Collaborative Storytelling between Robot and Child: A Feasibility Study

Ming Sun[1], **Iolanda Leite**[2], **Jill Fain Lehman**[1], **and Boyang Li**[1]
[1]Disney Research, Pittsburgh PA, USA
[2]KTH Royal Institute of Technology, Stockholm, Sweden
[1]{ming.sun, jill.lehman, albert.li}@disneyresearch.com, [2]iolanda@kth.se

ABSTRACT
Joint storytelling is a common parent-child activity and brings multiple benefits such as improved language learning for children. Most existing storytelling robots offer rigid interaction with children and do not contribute to children's stories. In this paper, we envision a robot that collaborates with a child to create oral stories in a highly interactive manner. We performed a Wizard-of-Oz feasibility study, which involved 78 children between 4 and 10 years old, to compare two collaboration strategies: (1) inserting new story content and relating it to the existing story and (2) inserting content without relating it to the existing story. We hypothesize the first strategy can foster true collaboration and create rapport, whereas the second is a safe strategy when the robot cannot understand the story. We observed that, although the first strategy creates heavier cognitive load, it was as enjoyable as the second. We also observed some indications that the first strategy may mitigate the difficulties in story creation for young children under the age of 7 and encourage children to speak more. This study suggests that a mixture strategy is feasible for robots in collaborative storytelling, providing sufficient cognitive challenge while concealing its shortcomings on natural language understanding.

Author Keywords
conversational agents; conversational storytelling; child-robot interaction

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION
Storytelling is one of the most common parent-child activities. Modern research finds that it offers substantial benefits to the child, including broadened vocabulary, increased complexity of produced sentences, better narrative comprehension, and accelerated cognitive development [6, 19, 21]. In addition to exposure to new words, conversational storytelling provides an exercise for using language to express concrete meaning with social feedback, which facilitates the acquisition of early literacy [9], which in turn is predictive of later academic success [12]. This fits into the broad pedagogical argument that early intervention at preschool ages can be immensely rewarding [16, 11].

Parents and guardians may not always have the time or energy to engage their children in joint storytelling, despite their best intentions. Recently, interactive, virtual, robotic, and toy characters have emerged as viable play-pals for children [20, 29, 27]. A mixed-initiative storytelling task involves a computational character constructing one or more stories together with a child [3, 24]. Given the benefits of joint storytelling, incorporating storytelling capabilities into interactive characters can bring pedagogical benefits and improve the utility of these characters.

However, traditional storytelling robots or virtual characters interact with children in a rigid and sparse way (e.g., [3, 27]). The robot usually makes little contribution *during* the child's story and waits for the child to completely stop before telling its own story. In contrast, storytelling activities between children and parents or peers tend to contain more interruptions, questions, affirmation, backchanneling behaviors, negotiation over the story content, and so on.

In this paper, we envision a robot capable of performing collaborative storytelling with natural, fluid and fine-grained interaction. For example, when the child experiences difficulty in continuing the story, the robot may encourage further storytelling with questions (e.g., "What happens next?" or "Now your hero has come to a forest. What if a tiger suddenly appears?"), backchanneling, and other social responses, which are known to improve robot-child interaction [5] and argued to be essential for certain types of language learning [15]. Thus, our design goal is to encourage children to tell their own stories by establishing rapport, maintaining engagement, and offering scaffolding, so as to create an experience more entertaining and more engaging than traditional storytelling robots.

Nevertheless, the vision of a collaborative storytelling robot is limited by the current capabilities of artificial intelligence (AI). Despite recent progress, AI remains imperfect in recognizing children's speech and understanding the semantics of natural language. Imperfect speech recognition and natural language understanding imply that the robot may not respond to children in a semantically coherent manner. With these impeding factors, it remains an open question whether fluid

IDC'17, Jun 27–30, 2017, Stanford, CA, USA

ISBN 978-1-4503-4921-5/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078072.3079714

collaborative child-robot storytelling is feasible or is perceived as valuable by children.

CURRENT WORK

In this paper, we report a Wizard-of-Oz experiment designed to test the feasibility of the envisioned form of collaborative storytelling. Two adult experimenters worked with 78 children in an oral storytelling task. The experimenters took the place of the robot to compare the effects of two story collaboration strategies. The collaboration happened when the experimenters introduce five objects and characters throughout the story. The first, contextual strategy is to relate the inserted object or character to the existing story. The second, non-contextual strategy is to avoid relating to the existing story and use mechanical utterances like "Let's include a kitten in the story." Although the first strategy seems ideal, when the robot does not understand the child's story, it has to use the second strategy.

With this experiment, we aim to answer the following research questions:

- What are the effects of non-contextual and contextual prompts on children's storytelling experience?
- Are those effects in the first question specific to children's age and gender?

The effects were measured in terms of (1) the complexity of children's language and fluency in storytelling, (2) recall of story elements in story retelling, and (3) children's self reports of enjoyment. The reason for studying the effects of age and gender is that these factors are substantially correlated with language development for participants in the age group (4-10 years old) that we are interested in [4, 14].

We find that the self-reported enjoyment remained high across the two conditions with no statistically significant difference. The use of contextual prompts caused slight degradation of short-term language performance. This could be explained as the contextual prompts imposed a high cognitive load on children, which created difficulties for young children and boys. We also observe occasional evidence that the contextual prompts, when used in the right situation, provided scaffolding. We conclude that both strategies can be useful for practical interaction design.

