Exploring Young Children's Engagement in Joint Reading with A Conversational Agent

Ying Xu
University of California, Irvine
Irvine, United States
ying.xu@uci.edu

Mark Warschauer
University of California, Irvine
Irvine, United States
markw@uci.edu

ABSTRACT

Joint book reading is a highly routinized activity that is nearly universal among families. Conversational agents (CAs) can potentially act as joint-reading partners by engaging children in story-related, scaffolded conversations. In this project, we develop a CA reading partner that incorporates components of effective conversational guidance (i.e., questions to stimulate thinking, specific feedback, and adaptive scaffolding) and examine children's interactions with this CA. We identify patterns in children's language production, flow maintenance, and affect when responding to the CA. We then lay out a set of affordances and challenges for developing CAs as conversation partners. We propose that, rather than attempting to develop CAs as an exact replicate of human conversational partners, we should treat child-agent interaction as a new genre of conversation and calibrate CAs based on children's actual communicative practices and needs.

Author Keywords

Conversational agents; joint reading; conversation analysis; design; early literacy

CCS Concepts

•Human-centered computing → Natural language interfaces; Empirical studies in HCI; •Social and professional topics → Children;

INTRODUCTION

Joint book reading is a highly routinized activity engaged in by families across cultures. Joint reading provides a focused and interactive literacy environment, which is believed to boost children's language development and long-run academic success [44]. One key ingredient to such benefits is the meaningful conversation between the child and parent during joint reading [9, 14, 23]. Through back-and-forth conversation, children focus their attention, express their thoughts, and critically reflect on the topic being discussed [3, 28, 36, 68]. However, this kind of conversation is not common: it does not come natural for parents to pause the story, ask questions, and then further discuss with their children [69]. Many parents either assume that children can understand well from simply listening to a story, or they lack the skills or time to incorporate such interactive opportunities [23].

In recent years, the rapid development of artificial intelligence (AI) has made conversational agents (CAs) more capable of simulating natural interpersonal interactions [49]. CAs in the form of smart speakers are prevalent in many homes, and children readily interact with and accept these devices as part of their daily lives. Studies suggest that children respond to CAs socially and treat CAs as companions or guides [46, 60, 62]. Children's social reactions to CAs raise the question: Can CAs serve as suitable language partners for children in joint reading activities, complementing the role of parents or other mentors?

In fact, numerous voice-based apps, termed "Skills" on the Amazon platform or "Actions" on the Google platform, have been developed and made available. Many of these apps target young children and purport to enrich children's learning. Such apps are designed to engage children in a variety of conversations; they can tell stories, play games, recite lessons, or quiz children. However, CAs on the market are not usually designed with a clear theoretical rationale for meeting children's unique learning and communication needs. In addition, little research has been carried out on understanding how children respond to the conversational design features of a CA language partner. When considered together, these two deficiencies point to an unproductive development cycle devoid of research that supports the intended educational goal [27, 45, 66].

In this project, we first discussed key components of an effective reading partner, and then designed a CA reading partner incorporating these components. In particular, we developed a smart-speaker CA narrating a picture book while engaging children in story-related, scaffolded conversations in order to facilitate comprehension and engagement. The CA partner was tailored to children aged 3 to 6 years as children in this age group are not able to read independently [39] but typically have sufficient oral language skills for productive oral communication [8]. We then conducted an observational study of 33 children's individual interactions with the CA and explored how such interactions were influenced by particular CA design features. In our analysis, we approached conversation as an interaction between two parties (i.e., the child and the CA) and focused on the children's responses to the CA in

IDC '20, June 21–24, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7981-6/20/06 ...$15.00.
http://dx.doi.org/10.1145/3392063.3394417

three dimensions that are traditionally identified as revealing engagement levels in conversations [7, 26, 31, 53, 63]. These three dimensions are **language production** that captures the quantity of children's vocalization, **flow maintenance** that details the semantic and temporal appropriateness of children's responses, and **affect** that indicates children's emotional engagement during the conversation.

We seek to answer the following question: *How do children respond to a CA reading partner during conversation, in terms of children's language production, flow maintenance, and affect?* We also note significant developmental differences within children aged 3 to 6 years, and thus further ask: *Do the younger children within this age group (3- to 4-year-olds) respond to the CA reading partner differently than do older children (5- to 6-year-olds)?*

RELATED WORK

Characteristics of an Effective Reading Partner

Children's learning opportunities from joint reading are influenced by the quality of the verbal interactions they have with parents [56]. A parent, as the more experienced language partner, usually guides the conversation by posing questions, commenting on children's responses, and adjusting the dialogue to the child's developmental level [61, 65]. According to Snow [59], parent's conversational guidance should contain the following three components: **questions** that open up the conversation and invite thoughtful responses from the child; **semantically contingent feedback** that continues topics introduced by the child's preceding utterances; and **language scaffolding** that reduces the degrees of freedom in language exchange to lessen the cognitive load needed for the conversation. These three components of conversational guidance together support children and engage them in more cognitively and linguistically beneficial interactions [5].

Parent-Child Conversations in Joint Reading

Parents are found to vary in their use of questions, feedback, and scaffolding during joint story-book reading [42]. Such variations in conversational guidance appear to be associated with how children engage in the conversations with their parents and how much children learn from the reading [22, 48].

First, parents appear to pose different kinds of questions during reading: Some parents tend to utilize open-ended questions, whereas others more frequently ask yes-or-no questions or merely make directive comments (e.g., "Turn the page.") [13]. These different prompting strategies elicit differing responses from children. In general, open-ended questions have been shown to encourage children to engage in deep-level processing and generate more sophisticated responses than other types of questions [41]. Moreover, some studies suggest that parents should ask open-ended questions at different levels of cognitive demand [31]. Easier questions help children construct a basic understanding of story facts and lay the groundwork for harder questions that encourage children to predict what will happen next, relate story elements to personal experience, and make inferences based on what they know.

Secondly, how parents reply to children's responses is also important. Parents' direct and specific feedback helps children clarify their own confusion and increases children's engagement [24, 30]. As such, it is recommended that parents repeat, validate, and elaborate on what the child says [50]. A parent's neutral or vague response (e.g., "Umm", "Ok, I hear you.") may not ease confusion and may even lead children to perceive their parents as inattentive, discouraging the child from further participation.

Thirdly, some parents are also aware of the benefits of using language scaffolding [15, 58]. They actively and constantly adjust the conversations to the developmental level of the child [37]. The aim of scaffolding is to fully engage children in the conversations by easing the obstacles a child may have when responding to parents' prompts. Common scaffolding techniques include using language that matches the child's level of comprehension or providing hints and options that prime the child to maintain the topic [18].

This line of research in traditional reading environments may help inform research on the dialogic interactions between a CA and a child. In particular, it points to how an effective language partner may increase children's learning through engagement in conversations.

Intelligent Systems as Reading Partners

A group of studies, most of which focused on robots, utilized intelligent systems with voice interface to engage children in joint reading activities. For example, Kory and Breazeal [35] developed a storytelling robot for preschool children's oral language development. The study found that children learned the vocabulary words that the robot had introduced in their conversation. Similarly, Conti and colleagues developed a robot that could tell children stories with expressive behaviors and found that children can memorize the story they heard from robots as well as from a human reader [12]. However, these experimental robot systems can only be used for narrowly specific scenarios, thus are rarely adopted by the general public. On the contrary, **conversational agents (CAs)** in a smart speaker form, such as Google Home and Amazon Echo, are already used by many families and children as consumer-oriented voice assistants. Yet little research has been devoted to embracing CAs for early literacy learning purposes.

