrapping of phrases: Evidence from prelinguistic in-
 1443 fants. *Journal of Memory and Language*, 49, 249–267.
- 1444 Speer, S. R., & Ito, K. (2009). Prosody in first language acquisition–
 1445 Acquiring intonation as a tool to organize information in conversation.
 1446 *Language and Linguistics Compass*, 3, 90–110.
- 1447 Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann,
 1448 T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E. et al. (2009).
 1449 Universals and cultural variation in turn-taking in conversation. *Proceed-*
 1450 *ings of the National Academy of Sciences*, 106, 10587–10592.
- 1451 Stivers, T., & Rossano, F. (2010). Mobilizing response. *Research on Language*
 1452 *and Social Interaction*, 43, 3–31.
- 1453 Takahashi, D. Y., Narayanan, D. Z., & Ghazanfar, A. A. (2013). Coupled
 1454 oscillator dynamics of vocal turn-taking in monkeys. *Current Biology*, 23,
 1455 2162–2168.
- 1456 Thorgrímsson, G., Fawcett, C., & Liszkowski, U. (2015). 1- and 2-year-olds’
 1457 expectations about third-party communicative actions. *Infant Behavior*
 1458 *and Development*, 39, 53–66.

- 1459 Tice (Casillas), M., & Henetz, T. (2011). Turn-boundary projection: Looking
1460 ahead. In L. Carlson, C. Hlscher, & T. Shipley (Eds.), *Proceedings of the*
1461 *33rd Annual Meeting of the Cognitive Science Society* (pp. 838–843).
- 1462 Toda, S., & Fogel, A. (1993). Infant response to the still-face situation at 3
1463 and 6 months. *Developmental Psychology*, *29*, 532–538.
- 1464 Tomasello, M., & Brooks, P. J. (1999). Early syntactic development: A
1465 construction grammar approach. In M. Barrett (Ed.), *The development of*
1466 *language* (pp. 161–190). Psychology Press.
- 1467 Weisleder, A. (2012). *Richer language experience leads to faster understand-*
1468 *ing: Links between language input, processing efficiency, and vocabulary*
1469 *growth*. Ph.D. thesis Stanford University.
- 1470 Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and
1471 female infant-directed talk: A developmental study of attentional and af-
1472 fective responsiveness. *Canadian Journal of Psychology/Revue Canadienne*
1473 *de Psychologie*, *43*, 230–246.
- 1474 Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H.
1475 (2006). Elan: a professional framework for multimodality research. In
1476 *Proceedings of LREC*.