RELATED WORK

Mixed-initiative storytelling with children has been a popular topic in interaction design. As technology progresses, the form of interaction has become more natural over time. In early systems like Rosebud [10], children typed stories into a computer. Later systems begin to recognize speech, but the turn-taking mechanism remains relatively simple and rigid. The most common form is to let the robot tell one story after listening to the child's story. The SAGE system [3], for example, first listened to a child's story and responded with a traditional tale in an effort to impart wisdom. The Sam system [24] was an embodied storytelling virtual character, who was a peer to preschool children and told stories around a figurine and a toy castle. The child was asked to tell his/her own story after Sam had finished. In a similar mode of interaction, the storytelling robot in [27] made use of language that was adaptive to the children's level to facilitate learning. In this paper, we push the boundaries of interaction even further and aim for a highly fluid process of co-storytelling where the computational partner is allowed to contribute new characters and objects during children's storytelling.

Most similar to our work, Tewari and Canny [26] tested the feasibility of interactive virtual character carrying out a question-and-answer conversation with preschool children using a Wizard-of-Oz experiment. Our work is similar in that the collaborative storytelling activity makes use of, but is not limited to, question and answer. The conversation in our experiment revolved around the creation of a story, whereas [26] focused on fact-finding.

Another type of story-based interaction consists of robots participating as actors rather than storytellers. Plaisant *et al* [22] used remotely controlled robots to act out stories written by children in order to help children with rehabilitation. GENTORO [25] utilized a robot controlled by hand-held projectors to perform as a character in the story authored by children.

Existing works also studied other issues related to storytelling with children. Benford *et al.* [2] designed graphical interfaces for encouraging children to collaborate in creating stories. StoryMat [7] recorded children's voices to allow collaborative storytelling across time. KidPad [8] is a collaborative story authoring tool that provides children with the ability to draw, type, and insert hyperlinks. Al Moubayed *et al.* [1] considered the problem of providing proper feedback and backchanelling to human storytellers. Their system relied on analysis of video and audio signals of storytellers to synthesize facial expressions and motion behaviors for a virtual human's head and AIBO, a dog-like robot, but SAIBA did not attempt to understand the semantics of stories. Leite and Lehman [17] studied children's sense of privacy in the presense of a seemingly prescient robot during conversational storytelling.

EXPERIMENT

To test the feasibility of the envisioned form of collaborative storytelling and answer the research questions introduced earlier, we designed a task which contained five steps organized in three phases and involved one child and one experimenter at a time. Figure 1 illustrates a general outline. In the first phase (including 1.1 to 1.3 in Figure 1), the child and the experimenter collaboratively created one oral story. In the second phase, the child rated his or her experience in the storytelling activity. In Phase 3, the child was asked to retell the story to a wizarded robot. Two experimenters experienced in interacting with young children took turns to be assigned to an incoming child. The same experimenter went through all three phases with the assigned child. Details of these three phases are described in subsequent subsections.

We recruited 78 children from 4 to 10 years old through postings in physical and online community bulletin boards. Participants were randomly assigned to two conditions (experimental and control), while maintaining the balance of female and male participants across age groups (See Table 1). In total, we had 40 male children and 38 female children, of which 42 are

Table 1: Demographic information of child participants in the experiments.

Age	Control		Experimental		Total
	M	F	M	F	
4.0 - 6.9	11	9	12	10	42
7.0-10.9	8	9	9	10	36
Total	19	18	21	20	78

between 4.0 and 6.9 years old and 36 are between 7.0 and 10.9 years old.

Among the 78 participants, 7 children did not verbally engage in storytelling activity (some used body movements and vocal sounds to tell the story). One participant's speech could not be understood by the data annotator. Two participants' sessions were not recorded due to a software issue. After careful inspection, experimenters did not use enough correct type of prompts in seven participants (more details later). After discarding those data, the analysis for storytelling (Phase 1) used data from the remaining 61 children. For self-reported post-storytelling evaluations, we recorded 76 responses among the total 78 participants. Among the 61 children who successfully participated in the storytelling, we recorded 60 story retelling sessions. We perform analysis of the experiments based on available data.

Phase 1: Collaborative Storytelling

The main activity of Phase 1 is the collaborative telling of a story involving a hero (*Dragon Lady, Flying Toaster, Windstorm*, or *Red Octopus*), a villain (*Trash Can Guy*), four auxiliary entities (a cat, a fork, a pair of roller skates and a bottle of magic potion) and a scene (forest or marketplace). Storytelling took place inside a room instrumented with microphones and a camera. Figure 2 shows the setup. A tablet app (concealed from the child) was developed to notify the experimenter of the current child's condition (experimental vs. control) and the remaining time before the introduction of the next character or object. In order to ground concepts and establish joint attention, which is believed to improve learning [9], we printed the two scenes on letter-sized paper and presented all characters and entities as small cards on stands. The cards remained hidden from the child until they were introduced into the story.

The experimenter and the child first engaged briefly in small talk to build rapport. Immediately after that, the child was asked to select a super hero from four candidates as the main protagonist of the story, as well as a background scene for the story. After the selection, the child was prompted to begin telling the story. During the storytelling, the experimenter would verbally and physically introduce the villain and the auxiliary entities one by one to the child.