Child-CA Conversations

A growing body of research has documented positive conversation experiences that children have had in their everyday lives with CAs, mostly general voice assistant tools such as Google Assistant or Amazon Alexa. Druga and colleagues found that preschool-aged children interacted naturally with CAs [16]. Another study [55] revealed that children's conversation with CAs involved a wide-range of topics, including asking smart speakers questions about homework and requesting speakers to play music or even skip a song they did not like [55]. Lovato and colleagues found that children turned to the speaker for information on language, culture, science, math, etc [40]. Some studies also revealed that children perceived CAs as a friendly, trustworthy, and safe language partner [16, 40]. Taken all together, these studies provide important evidence as to the feasibility of harnessing CAs to support children's learning through conversations. Moreover, we expect that CAs that are

specifically designed to engage in guided conversations may better support child-focused educational experiences [64].

Despite the promising future of CAs, research has also found that children sometimes encounter challenges when interacting with CAs. These challenges stem from CAs misinterpreting children's speech or providing responses that are not age-appropriate to children [32]. Sciuto and colleagues reported that parents sometimes observed unsuccessful child-agent interactions when the voice devices failed to understand children's speech [55]. A related line of research has focused on scaffolding strategies that could help remediate communication breakdowns. Scaffolding may come from more capable family members [10]. The interface itself may also be designed to provide this kind of scaffolding. Indeed, in a study on children's conceptualization of intelligent interfaces, children expected the interfaces to be able to recognize their different abilities and adjust the conversation appropriately to the child [67]. However, little research has examined how to design and embed such scaffolding feature in voice products.

In summary, these in-the-wild studies have demonstrated how young children engage in voice interfaces in their everyday lives. What is not yet clear is what aspects of CAs' conversational design features have contributed to children's natural conversation or communication obstacles with CAs.

DEVELOPMENT OF THE CA READING PARTNER

Our CA reading partner, which engages children in story-related conversations, was designed around the three components of conversational guidance discussed in the preceding section. The CA itself contains no visual element but is designed to be used alongside a printed picture book. This combination increases children's print exposure and potentially enhances their engagement and learning.

The CA, deployed in a Google Home Mini device, was built upon Google's Dialogflow open source client library. The CA learns to understand children's responses both from the pre-trained language models already built into this engine as well as training phrases that we provide, which are sample phrases of what children may say as a response to a particular prompt. The CA is able to learn from a small set of training phrases and naturally expand them to many more similar phrases so that children's voice input can be accurately interpreted. Figure 1 displays the general workflow of the child-CA communication. Children were first invited to respond to an open-ended question (**Initial Prompt** hereafter) and received feedback for providing an answer that the CA could interpret. If the response could not be understood by the CA, the CA would ask children a scaffolded follow-up question (**Follow-up Prompt** hereafter) and then would give feedback based on the child's response. If the CA could not understand a child's response to the Follow-up Prompt, the CA would give the child vague, generic feedback that explained the question but did not directly address the child's answer.

Developing Initial Questions

We included a total of ten open-ended questions (i.e., Initial Prompts) related to the story, with varying levels of difficulty. Seven are easier, fact-based questions, with the other three

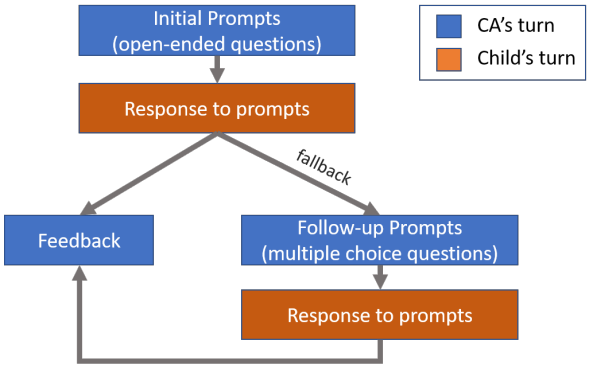

Figure 1. Child-CA Dialogue Flow. Blue textbox represents the CA's turn. Orange textbox represents the child's turn.

being more difficult, inferential questions. Sample fact-based questions include, "*What places did the bears search on that island?*" and "*What did Mama bear do after the bears got home?*" Sample inferential questions include, "*Why do you think the bears were afraid of their mother?*" and "*Why do you think the bears stopped at the island?*" These questions were reviewed by two outside experts on children's literacy education. We also field tested these questions with five children where a human experimenter read the story and asked these questions, and some minor revisions were made after the field testing. Nevertheless, asking children open-ended prompts without a small range of pre-selected response options may make it challenging for the CA to directly follow up with the child's responses. We addressed this challenge through collecting a large number of possible user input and then creating fine-grained categorization of the possible inputs.

Developing Feedback

For each conversational prompt, we predefined categories, or intents, that we wanted the CA to classify the child responses into. Given that children were likely to respond to particular questions in many distinct ways, we included multiple intent categories for anticipated answers. After the CA classified the child's responses into one of the intent categories, differentiated specific feedback was given based on the classification. For example, one question asked, "*What do you think is going to happen with the weather?*" and the correct answer was inclement weather. We created multiple relevant categories related to inclement weather, including rainy, windy, stormy, bad, dark, scary, and cloudy. Any of these responses was considered correct, but the CA provided differential specific feedback to each of them. The feedback first played a rising sound effect and praised children for the correct answer by saying "*Yes, I think so too! It seems that the weather is going to turn bad and a storm is coming*" and then followed with the specific weather the child had mentioned, such as "*Storms are very cloudy,*" "*A storm brings rain,*" or "*Storms sometimes are scary.*" Multiple intents were also created for possible incorrect answers. For example, if a child provided an incorrect answer "the weather's going to be sunny", the CA played a falling sound effect and said "*Umm, I don't think so. It seems*

*that the weather is going to turn bad and a storm is coming. Storms are not **sunny**."*

Developing Scaffolding Mechanism

If a CA failed to categorize children's response as any of the pre-defined intent categories (i.e., categorized the response as "fallback"), the CA triggered a scaffolding mechanism where children were provided with an additional opportunity to answer the question, with the Follow-up Prompt rephrasing the Initial Prompt into a multiple-choice format. This scaffolding mechanism facilitated the conversation flow within the context in cases when children had difficulty participating in the on-topic conversation or the CA did not understand the exact utterance of a child. Using the question "*Why do you think the bears stop at this island?*" as an example, if a child provided answers that the CA could not categorize, the CA gave the child a second chance by asking a multiple-choice question, "*Is it because the bears think they can find a blue seashell there, or is it because the island is a fun place to play?*" When the child's response could not be categorized a second time, the CA provided generic feedback that did not directly address the child's response, but included the correct answer, "*The bears stopped at this island because they think they can find a blue seashell there*," and then the CA continued the story.

Optimizing the Language Model

This language training model was developed and optimized through three months of field testing involving 20 children. We collected data on children's responses in order to modify the intents (e.g., added more intents to encompass other response categories) and included more training phrases to increase the accuracy of intent classification.