In order to minimize confounding factors, the collaboration in this storytelling activity was limited to the experimenter introducing entities in a fixed sequence into the story — first the villain, *Trash Can Guy*, followed by the cat, the fork, the pair of roller skates, and the potion, in roughly one minute intervals. The corresponding entity card was shown to the

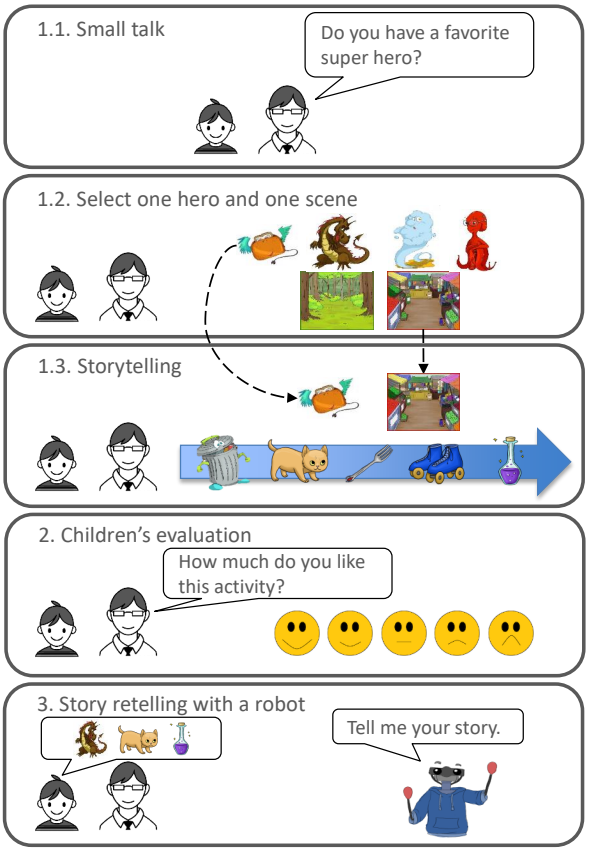


Figure 1: The outline of the procedure. The first phase starts with a small talk (1.1) and the selection of a hero and a scene (1.2). In the main activity (1.3), the experimenter collaborates with the child in telling a story by introducing five entities in roughly one-minute intervals. In the second phase, the child evaluates his or her experience. In the third phase, the child retells the story to a wizarded robot.

child at the same time. After the introduction of a new entity, the child was not restricted to exclusively talk about it. As an effort to maintain consistency across the two experimental arms, we utilized a time window for introducing new entities. The time window was from 50 seconds to 70 seconds (i.e., 1 minute $\pm$ 10 seconds) after the last entity's introduction. This gave the experimenter some flexibility in choosing the best moment to introduce an entity. This time window was displayed on the tablet to the experimenter, but not visible to the child.

In addition to the *introduction* of a new entity, we designed another type of collaborative storytelling: *encouragement*. The experimenter may encourage the child to create more content, if the child stopped his or her story prematurely before the one-minute time limit.

Either type of experimenter utterance may be *contextual* or *non-contextual*. As mentioned earlier, in the control condition, the experimenter was not allowed to refer to the story in her introduction or encouragement utterances. We call these utterances non-contextual. For example, as non-contextual en-

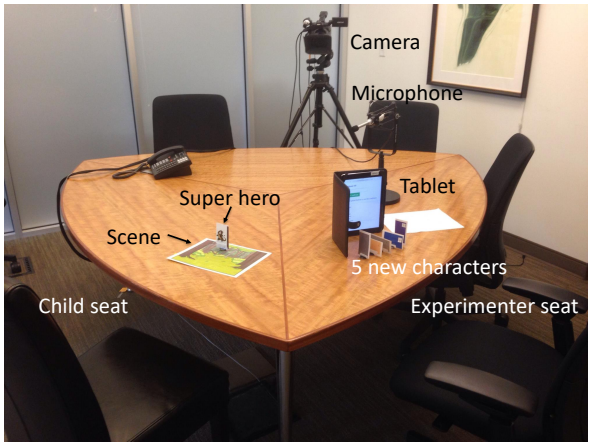


Figure 2: Experimental setting: the child and the experimenter (exp) were seated next to each other. A web-based interface on a tablet indicated the timing for introducing new objects and characters to the experimenter, but was concealed from the child.

couragement, the experimenter can say "What happens next?". As an introduction, she may say "Let's include a kitten." In comparison, in the experimental condition experimenters were instructed to use prompts related to the story thus far. Examples of such utterances include encouragements like "Where did the kitten hide?", "What happened after Trash Can Guy saw the kitten?", or introductions like "Let's say while the kitten was running away from the villain, it found a magic potion in the forest." We call these experimenter utterances contextual. As the experimenters do not always have time to ponder over their utterances in the fast-paced storytelling task, some non-contextual utterances may occasionally occur in the experimental condition.

The rationale for this contextual versus non-contextual dichotomy is to simulate the effect of imperfect natural language understanding and imperfect speech recognition. While a robot can be pre-authored with a set of non-contextual prompts in the control condition, sophisticated comprehension models are needed to generate contextual prompts in the experimental condition. If the robot cannot understand the story well, its best strategy is to avoid referring to the story in order not to make mistakes.

Phase 2: Children's Evaluation of Experience

In this phase, we focus on whether or not children in the control condition (mostly non-contextual prompts) would experience storytelling activity differently from those in the experimental condition (mostly contextual prompts). After the collaborative storytelling activity, participants rated their experience using the Smileyometer instrument [23], which communicates the idea of Likert scale using smiley faces. The children were asked to choose one face for each of the following two questions:

Q1: How much do you like this activity?

Q2: How brave was your superhero?

Prior to the storytelling activity, the children were briefly trained to understand the smiley scale: they were first asked about a few daily events (such as favorite vegetable, the feeling when somebody stepped on their toes). Experimenters helped them associate their feelings with appropriate smiley faces.

Phase 3: Story Retelling

In the third phase of the experiment, the child was led away from the room to interact with a puppeteered robot. The robot asked the child to retell the story. The experimenter stood next to the child and provide assistance as necessary. One type of assistance was explaining what the child was expected to do (e.g., "you can just tell the robot your story"). The second type of assistance happened when the child struggled to remember his/her story. In this case, the experimenter may provide a minimum hint by saying, for example, "I think you have a flying toaster in your story." The interaction among the child, the robot and the experimenter was recorded and transcribed into text. Phase 3 allows us to measure the child's memory of the collaboratively created story, as recall may be related to language learning, which is believed to be a major benefit of storytelling [6, 19, 21]. The design of Phase 3 is part of a larger study. The robot was puppeteered by a researcher experienced with child-robot interaction, who followed a preauthored script. The controls include hand-held controllers and a virtual reality headset connected to two cameras mounted on the robot's head. We refer interested readers to a parallel publication [18] for more details.

Data Annotation

One experienced annotator annotated the storytelling (Phase 1) video with the ELAN software package [28]. Using the display of waveforms, the annotator marked the utterance boundaries with approximately 100ms silence at the beginning and the end. She transcribed each utterance into text and associated each experimenter utterance to indicate (1) whether an utterance referred to some elements in the story (CON vs. NONCON) and (2) whether the utterance introduced a new character into the story or simply encouraged a child to continue (INTRO vs. ENC). See Table 2 for a detailed explanation of these functions. The following conversations were discarded: small talk before the storytelling (i.e., discussion on the child's favorite super hero) and chat after the end of a story, as well as experimenter utterances which served as acknowledgments, empathy or back-channel (e.g., "Red octopus is my favorite too!", "Ok, so the cat is gone.", "++uh-huh++").

For children's self evaluation of the experience (Phase 2), we converted the smiley faces selected by the children to ratings from 1 to 5, where 1 indicates the most negative and 5 indicates the most positive.

We manually analyzed the transcripts of story retelling (Phase 3) to count the characters and objects recalled by the participants. We looked for mentions of the five entities and the protagonist introduced in storytelling (Phase 1) and ignored linguistic variations. For example, we considered the "cat" character as successfully recalled as long as the child mentioned any of the words "cat", "kitten" or "kitty". Table 3

Table 2: Functions of each experimenter utterance.

Function	Explanation	Examples
CON-INTRO	*Contextually* introduce a **new character** to the child.	"What will happen if *the kitten* finds this **fork**?" "But wait. *He lay down* and he found a **fork** underneath him." "But *when he was at the supermarket*, he found the **roller-skate**."
NONCON-INTRO	Non-contextually introduce a **new character** to the child.	"Let's include a **kitten** in the story." "Let me give you a **magic potion** and now you can end your story."
CON-ENC	*Contextually* encourage the child to continue with the story.	"*OK, kitty and dragon are hanging around the forest*. What will happen?" "What happened after *trash can appeared in the market place and flying toaster was all around flying in the market place*. Where did they go next?"
NONCON-ENC	Non-contextually encourage the child to continue with the story.	"Go for it!", "No, no, no. This can't be the end. We still have time." "So what happened next?"

shows an example of this process. The hero character, *Windstorm*, was not considered as a successful recall because it was directly hinted by the experimenter in Turn 5. In Turn 8, the three characters mentioned by the child himself were counted as successful recalls.

RESULTS

In this section, we report our analysis of the experiment. We first examine linguistic measures, followed by enjoyment and story recall.

Linguistic Measures

We define an *exchange* as two turns in the storytelling, in which the experimenter and the child each speak once. Note that both the experimenter's turn (or prompt) and the child's turn in one exchange can contain more than one utterances. The experimenter always initiated the conversation, so her prompt always appeared before the child's turn in an exchange. An example of turn segmentation as well as exchange segmentation is shown in Table 4. The experimenter prompt in one exchange was annotated as either *contextual* or *non-contextual* based on experimenter utterance function — a prompt is considered contextual if at least one of the experimenter utterances in that prompt was annotated as contextual introduction of new character (CON-INTRO) or contextual encouragement (CON-ENC). Otherwise, this prompt is considered non-contextual.

Experimenters were asked to use either contextual or non-contextual prompts according to the experiment condition. However, due to time pressure, experimenters sometimes were not able to meet this requirement and used the wrong type of prompt. For 61 out of the 68 participants, the experimenter used the correct type of prompts more than 70% of the exchanges. In order to keep our experiment strictly between-subject, we used data from these 61 participants and computed the average performance for each individual.

We propose four linguistic measurements of the quality of the children's storytelling. As measures of language complexity, we used the *duration* of the child's entire response in one exchange and the *mean length per utterance* (MLU), which is the number of words per utterance. We computed two measures of fluency as follows: *Content speech rate* is defined as the number of words per minute, excluding stopwords (e.g., "the", "that") and fillers. *Response latency* is defined as the time the child spent before responding to an experimenter prompt. These four measures were automatically computed based on the human annotations of the stories.