METHOD

Participants

Thirty-three children aged 3 to 6 years (none of whom had participated in the prior field testing) were recruited from childcare centers in a research university community (see Table 1 for demographic information). The mean age of the participants was 4.5 years, and 19 of them (58%) were girls. Twenty-three children (70%) spoke only English at home. According to parent reports, 30% of the children had never interacted with a CA, 27% had done so monthly, 12% had done so weekly, and 30% had daily interaction with a CA. We divided these children into two groups based on their age. The younger group (3- and 4-year-olds) consisted of 16 children with a mean age of 3.8 years, and the older group (5- and 6-year-olds) consisted of 17 children with a mean age of 5.2 years.

Study Procedure

Children met individually with a trained researcher in a designated quiet area at their school. Children were looking at a hard copy of the story book and were encouraged to take responsibility for turning pages when the narration of a page was finished. The researcher sat beside the children but interfered only when/if technical issues interrupted the reading. In the case that children asked questions or initiated comments, the experimenter simply addressed the question or replied "okay,"

	Full sample	Younger group	Older group
Age in years	4.5 (0.8)	3.8(0.4)	5.2(0.4)
Female	58%	59%	56%
English only	70%	76%	62%
CA use			
Daily	30%	13%	47%
Weekly	12%	25%	0%
Monthly	27%	31%	24%
Never	30%	31%	29%
N	33	16	17

Table 1. Participant Information (Standard deviation in parentheses)

but avoided elaborating or extending the conversation. The reading session lasted approximately 15 minutes per child. The reading sessions were video-taped for future analysis. See Figure 2 for the setup of our study session.

Figure 2. Study Session Setup

Coding Framework

The development of a coding framework was guided by prior research that collectively emphasizes the verbal and non-verbal aspects of conversations [7, 26, 31, 53, 63]. The resultant coding framework consists of three dimensions, namely **language production**, **flow maintenance**, and **affect**. These three dimensions are believed to work in conjunction to signify the extent to which a speaker is engaged in meaningful and productive communication [54]. Prior research has included some or all of these dimensions to analyze children's communication with voice interfaces such as robots and other CAs [4, 51, 57]. Below, we will detail how each of the dimensions was informed by prior work, and how the coding was operationalized.

The first coding dimension was **language production**, which captures a child's production of verbal responses to the CA's prompts. As suggested by Brennan, active verbal responses are generally prerequisite for fluid conversation [7]. Thus, we coded whether a child verbally responded to the prompt. In addition, studies suggested that the word length of a response is one of the most important indicators of conversation engagement [31]. We therefore also coded the total number of words in each of the children's responses.

The second dimension was **flow maintenance**, which focuses on the semantic flow and temporal flow of the conversation.

According to Wanska, to maintain the semantic flow, a speaker needs to respond to his partner in a topically relevant way [63]. This indicates that a speaker is monitoring the content of his partner's statement and making an effort to link his own response to his partner's [63]. According to Heldner, to maintain temporal flow, a speaker's timing of responses should follow a turn-taking pattern without any overlapping speech or any silence between turns (i.e., no-overlap-no-gap) [26]. We therefore coded the topic relevance and timing of children's responses. For example, in response to the question "*What shape is the island the bears need to look for,*" a relevant answer would be a shape (e.g., triangle) or some recognizable object (e.g., hat, crown). Responses that were not considered relevant included those that did not reference some shape or did not stay within the broader theme of the story. The timing of response included two codes: whether a child responded too quickly (before the CA came to a full stop) and whether a child responded with a substantial delay (after approximately 2 seconds when CA believed the child was giving up their turn).

The third dimension was **affect**, which focuses on children's varied emotional responses throughout the conversation. According to Ruusuvuori, a speaker's emotional engagement in a conversation (both when speaking and being spoken to) can be revealed through several affective markers, including laugh tokens, lexical choices, tones of voice, and facial expressions [53]. We examined these markers during children's responses to the CA and when they were listening to the CA's feedback and then categorized children's affective state as belonging to one of four categories: positive, negative, confused, and neutral. These four categories are believed to be salient affective states as children engage in learning processes [25]. Positive emotion was identified through the presence of any of the following: positive facial expressions, positive body cues, presence of laugh, rising tone, or positive connotations. Negative emotion was identified through the presence of any of the following: negative facial expressions, negative body cues, falling tone, or negative connotations [2]. Confusion was identified through any facial expressions (e.g., eyebrow raise-arched, side mouth stretch) or verbal expressions (e.g., "Umm?" "Why?") that indicated confusion [52]. Neutral emotion was coded when no significant signs of emotion were present [38].

Coding Procedure

Our primary data sources were the video-taped interaction sessions and their transcriptions. The unit of analysis is a child's response to a single prompt. If a child successfully answered an Initial Prompt, they would not receive a Follow-up Prompt for that same question. In total, we analyzed 330 responses to Initial Prompts and 205 responses to Follow-up Prompts, thus resulting in a total of 535 coding fragments. For each coding fragment, we coded the three dimensions of communication and included detailed notes for each dimension. This process generated both quantitative and qualitative coding data, enabling statistical analyses accompanied by contextual evidence.

We established the reliability of the coding using two coders who were informed with the overall objective of the study to examine children's engagement with a CA reading partner. Coder A coded and took notes on all of the videos, while Coder B coded a subset of the videos (30%) to establish the inter rater reliability. Coders met once every week to compare codes and discuss any discrepancies in coding. The inter-rater reliability (Cohen's Kappa for categorical codes and Interclass correlation for numeric codes) between Coder A and Coder B for each item is between 0.88 and 1. To establish reliability of the qualitative coding, Coder B reviewed the notes initially taken by Coder A and discussed any disagreements or necessary clarifications. This process was repeated until both coders agreed that the notes accurately reflected the actual interactions.

RESULTS

In this section, we first detail the CA's performance in order to demonstrate the CA's accuracy as a language partner. We then answer our first research question by presenting statistics from the quantitative coding along with descriptive notes contextualizing the statistics. In addition, we answer our second research question using an ANOVA analysis for numeric coding data (i.e., response length in words) and Chi-square analyses for the rest of the coding items with categorical data to determine whether a significant difference exists between the younger children and the older children along the three coding dimensions.

CA's Performance

The performance of the CA was determined by how successfully the CA could categorize children's responses into pre-defined intent categories. There were three possible outcomes: accurate categorization, inaccurate categorization, or categorization failed. "Accurate categorization" indicates that the CA was able to categorize a child's response to a pre-defined intent, and this categorization was accurate. "Inaccurate categorization" indicates that the CA was also able to categorize a response, but this categorization was inaccurate. "Categorization failed" indicates that the CA was not able to categorize a child's response as any of the pre-defined intent categories.