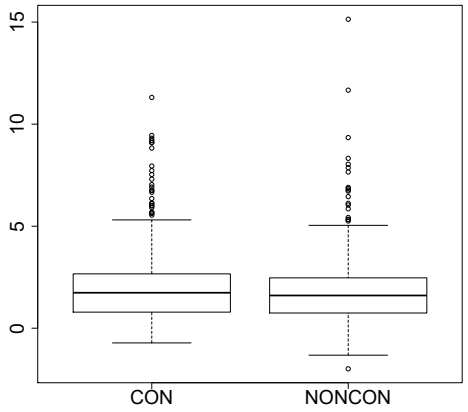


Figure 3: Box-and-whisker plots for experimenter latency across contextual (CON) and non-contextual (NONCON) groups.

Experimenter latency is defined as the amount of time the experimenter waited before she started her prompt in an exchange. It may be a confound variable because it indicates whether the experimenter prompt is well timed or not, which may have effects on the storytelling activity. We find this variable to be well controlled across contextual and non-contextual groups, as their distributions are very similar (see Figure 3). To see if values of experimenter latency in the contextual group and the non-contextual group are identically distributed, we first check if they can be described by Gaussian distributions or Log-Normal distributions using the Shapiro-Wilk test, yielding negative answers for both. Using the non-parametric Wilcoxon-Mann-Whitney test and Kolmogorov-

Table 3: A transcript for a story retelling. [unint] indicates unintelligible speech.

Turn ID	Transcription	Recalled objects
1	Robot: "Was storytelling fun?"	[]
2	Child: "Yeah."	[]
3	Robot: "That's great. We like to keep you happy. Tell me your story."	[]
4	Child: "++uh++, so it was long, but one part was that... ++um++ ..." (*Child turned to the experimenter*)	[]
5	Exp: "Who was your, who was your hero? Windstorm."	[]
6	Child: "Yeah."	[]
7	Exp: "One part was windstorm, and what happened?"	[]
8	Child: "**Kitty** came to ... and to scare the **fork** and the ... and the [unint], and **trash can**."	[cat, fork, villain]
9	Robot: "That was a very interesting adventure. Is there time for me to tell a story?"	[]
10	Exp: "No. Sorry about that. We have to go play the next game."	[]

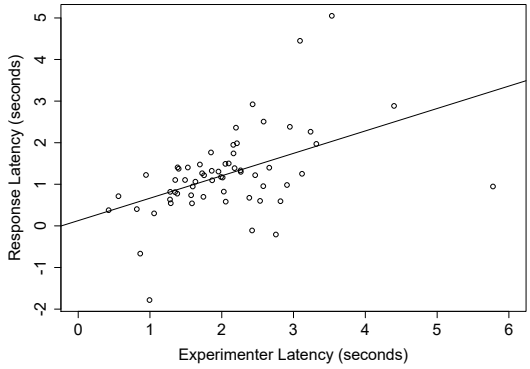


Figure 4: The relationship between experimenter latency and children's response latency.

Smirnov test, we cannot reject the null hypothesis that the two distributions of experimenter latency are the same. Given our reasonably large sample sizes ($N_{contextual} = 385$ exchanges, $N_{non-contextual} = 348$), this suggests the contextual and non-contextual groups are quite similar in this aspect.

We conduct analysis of covariance (ANCOVA) to determine the difference between contextual and non-contextual prompts on children's language performance, controlling for experimenter latency. We do not find significant interaction between contextualization and experimenter latency across our measurements. We find that contextualization significantly impacted the duration of children responses ($F = 8.7$, $p = .005$). Non-contextual prompts yielded longer responses ($M = 11.80$ seconds, $SD = 7.09$) compared to contextual prompts ($M = 6.30$, $SD = 5.91$). For MLU, non-contextual prompts encouraged longer sentences ($M = 9.07$ words/utterance, $SD = 3.14$) compared to contextual ones ($M = 6.89$, $SD = 2.97$).

Experimenter latency is associated with content speech rate ($F = 5.3$, $p = .025$) and children's response latency ($F = 18.1$, $p < .001$). As shown in Figure 4, longer experimenter latency was associated with worsen language performance (e.g., longer children response latency).

Age Effects

We first separate the children into a younger group who are between 4.0 and 6.9 years old and an older group whose ages are between 7.0 and 10.9. After that, we perform a two-way ANCOVA to investigate the impact of contextualization on children's verbal responses controlling experimenter latency.

In the younger age group, we find significant interaction between contextualization and experimenter latency ($F = 8.6$, $p = .007$) on duration. The interaction is shown in Figure 5a. For control group (non-contextual prompts), longer experimenter latency is associated with shorter responses. However, when providing contextualized prompts, this trend can be reversed (see the red dashed line in Figure 5a). We find a main effect of contextualization ($F = 11.9$ $p = .002$) on duration where non-contextual prompts yielded longer responses ($M = 10.39$ seconds, $SD = 7.37$) compared with contextual prompts ($M = 4.06$, $SD = 2.30$). For MLU, non-contextual prompts significantly encouraged longer sentences ($F = 4.6$, $p = .040$). On average, children produced 7.65 ($SD = 2.71$) words per utterance after non-contextual prompts. But after contextual prompts, they produced 5.57 ($SD = 2.39$). For content speech rate, we find that longer experimenter latency is associated with significantly slower yield of meaningful content ($F = 5.3$, $p = .028$). For response latency, we find significant interaction between contextualization and experimenter latency ($F = 6.5$, $p = .016$). As the red dashed line in Figure 5b shows, contextual prompts can help attenuate the negative impact of long response latency.