As displayed in Table 2, the majority of the responses to Initial and Follow-up Prompts were accurately categorized by the CA, with 76.7% and 83.7% accuracy, respectively. Inaccurate categorization occurred very rarely for Initial or Follow-up Prompts, with 0.3% and 0.7% inaccuracy. All instances of inaccurate categorization were due to the CA's inaccurate speech-to-text translation. Another 22.7% of responses to Initial Prompts and 15.6% of responses to Follow-up Prompts were identified as "categorization failed." There were three reasons for categorization failure. First, children's verbal responses were absent or incomplete. For example, children nodded their head to indicate "yes," shook their head for "no," or shrugged their shoulders for "I don't know." Children sometimes provided a verbal response that could only be understood when combined with non-verbal expressions. For example, saying "This one," and pointing to the picture at the same time. Second, a child's response was not anticipated. For example, a child answered "*dinosaur*" to the question "What shape is

	Accurate categorization		Inaccurate categorization		Categorization failed	
	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts
Full sample	76.7%	83.7%	0.3%	0.7%	22.7%	15.6%
Younger	70.3%	80.2%	1.6%	1.6%	28.1%	18.2%
Older	80.1%	86.5%	1.1%	1.0%	18.8%	12.5%
Age difference	Initial Prompts: $\chi^2(2) = 9.49$, $p < 0.001$; Follow-up Prompts: $\chi^2(2) = 6.68$, $p < 0.05$					

Table 2. CA Performance in Intent Categorization

the island the bears need to look for?", with dinosaur outside of the overall theme of the story and only brought up by this single child. Not surprisingly, Follow-up Prompts resulted in a higher rate of intent detection than Initial Prompts, largely due to the more restricted questions that eliminated the likelihood of a child providing unanticipated answers. Third, the voice response was translated incorrectly to text. One example for this case was the CA mis-registering a child's correct answer of "*shell*" to "sound," thus leading to an out-of-context response.

The CA appeared to perform better with the older group of children. Young children's utterances being categorized with a lower success rate may primarily be due to young children's less articulate pronunciation and higher likelihood of providing unanticipated answers.

Language Production

Presence of verbal expressions

Children actively responded to the CA with verbal expressions: they verbally responded to over 85% of the CA's prompts (see Table 3). The response rate for Follow-up Prompts (89.3%) was higher than that for Initial Prompts (86.2%), probably due to the scaffolded nature of the Follow-up Prompts. We also found that older children were more likely to verbally respond to the prompts.

When children did not respond verbally to a prompt, they almost always instead relied on non-verbal expressions. Non-verbal responses were quite common when children did not know the answer (e.g., shrugging, shaking head). Children also sometimes gestured to convey information (i.e., pointing to an image in the book). Since the CA was not able to understand such responses, they triggered the CA's programmed scaffolding mechanism. Only a few of children's failures to respond verbally were due to the child's disengagement or intentional avoidance. For example, one child became distracted and looked at the ceiling, missing the question altogether. Another child appeared to realize that he would receive a multiple-choice question if he did not answer the Initial Prompts. After attempting two questions, he stopped responding to any of the Initial Prompts and instead waited for the CA to give him scaffolded questions, all of which he answered correctly.

Response length

The average length of responses to Initial Prompts was 4 words, and the average for Follow-up Prompts was 2 words (see Table 3). Initial Prompts generally solicited longer responses that tended to be complete sentences or phrases. For example,

	Verbal expressions		Response length	
	Intitial Prompts	Follow-up Prompts	Intitial Prompts	Follow-up Prompts
Full	86.2%	89.3%	4.1	2.4
Younger	80.0%	82.3%	4.2	2.4
Older	91.1%	97.7%	4.1	2.4
Age difference	$\chi^2(1) = 7.53$ $p < 0.01$	$\chi^2(1) = 10.28$ $p < 0.01$	$F(1, 281) = 0.08$ $p = 0.78$	$F(1, 180) = 0.07$ $p = 0.80$

Table 3. Language Production

when asked the question "*What do the bears ride on to travel across the sea?*" most children's responses included a verb or a preposition; rather than simply replying "sailboat," children replied "*ride on a sailboat*" or "*on a sailboat*". Follow-up Prompts tended to result in children giving shorter responses, and many children tended to give single-word responses. For example, when the question on the bears' transportation was rephrased as "*Do the bears ride on a sailboat or do they swim across the sea?*" children tended to respond by simply saying "*sailboat*" or "*swim*," rather than "on a sailboat" or "swim across the sea." Among both Initial and Follow-up Prompts, older children and younger children generated responses of comparable length.

Flow Maintenance

Topic relevance

In our observation, children were able to directly answer the majority of questions (see Table 4). Children were much more likely to generate relevant responses to Follow-up Prompts (89.6%) than to Initial Prompts (76.7%). The increase of topic relevance among Follow-up Prompts suggests that our scaffolding mechanisms worked well to support children's communication. For example, when asked "*Where did the bears find the blue seashell?*" one child provided an answer that was topically irrelevant to the story (an answer about a dinosaur). The CA then asked, "*Did they find it on an island or did they find it under the sea?*" The child responded appropriately, and the conversation flow was maintained.

When looking at the topic relevance by age group, we found that, unsurprisingly, older children were better able to directly answer the Initial Prompts, which were open-ended questions. However, with the scaffolding prompts, the age difference in topic-relevance became non-significant. The topic relevance of younger children's responses increased by 23 percent (from 65% to 88%). For older children, scaffolded prompts only slightly increased the already high proportion of relevant responses by 5 percent (from 87% to 92%).

	Response relevant to the qustion	
	Initial Prompts	Follow-up Prompts
Full sample	76.7%	89.6%
Younger	64.9%	87.8%
Older	86.5%	91.7%
Age difference	$\chi^2(1) = 17.12$ $p < 0.001$	$\chi^2(1) = 0.38$ $p = 0.54$

Table 4. Topic Relevance in Children's Responses

	Gaps and pauses		Rushed responses	
	Intitial Prompts	Follow-up Prompts	Intitial Prompts	Follow-up Prompts
Full	21.6%	11.5%	8.4%	24.2%
Younger	29.7%	13.3%	5.5%	31.6%
Older	14.8%	9.5%	11.0%	15.5%
Age difference	$\chi^2(1)$ = 8.28 p <0.01	$\chi^2(1)$ = 0.31 p =0.58	$\chi^2(1$ = 5.88 p <0.05	$\chi^2(1)$ = 6.15 p <0.05

Table 5. Timing of Children's Responses

Timing of response

Gaps and pauses. Children needed time to organize their thoughts when answering a question, and this resulted in children sometimes not initiating a response within a short period of time or beginning a response but pausing to think before completing it. Because CAs must rely solely on the duration of gaps and pauses to determine when a child's turn is either abandoned or completed, CAs would simply miss the utterances spoken after the gaps or pauses. Overall, gaps and pauses were observed among responses to 21.6% of Initial Prompts, and this number was 11.5% for Follow-up Prompts (see Table 5). The occurrence of gaps and pauses was higher among the Initial Prompts than among the Follow-up Prompts with multiple choices, probbably becuase Initial Prompts were generally more challenging for children. As expected, younger children had significantly more gaps and pauses in responding to the open-ended Initial Prompts than older children. Younger children were observed to have gaps and pauses among 29.7% of Initial Prompts while older children had gaps and pauses among 14.8% of Initial Prompts. This difference was probably due to younger children's less advanced reading comprehension. With the Follow-up Prompt, the frequency of gaps and pauses became more similar between younger and older children (13.3% for the younger group and 9.5% for the older group).

Rushed responses. Children sometimes responded to a question too quickly, before the CA fully completed its turn. This kind of rushed response was observed among 8% of Initial Prompts but among 24% of Follow-up Prompts (see Table 5). The commonality of rushed responses to Follow-up Prompts may be due to the questions' lower difficulty or wording that contained the original open-ended question and a set of possible answers in question form. For example, a Follow-up Prompt asked "*What did the bears break? Did they break a blue seashell or did they break a honey jar?*" and one child responded "blue seashell" immediately after the first part of the question, given that the questioning tone invited the child to respond. The CA did not register the child's answer and continued to complete the scaffolded question. The child then replied "yes" immediately after the CA mentioned the blue seashell, but the CA also did not hear this response. Younger children appeared to have more trouble determining when the CA had completed its question and could register the child's response for the Follow-up Prompts that contained questioning tone in the middle of them. Specifically, younger children's rushed responses occured among 31.6% of Follow-up Prompts while older children's only occured among 15.5% of Follow-up Prompts.

Affect

Affect while responding

As shown in Table 6, most of the time, children expressed no emotion at all when responding to the CA. Children showed neutral affect among 74.7% of Initial Prompts and 85.6% of Follow-up Prompts. These neutral affective states were categorized by a lack of facial expressions and body gestures, a flat tone of voice, and matter-of-fact word choices. This lack of affect may be due to the design of our CA's prompts: these prompts primarily asked about specific content in the story, thus leaving little room for children's emotional expression.

Nevertheless, positive emotional responses were not uncommon, which was observed among 25.3% of Initial Prompts and 14.4% of Follow-up Prompts. Children sometimes exhibited pride in having given what they were confident was a correct response. For example, when a child was recalling a set of places where the bears had searched on an island, she nodded her head and clapped her hands with every additional place she recalled. Another child smiled and nodded to herself once she finished answering, as if she was satisfied with her own answer. Positive emotions were also observed during some "distracted moments" when children's comments expressed excitement over something tangentially related to a detail in the book but not directly related to the prompt. For example, when asked, "Why do you think the bears stopped at this island?" a child commented excitedly while pointing at a Ferris wheel on that page, "*This island has a Ferris wheel! I saw a Ferris wheel! I have been on a Ferris wheel.*" Although these conversational moments may, on the one hand, indicate a child's disengagement with the story, the child sharing a related personal experience with the CA may, on the other hand, suggest that the CA's prompt did provide children an opportunity to express their enthusiasm. We observed less neutrality and more positivity among Initial Prompts, which may be due to that Initial Prompts were less restricted, thus allowing children to include their feelings and attitudes.

Young children appeared to be more likely to show positive affect as they responded to the Initial Prompts than older children. When answering Initial Prompts, younger children had positive affect among 31.1% of their responses while older children showed positive emotion among 21.8% of their responses. This may be due to that younger children would have a greater sense of accomplishment for answering a question that seemed to be challenging for them. It may also be related to younger children's tendency to insert information that interests them in the conversation. As expected, when it comes to the Follow-up Prompts that were generally easier

and more restricted, the age difference between positive (or neutral) affect diminished.

We did not observe any negative affect or confusion during children's responses to the CA's prompts, either to Initial or Follow-up Prompts. However, one caveat to interpret the absence of negative affect and confusion was that we only examined the affect while children were actually responding to the CA. It is possible that some children who were not responding were confused by a particular question or did not feel like answering the question.

	Postive affect		Neutral affect	
	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts
Full	25.3%	14.4%	74.7%	85.6%
Younger	31.1%	15.3%	68.9%	84.7%
Older	21.8%	13.4%	78.2%	86.6%
Age difference	Initial Prompts: $\chi^2(3) = 8.05$, $p < 0.05$; Follow-up Prompts: $\chi^2(3) = 2.42$, $p < 0.49$			

Table 6. Affect in Children's Responses to the CA (Negative affect or confusion was not observed.)

Affective reaction to feedback

We analyzed children's reactions to CA feedback resulting from two types of CA intent categorization (i.e., accurate and failed categorization) described in the "CA's Performance" section above. The CA's accurate categorization of a child's response resulted in feedback that was appropriate and specific, while the CA's failure to categorize a child's response resulted in feedback that was vague and generic. Instances of inaccurate categorization were not examined here since they occurred very rarely in our study.

We first looked at children's reactions to specific feedback to their correct and incorrect responses (see Table 7 for correct responses and Table 8 for incorrect responses). When children received specific feedback that indicated their answer was correct, they typically expressed positive emotion (e.g., laughing, cheering, clapping, dancing, saying "Yay!"). This positive emotion was observed among 75.2% of Initial Prompts and 73.4% of Follow-up Prompts. However, children's affect was less impacted when they received specific feedback for an incorrect response, as neutral affect was observed among 82.4% of Initial Prompts and 83.4% of Follow-up Prompts. Negative emotion only occured among 15 percent of feedback that indicated a child answered a prompt incorrectly (15.9% of Initial Prompts and 15.3% of Follow-up Prompts). In a few cases, children also showed confusion after receiving feedback for their incorrect answers (1.7% of Initial Prompts and 1.3% for Follow-up Prompts).

Interestingly, compared to older children, younger children's emotion was more likely to be enhanced when they received positive feedback from the CA for their correct answer. However, younger and older children reacted in a similar way when they received feedback indicating their answer was incorrect.

We then looked at children's reactions to the CA's vague and generic feedback (i.e., feedback that did not address children's response at all and that instead simply told them the answer

	Postive affect		Neutral affect	
	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts
Full	75.2%	73.4%	24.8%	26.6%
Younger	82.3%	81.8%	17.7%	18.2%
Older	67.9%	66.5%	32.1%	33.5%
Age difference	Initial Prompts: $\chi^2(2) = 11.65$, $p < 0.01$; Follow-up Prompts: $\chi^2(2) = 11.52$, $p < 0.01$			

Table 7. Children's Affective Reactions to Receiving Positive Feedback from the CA (Negative affect or confusion was not observed.)

to a question) resulting from failed categorizations (see Table 9). We rarely observed any emotional changes when children received this kind of feedback (i.e., neutral affect, 94% for Initial Prompts and 90% for Follow-up Prompts), regardless of whether children actually answered the question correctly or incorrectly. Occasionally, we observed confusion among children (6.5% for Initial Prompts and 9.9% for Follow-up Prompts), especially those who appeared confident about their response. For example, one child correctly answered that the bears were "*hugging each other*," but the CA mistranslated "hugging" as "hacking." The CA then replied, "*The bears were hugging each other to make themselves feel better*," and the child commented, "*Why? Why didn't it say I'm right?*" This pattern was similar across age groups.

DISCUSSION

In this paper, we describe both the design of a CA that can engage children in joint reading and a user study of how children interacted with this CA. Based on our findings, we now turn to a discussion of how automated conversational interfaces could play the role of language partners for young children and how to best design such interfaces.

Leverage CA's Natural Language Ability

Our study suggests that, if designed properly, a CA can perform satisfactorily as a joint reading partner for children. In our observation, children actively participated in conversation with the CA and frequently generated on-topic responses. Children were generally able to respond to the CA within the proper time frame. Children also showed positive affect while speaking to the CA or listening to the CA's feedback. We attribute children's engagement to how we designed the CA to invite children's verbal engagement and respond to children.