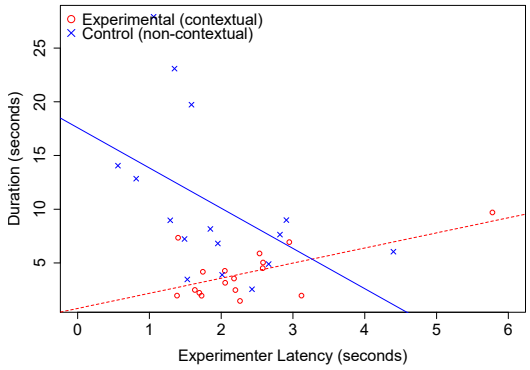
For older children, we do not find any statistically significant interaction or main effects on duration, MLU, and content speech rate. However, we find experimenter latency to be strongly associated with children response latency. Again, longer experimenter latency co-occurs with longer children response latency.

Gender Effects

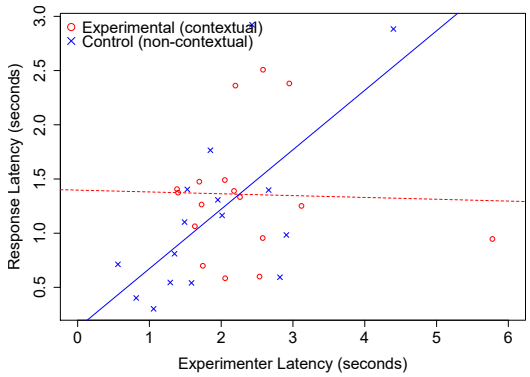
We separated participants to boys and girls and conducted Type-II ANCOVA within each group. We first report the results on boys. For duration, there is a statistically significant difference between contextual and non-contextual prompts ($F = 11.8$, $p = .002$). Non-contextual prompts led to children generating longer responses ($M = 11.83$ seconds,

Table 4: A transcript for the three steps in Phase 1. C denotes the child and E denotes the experimenter. Subscripts denote different turns.

1.1 Small Talk	E_1	Do you like superhero stories?
	C_1	Yes.
	E_2	Who is your favorite superhero?
	C_2	Spiderman.
	E_3	Tell me about his adventure.
	C_3	He is ...
		...
1.2 Hero and Scene Selection	E_5	There are red octopus, windstorm, flying toaster, and dragon lady. Who do you want to be the hero of your story?
	C_5	Toaster.
		...
1.3 Story-telling	E_8	One day, flying toaster was in the market when the villain, trash can guy, showed up. What happened?
	C_8	The trash can started to eat all the fruits and vegetables. Or at least all the stuff.
	E_9	And then what happened? (*story continues ...*)
	E_{12}	Alright, well, let's see what will happen if somebody has roller skates. (*story continues ...*)
	E_{24}	Is there anything else that happens?
	C_{24}	[Toaster] brings it [magic potion] to his house and use it the next time he needed to save somebody. The end.



(a) Duration



(b) Response Latency

Figure 5: The interaction between contextualization and experimenter latency on young children's language performance.

$SD = 7.55$) compared with contextual prompts ($M = 4.52$, $SD = 2.35$). For MLU, non-contextual prompts are statistically significantly better as well ($F = 4.4$, $p = .046$). On average, children's utterances are 8.74 words long ($SD = 2.48$) after non-contextual prompts. However, they are only 6.50 ($SD = 3.14$) after contextual prompts. For content speech rate and latency, we find a statistically significant main effect of experimeter latency ($F = 4.9$, $p = .037$ for content speech rate; $F = 13.2$, $p = .001$ for children response latency). Longer experimenter latency is associated with slower content speech rate and longer child latency.

For girls, we do not find significant interaction or main effects on duration, MLU, and content speech rate. We find a main effect of experimenter latency ($F = 6.4$, $p = .017$) on child response latency. Again, longer experimenter latency is associated with worsen response latency.

Enjoyment

We collected ratings of enjoyment using the Smileyometer instrument [23]. The data we obtained from this evaluation are skewed; most participants rated the aforementioned two questions with the most positive rating, 5. Given the ordinal nature of the ratings, we use ordinal logistic regression to investigate the impact of gender, age group, condition (experimental or control) as well as their interactions on those ratings. For Q1, there is no significance found. For Q2, younger children rated the braveness of their superheroes significantly higher ($p = 0.04$).

Character/Object Recall

We also measured the recall of the six individual story entities in story retelling. Among the aforementioned 61 children who successfully participated in storytelling in the correct conditions, we lost 1 recording of story retelling due to technical issues. Using Fisher's exact test to compare the control group and the experimental group, we do not find any statistical significance.

Separating the participants into two age groups, we observe that older children (7-10 years old) in general recall more than younger children. For hero characters, children in old group statistically significantly recalled more than children in young group ($p = .021$). For different gender groups, as shown in Figure 7b, girls tended to recall more than boys. The difference is significant for the potion object ($p = .016$). Participants tended to recall earlier characters like heroes and villains than entities introduced later.