Inviting children's responses

Our CA used a combination of open-ended questions (i.e., Initial Prompts) and multiple-choice questions (i.e., Follow-up Prompts), which worked together to support children's interactions with the CA. The initial open-ended questions were designed to encourage children's free expression of their thoughts related to the story [1,21]. Indeed, we found that children articulated their understandings more fully and in a more grammatically complex way when responding to open-ended questions. Interestingly, we also observed that children's responses to the open-ended questions commonly involved some personal connection the child had to the topic. Although these responses sometimes did not directly answer the question, they were almost always accompanied by children's increased affective engagement, and we believe this engagement makes the

	Postive affect		Neutral affect		Negative affect		Confusion	
	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts
Full sample	0.0%	0.0%	82.4%	83.4%	15.9%	15.3%	1.7%	1.3%
Younger	1.4%	0.0%	80.2%	79.5%	16.8%	17.9%	3.0%	2.6%
Older	0.0%	0.0%	85.6%	86.5%	14.4%	13.5%	0.0%	0.0%
Age difference	Initial Prompts: $\chi^2(3) = 7.16$, $p = 0.07$; Follow-up Prompts: $\chi^2(3) = 7.73$, $p = 0.05$							

Table 8. Children's Affective Reactions to Receiving Negative Feedback from the CA

	Neutral affect		Confusion	
	Initial Prompts	Follow-up Prompts	Initial Prompts	Follow-up Prompts
Full	93.5%	90.1%	6.5%	9.9%
Younger	94.2%	92.2%	5.8%	7.8%
Older	93.0%	88.9%	7.0%	11.1%
Age difference	Initial Prompts: $\chi^2(3) = 0.72$, $p = 0.87$; Follow-up Prompts: $\chi^2(3) = 6.67$, $p = 0.08$			

Table 9. Children's Affective Reactions to Receiving Vague, Neutral Feedback from the CA (Positive or negative affect was not observed.)

joint-reading experience more relatable for the children [34]. However, we also note that this excitement may not be directly linked to the topics being discussed [17], but may, in general, reflect children's enjoyment of having open-ended conversation with a digital learning partner.

Despite the benefits of using open-ended questions, this approach is not without costs. While open-ended questions can stimulate thinking, responding to them may also require more cognitive resources, sometimes exceeding a child's capacity. Additionally, the freedom in formulating a response may lead children astray from the topic at hand. As such, broadly open-ended questions may lead children to either not answer the questions or answer them with lower topical relevance. Moreover, unlike humans, CAs' response quality may degrade substantially when discussing unrestricted topics; a CA's performance is largely reliant on the designers' ability to predict children's responses, and open-ended questions can result in greater variation and unpredictability. Indeed, in our observations, children's responses to unrestricted Initial Prompts contained more unpredicted content, resulting in the CA's decreased accuracy rate when categorizing responses for Initial Prompts.

In order to balance the drawbacks of open-ended questions, we introduced more restricted multiple-choice questions as Follow-up Prompts. These types of questions can help ease the potential cognitive obstacles and redirect children's attention to the story content. They also have benefits for the CA's performance, since they keep children's likely responses within what the CA is capable of categorizing, thereby preventing possible conversation breakdown. This strategy of including restricted questions as a way to recover from impediments in the preceding conversational turns resonates with the notion of adaptability commonly suggested in conventional educational pedagogy [19, 20]. However, adaptability traditionally involves the more experienced language partner tailoring the conversation to meet the child's ability; in CA-child communication, including restricted questions also adjusts the child's response to accommodate the CA's ability in understanding.

Responding to children

The CA was intended to provide specific feedback to children's responses. Specific feedback acknowledges what a child has said and then moves the conversation forward based on the child's input. The CA's capability of providing specific feedback depended on how accurately the CA could interpret children's responses and map those to intent categories. As discussed before, we attempted to achieve this goal through creating fine-grained categorization of children's possible language input so that the CA could provide precise feedback based on children's utterance. In our observations, we found that specific feedback kept children emotionally engaged, while vague feedback resulting from failed categorization generally did not facilitate engagement.

While the value of specific feedback has been emphasized in adult-child communication, we think that specific feedback is especially important in human-machine interaction. As voice interfaces cannot provide other social cues (e.g., eye-gaze, facial expression) through which children can infer that their responses have been correctly registered, feedback plays a role to reassure children that the agent understands their input. This viewpoint has been confirmed in multiple HCI papers that examined how CAs should respond to users to demonstrate CAs' good listenership and understanding [6, 11].

Interaction Challenges Inherent to Voice Interfaces

While CAs' natural language abilities make it possible to simulate a human reading partner, there exists some interaction challenges inherent to voice interfaces. However, it is still possible to improve children's conversation experiences through optimizing the conversational design.

First, CAs in the form of a smart speaker do not have the capability to capture and interpret non-verbal expressions. Children were not fully aware of this inability, and the children thus tended to use both verbal and non-verbal communication when responding to the CA. On the one hand, children's engagement in multi-channel expressions is similar with what has been identified in child-parent conversations during story reading, suggesting that the CA elicited children's natural responses. On the other hand, children's use of non-verbal expressions may lead to the CA's failure to register their responses, and the conversation flow can suffer. One possible way to eliminate this issue would be having the CA emphasize its ability to listen and prime the children to respond verbally. For example, the CA may explicitly include the words "tell" or "say"

within questions, such as "*Please tell me whether the bears broke a blue seashell or a honey jar.*" The CAs' reliance on speech may actually be positive, since this reliance—once understood by children—encourages children to practice verbal communication vital for their language development.

Second, research on human-to-human communication suggests that the most common turn-taking pattern is one-party-at-a-time and that speaker changes typically occur without any silence in between and without any overlapping speech (i.e., no-gap-no-overlap). Although infrequent, human-to-human interactions can sometimes involve some overlap and gaps, but CAs are not currently capable of allowing this flexibility. In communicating with CAs, the turn-taking schema must be followed rigidly. As such, awareness of conversational timing is especially important for interacting with CAs. Unsurprisingly, we observed that children sometimes did not follow the "no-gap-no-overlap" rule. This violation may be, in part, due to children's unfamiliarity of CA's restrictions, and may also in part arise from the design of question prompts. For prompts that elicit responses from the children prior to the prompt's completion, we suggest avoiding any questioning tone if there is no intention to immediately invite a child's response. For example, our CA originally asked, "*Did they break a blue seashell (?) or did they break a honey jar?*" This could be rephrased as "*The bears broke something: a blue seashell or a honey jar. Which one did they break?*" As for gaps, the question prompts that are particularly difficult led to gaps before or pauses during a child's response. We suggest that developers ensure the difficulty level of all prompts is such that children can maintain relatively constant responses. This suggestion is consistent with the literature on parent-child interaction that recommends parents ask questions within a child's "zone of proximal development" [33, 47].

Age Differences in Engagement

Younger children's communication patterns differed from those of older children. The age difference in language production and flow maintenance was expected, given young children's less developed language skills and reading ability. This is in line with many studies in adult-child joint-reading that suggest younger children are less able to generate verbal responses that are topically relevant to the conversation [43]. In addition, younger children appeared to face more challenges when interacting with the CA, mostly due to their lack of awareness of the unique nature of artificial voice interfaces [10, 66]. Despite younger children's obstacles in participating in the conversation, the younger children seemed more interested in the CA reading partner than did the older children. One possible explanation is younger children's increased tendency to perceive CAs as human-like social beings, whereas older children tend to view CAs simply as machines [29]. Younger children may thus approach their interactions with the CA with greater enthusiasm.

We also observed that the age difference among responses to Follow-up Prompts were generally smaller than those to Initial Prompts. In particular, Follow-up Prompts increased younger children's relevant responses by 23 percent as compared to the 6 percent increase among older children. This further suggests the CA's scaffolding mechanism gears towards supporting younger children who are more in need of it.

Limitation and Future Work

First, while our study has proven feasibility of using a CA to simulate parent-child interaction during joint-reading activities, a follow-up experimental study should be conducted to compare children's engagement with the CA partner compared to their engagement with a human partner, in order to examine the effectiveness of the CA partner. Second, as it is unclear how the effectiveness of unimodal CAs (speech-only) compares to robots capable of carrying out multimodel interactions, future studies may explore whether CAs and robots with the same conversation design result in children's different patterns of engagement.

CONCLUSION

In this project, we developed and tested a CA reading partner that can read storybooks to children and actively engage them in conversations relevant to the story. The main goal of this CA is to simulate a conversation-rich, interactive reading experience similar to that of an adult partner guiding meaningful language exchange. Through examining children's conversations with the CA, we identified patterns in how the design of the CA's questions, feedback, and scaffolding influences children's responses. Overall, children responded to the CA's conversational guidance in many ways consistent with the literature on parent-child communication. Children's natural communication with CAs is encouraging and may indicate that CAs have, from the child's perspective, effectively simulated a dialogic partner, while children's assumptions of CAs' capabilities lead to some interaction challenges. As such, it is important to leverage what CAs have in common with human partners (i.e., the natural language ability) but also recognize CAs' own unique properties as artificially intelligent interlocutors. Our study suggests that, rather than attempting to develop CAs as an exact replicate of human conversational partners, we should treat child-agent interaction as a new genre of conversation and calibrate CAs based on children's actual communicative practices and needs.

ACKNOWLEDGEMENT

The authors would like to thank the delicated research assistants in the Digital Learning Lab at the University of California, Irvine, for assisting with data collection and data coding. We also thank the children, families, and childcare programs that participated in this study.

SELECTION AND PARTICIPATION OF CHILDREN

This study was reviewed and approved by the Institutional Review Board at the University of California, Irvine. Children were recruited from local childcare programs during school pick-up hours. A researcher explained the study procedure to the parents/caregivers and answered any questions. If interested, parents/caregivers completed a brief demographic survey and signed. During the study session, children were orally informed that they could discontinue the study anytime if they wished to do so. All the study sessions were carried out under supervision of a teacher in the children's schools.

REFERENCES

[1] Mary DeKonty Applegate, Kathleen Benson Quinn, and Anthony J Applegate. 2002. Levels of thinking required by comprehension questions in informal reading inventories. *The Reading Teacher* 56, 2 (2002), 174–180.

[2] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.

[3] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, and others. 2013. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction* 1, 2 (2013), 33–53.

[4] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 243.

[5] Barbara A Bradley and David Reinking. 2011. A formative experiment to enhance teacher-child language interactions in a preschool classroom. *Journal of Early Childhood Literacy* 11, 3 (2011), 362–401.

[6] Stacy M Branham and Antony Rishin Mukkath Roy. 2019. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 446–458.

[7] Susan E Brennan and others. 2005. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (2005), 95–129.

[8] Belinda Buckley. 2012. *Children's communication skills: from birth to five years*. Routledge.

[9] Nian-Shing Chen, Daniel Chia-En Teng, Cheng-Han Lee, and others. 2011. Augmenting paper-based reading activity with direct access to digital materials and scaffolded questioning. *Computers & Education* 57, 2 (2011), 1705–1715.

[10] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why doesn't it work?: voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 337–348.

[11] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and others. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[12] Daniela Conti, C Cirasa, Santo Di Nuovo, and A Di Nuovo. 2019. robot, tell me a tale!: A social robot as tool for teachers in kindergarten. *Interaction Studies* 20, 2 (2019), 1–16.

[13] Catherine Crain-Thoreson, Michael P Dahlin, and Terris A Powell. 2001. Parent-child interaction in three conversational contexts: Variations in style and strategy. *New directions for child and adolescent development* 2001, 92 (2001), 23–38.

[14] David K Dickinson, Julie A Griffith, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2012. How reading books fosters language development around the world. *Child development research* 2012 (2012).

[15] Susan E Dieterich, Mike A Assel, Paul Swank, Karen E Smith, and Susan H Landry. 2006. The impact of early maternal verbal scaffolding and child language abilities on later decoding and reading comprehension skills. *Journal of School Psychology* 43, 6 (2006), 481–494.

[16] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. Hey Google is it OK if I eat you?: Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*. ACM, 595–600.

[17] Roxanne A Etta and Heather L Kirkorian. 2019. Children's learning from interactive ebooks: Simple irrelevant features are not necessarily worse than relevant ones. *Frontiers in psychology* 9 (2019), 2733.

[18] Mary Ann Evans, Shelley Moretti, Deborah Shaw, and Maureen Fox. 2003. Parent scaffolding in children's oral reading. *Early Education & Development* 14, 3 (2003), 363–388.

[19] Jay Fogleman, Katherine L McNeill, and Joseph Krajcik. 2011. Examining the effect of teachers' adaptations of a middle school science inquiry-oriented curriculum unit on student learning. *Journal of Research in Science Teaching* 48, 2 (2011), 149–169.

[20] Lynn S Fuchs, Douglas Fuchs, and Norris Bishop. 1992. Instructional adaptation for students at risk. *The Journal of Educational Research* 86, 2 (1992), 70–84.

[21] Jamie Gazella and Ida J Stockman. 2003. Children's Story Retelling Under Different Modality and Task Conditions. *American Journal of Speech-Language Pathology* (2003).

[22] Roberta Michnick Golinkoff, Dilara Deniz Can, Melanie Soderstrom, and Kathy Hirsh-Pasek. 2015. (Baby) talk to me: the social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science* 24, 5 (2015), 339–344.

[23] Roberta Michnick Golinkoff, Erika Hoff, Meredith L Rowe, Catherine S Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language Matters: Denying the Existence of the 30-Million-Word Gap Has Serious Consequences. *Child development* 90, 3 (2019), 985–992.

[24] Eileen Haebig, Andrea McDuffie, and Susan Ellis Weismer. 2013. Brief report: Parent verbal responsiveness and language development in toddlers on the autism spectrum. *Journal of Autism and Developmental Disorders* 43, 9 (2013), 2218–2227.

[25] Simone E Halliday, Susan D Calkins, and Esther M Leerkes. 2018. Measuring preschool learning engagement in the laboratory. *Journal of experimental child psychology* 167 (2018), 93–116.

[26] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 4 (2010), 555–568.

[27] Kathy Hirsh-Pasek, Jennifer M Zosh, Roberta Michnick Golinkoff, James H Gray, Michael B Robb, and Jordy Kaufman. 2015. Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest* 16, 1 (2015), 3–34.

[28] John S Hutton, Kieran Phelan, Tzipi Horowitz-Kraus, Jonathan Dudley, Mekibib Altaye, Tom DeWitt, and Scott K Holland. 2017. Shared reading quality and brain activation during story listening in preschool-age children. *The Journal of pediatrics* 191 (2017), 204–211.