We further inspect the total number of entities recalled and the effect of gender, age and contextualization using a three-way type-II ANOVA. We find a marginal contextualization effect ($F(1,53) = 3.9$, $p = .055$). Children in the control group recalled more ($M = 2.69$, $SD = 2.05$) than those in the experimental group ($M = 1.79$, $SD = 1.45$). We also found a marginal gender effect ($F(1,53) = 3.4$, $p = .072$) that girls tended to recall more ($M = 2.68$, $SD = 1.87$) than boys ($M = 1.83$, $SD = 1.73$). We find neither statistically significant effects of age nor interaction between factors.

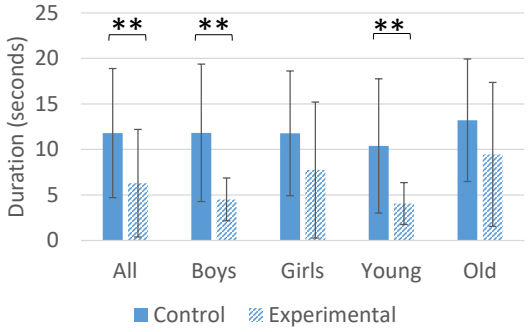
DISCUSSION

The goal of this experiment is to study the effects of contextual and non-contextual prompts. In the control group, the experimenters were not allowed to refer to the story content when they encouraged the child or introduced new entities into the story. In the experiment group, the experimenters were asked to refer to the story content whenever possible during the two types of utterances.

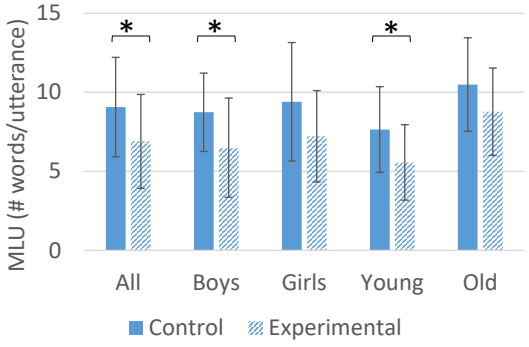
Measures on language complexity, including response duration and MLU, showed the children spoke less and spoke less fluently when contextual prompts were used. This may be explained by the hypothesis that collaborative storytelling imposes a heavy cognitive load on children. When an adult experimenter proposed a change to the story as a contextual prompt, the child must understand the proposed change to the story and incorporate this change into his or her own idea of the unfolding story. This extra work was absent in the non-contextual condition.

Upon detailed inspection, we note the reduction in response duration and MLU concentrated on children younger than 7 and boys, which are groups that are typically less well developed in speech production and, more generally, language abilities. For instance, Hyde and Lynn's meta-analysis of 165 studies [13] found speech production to be the area with the most female advantage, though the analysis was not limited to children. More recent studies [4, 14] noted young girls exhibit faster language development than young boys in areas like learning new meanings of familiar words and complexity of produced language. This agrees with the hypothesis that contextual prompts imposed heavier cognitive load on children than non-contextual prompts in the collaborative task.

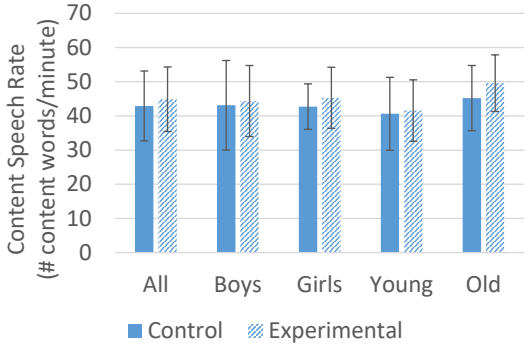
However, we also find that speaking less is not equivalent to using less content words. The fluency measures, especially the content word speech rate, in the two conditions appear rather similar. In addition, the self-reported enjoyment between the two conditions shows no statistically significant difference, so the additional cognitive load is unlikely to discourage children from participation. As long-term participation in a cognitive demanding task may accelerate cognitive development, these results can be considered to corroborate previous findings [6, 19, 21] that storytelling fosters children's cognitive development. Admittedly, the current experiment only measures short-term effects of collaborative storytelling. Further studies are needed to ascertain long-term effects of the proposed interaction paradigm.



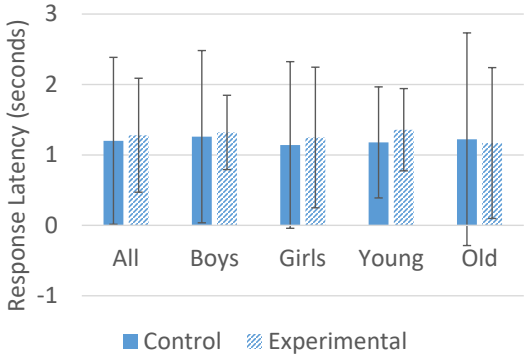
(a) Response Duration



(b) Mean Length of Utterance



(c) Content Speech Rate



(d) Response Latency

Figure 6: The effects of contextualization on (1) all participants, (2) different age groups, and (3) different gender. ** denotes $p < 0.01$. * denotes $p < 0.05$.