[29] Jennifer L Jipson and Susan A Gelman. 2007. Robots and rodents: Children's inferences about living and nonliving kinds. *Child development* 78, 6 (2007), 1675–1688.

[30] Ann P Kaiser, Mary Louise Hemmeter, Michaelene M Ostrosky, Rebecca Fischer, Paul Yoder, and Maureen Keefer. 1996. The effects of teaching parents to use responsive interaction strategies. *Topics in early childhood special education* 16, 3 (1996), 375–406.

[31] Jennifer Yusun Kang, Young-Suk Kim, and Barbara Alexander Pan. 2009. Five-year-olds' book talk and story retelling: Contributions of mother—child joint bookreading. *First Language* 29, 3 (2009), 243–265.

[32] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 82–90.

[33] Hengameh Kermani and Mary E Brenner. 2000. Maternal scaffolding in the child's zone of proximal development across tasks: Cross-cultural perspectives. *Journal of Research in Childhood Education* 15, 1 (2000), 30–52.

[34] Ann Ketch. 2005. Conversation: The comprehension connection. *The Reading Teacher* 59, 1 (2005), 8–13.

[35] Jacqueline Kory and Cynthia Breazeal. 2014. Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, 643–648.

[36] Huseyin Kotaman. 2013. Impacts of dialogical storybook reading on young children's reading attitudes and vocabulary development. *Reading Improvement* 50, 4 (2013), 199–204.

[37] Susan H Landry, Karen E Smith, Paul R Swank, Tricia Zucker, April D Crawford, and Emily F Solari. 2012. The effects of a responsive parenting intervention on parent–child interactions during shared book reading. *Developmental psychology* 48, 4 (2012), 969.

[38] Jukka M Leppänen and Jari K Hietanen. 2004. Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological research* 69, 1-2 (2004), 22–29.

[39] Christopher J Lonigan, Stephen R Burgess, and Jason L Anthony. 2000. Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Developmental psychology* 36, 5 (2000), 596.

[40] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, Do Unicorns Exist?: Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, 301–313.

[41] JOYCE H McNEILL and Susan A Fowler. 1999. Let's talk: Encouraging mother-child conversations during story reading. *Journal of Early Intervention* 22, 1 (1999), 51–69.

[42] Gigliana Melzi and Margaret Caspe. 2005. Variations in maternal narrative styles during book reading interactions. *Narrative inquiry* 15, 1 (2005), 101–125.

[43] Jon F Miller and Robin S Chapman. 1981. The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research* 24, 2 (1981), 154–161.

[44] Suzanne E Mol and Adriana G Bus. 2011. To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin* 137, 2 (2011), 267.

[45] Koichi Mori, Rafael Ballagas, Glenda Revelle, Hayes Raffle, Hiroshi Horii, and Mirjana Spasojevic. 2011. Interactive rich reading: enhanced book reading experience with a conversational agent. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 825–826.

[46] Julie Hagen Nilsen. 2019. *"It knows how to not understand us!" A study on what the concept robustness entails in design of conversational agents for preschool children*. Master's thesis.

[47] Anthony D Pellegrini, Gene H Brody, and Irving E Sigel. 1985. Parents' book-reading habits with their children. *Journal of Educational Psychology* 77, 3 (1985), 332.

[48] Elaine Reese, Diana Leyva, Alison Sparks, and Wendy Grolnick. 2010. Maternal elaborative reminiscing increases low-income children's narrative skills relative to dialogic reading. *Early Education and Development* 21, 3 (2010), 318–342.

[49] Rebekah A Richert, Michael B Robb, and Erin I Smith. 2011. Media as social partners: The social nature of young children's learning from screen media. *Child development* 82, 1 (2011), 82–95.

[50] Katherine E Ridge, Deena Skolnick Weisberg, Hande Ilgaz, Kathryn A Hirsh-Pasek, and Roberta Michnick Golinkoff. 2015. Supermarket speak: Increasing talk among low-socioeconomic status families. *Mind, Brain, and Education* 9, 3 (2015), 127–135.

[51] Ben Robins, Paul Dickerson, Penny Stribling, and Kerstin Dautenhahn. 2004. Robot-mediated joint attention in children with autism: A case study in robot-human interaction. *Interaction studies* 5, 2 (2004), 161–198.

[52] Paul Rozin and Adam B Cohen. 2003. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion* 3, 1 (2003), 68.

[53] Johanna Ruusuvuori. 2013. 16 Emotion, Affect and Conversation. *The handbook of conversation analysis* (2013), 330.

[54] J Alfredo Sánchez, Norma P Hernández, Julio C Penagos, and Yulia Ostróvskaya. 2006. Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states. In *Proceedings of VII Brazilian symposium on Human factors in computing systems*. 66–72.

[55] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, What's Up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 857–868.

[56] Tina Seidel and Manfred Prenzel. 2006. Stability of teaching patterns in physics instruction: Findings from a video study. *Learning and Instruction* 16, 3 (2006), 228–240.

[57] Candace L Sidner, Cory D Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*. 78–84.

[58] Lori Skibbe, Michelle Behnke, and Laura M Justice. 2004. Parental scaffolding of children's phonological awareness skills: Interactions between mothers and their preschoolers with language difficulties. *Communication Disorders Quarterly* 25, 4 (2004), 189–203.

[59] Catherine Snow. 1983. Literacy and language: Relationships during the preschool years. *Harvard educational review* 53, 2 (1983), 165–189.

[60] Luiza Superti Pantoja, Kyle Diederich, Liam Crawford, and Juan Pablo Hourcade. 2019. Explorations of Voice User Interfaces for 3 to 4 Year Old Children. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, LBW0177.

[61] Tamara Spiewak Toub, Vinaya Rajan, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2016. Guided play: A solution to the play versus learning dichotomy. In *Evolutionary perspectives on child development and education*. Springer, 117–141.

[62] Paul Vogt, Mirjam De Haas, Chiara De Jong, Peta Baxter, and Emiel Krahmer. 2017. Child-robot interactions for second language tutoring to preschool children. *Frontiers in human neuroscience* 11 (2017), 73.

[63] Susan K Wanska, Joanne C Pohlman, and Jan L Bedrosian. 1989. Topic maintenance in preschoolers' conversation in three play situations. *Early Childhood Research Quarterly* 4, 3 (1989), 393–402.

[64] Deena Skolnick Weisberg, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 2013. Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education* 7, 2 (2013), 104–112.

[65] Deena Skolnick Weisberg, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, Audrey K Kittredge, and David Klahr. 2016. Guided play: Principles and practices. *Current Directions in Psychological Science* 25, 3 (2016), 177–182.

[66] Ying Xu and Mark Warschauer. 2019. Young Children's Reading and Learning with Conversational Agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, CS10.

[67] Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and AJ Bernheim Brush. 2018. Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. 300–312.

[68] Andrea A Zevenbergen and Grover J Whitehurst. 2003. Dialogic reading: A shared picture book reading intervention for preschoolers. *On reading books to children: Parents and teachers* (2003), 177–200.

[69] Andrea A Zevenbergen, Sydnee Worth, Delaney Dretto, and Kelsey Travers. 2018. Parents' experiences in a home-based dialogic reading programme. *Early Child Development and Care* 188, 6 (2018), 862–874.