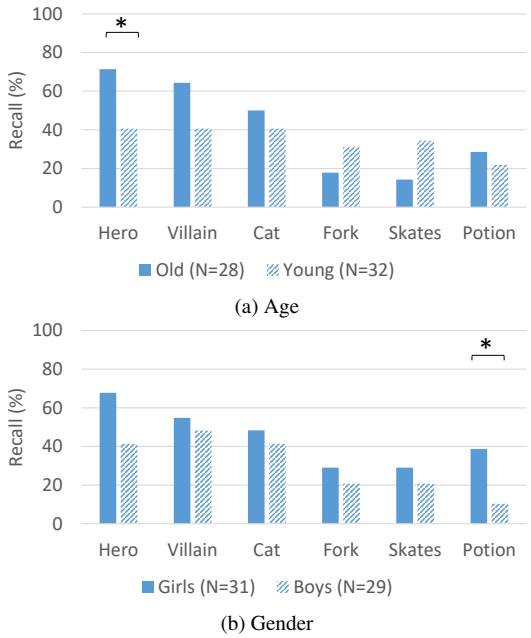
After the effects of individual factors are accounted for, we observe a statistically significant interaction between contextual / non-contextual prompts and experimenter latency among young children. This deserves further discussion. A longer experimenter latency is generally associated with a longer response latency from children (See Figure 4). We provide the hypothesis that this mostly happened when the child ran out of steam and the experimenter waited for the child. That is, a long experimenter latency was likely caused by a pause from the child, rather than a pause from the experimenter who could not find what to say. After receiving the experimenter's prompt, the child likely continued struggling, leading to a short response duration and a long response latency. However, when young children received contextual prompts after a long experimenter latency, the above trend mostly disappeared (See Figure 5). This suggests that a contextual prompt may have given the child something to talk about. We conclude the contexual prompts can provide valuable scaffolding when used in the right situation, such as when the child completely runs out of things to say.

In summary, we found contextual prompts to be cognitively demanding, if occasionally helpful, to children creating stories orally. The participants found the collaborative storytelling task very enjoyable and the two collaboration strategies had statistically insignificant effects on overall enjoyment. This strengthened our confidence in the collaborative storytelling task as an engaging form of interaction for language learning. As a design guideline, we propose that a practical robot should utilize both strategies depending on the child's cognitive development and the robot's ability to understand the story. Contextual prompts should be used to create optimal challenges to accelerate learning. On the other hand, non-contextual prompts can be used when the robot fails to recognize speech or understand the story semantics.

CONCLUSIONS

In this paper, we report a study for social collaborative storytelling between an adult and a child. The adult served as a surrogate for an envisioned storytelling robot, which can contribute new characters and objects to the child's story. Our aim was to investigate the effects of two collaboration techniques, which differ in whether they relate new content to existing story content or not. Relating new content to the existing story intuitively appears to be beneficial. However, it is not applicable when the robot fails to understand the child's story due to imperfect technology.

The experiment finds that both collaboration techniques can create enjoyable storytelling experiences. In general, the contextual technique led to degraded performance on young children and boys in the short term, which can be explained by the high cognitive this technique imposes on children. The experiment also suggests, however, properly positioned contextual prompts can help young children in story creation. Thus, instead of sticking to a single collaboration strategy, we recommend the combined use of two strategies. We believe this would provide the correct amount of cognitive challenge, and also allow a robot to have a smooth interaction even when its understanding of the story may be imperfect. Although



(a) Age

(b) Gender

Figure 7: The effects of experiment conditions, gender and age on recall of story characters. * denotes $p < 0.05$.

further studies are needed to measure long-term effects, this study is a stepping stone toward a natural form of collaborative storytelling between robots and children.

REFERENCES

1. S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J.C. Martin, S. Ondas, C. Pelachaud, J. Urbain, M. Yilmaz, and F. Thalés. 2008. *Multimodal feedback from robots and agents in a storytelling experiment*. Technical Report.
2. Steve Benford, Benjamin B. Bederson, Karl-Petter Akesson, Victor Bayon, Allison Druin, Pär Hansson, Juan Pablo Hourcade, Rob Ingram, Helen Neale, Claire O'Malley, Kristian T. Simsarian, Danaë Stanton, Yngve Sundblad, and Gustav Taxén. 2000. Designing Storytelling Technologies to Encouraging Collaboration Between Young Children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 556–563.
3. Marina Umaschi Bers and Justine Cassell. 1998. Interactive storytelling systems for children: Using technology to explore language and identity. *Journal of Interactive Learning Research* 9, 2 (1998).
4. Caroline Bouchard, Natacha Trudeau, Ann Sutton, Marie claude Boudreault, and Joane Deneault. 2009. Gender differences in language development in French Canadian children between 8 and 30 months of age. *Applied Psycholinguistics* 30, 4 (2009), 685–707.
5. Cynthia Breazeal, Paul L. Harris, David DeSteno, Jacqueline M. Kory Westlund, Leah Dickens, and Sooyeon Jeong. 2016. Young children treat robots as informants. *Topics in Cognitive Science* (2016), 1–11.

6. Adriana G. Bus, Marinus H. Van Ijzendoorn, and Anthony D. Pellegrini. 1995. Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research* 65, 1 (1995), 1–21.
7. Justine Cassell and Kimiko Ryokai. 2001. Making space for voice: Technologies to support children's fantasy and storytelling. *Personal and ubiquitous computing* 5, 3 (2001), 169–190.
8. Allison Druin, Jason Stewart, David Proft, Ben Bederson, and Jim Hollan. 1997. KidPad: A Design Collaboration Between Children, Technologists, and Educators. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 463–470.
9. A. Duranti and C. Goodwin. 1992. *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press.
10. Jennifer Williamson Glos. 1997. *Digital augmentation of keepsake objects: a place for interaction of memory, story, and self*. Master's thesis. Massachusetts Institute of Technology.
11. Michael J. Guralnick. 2016. Why Early Intervention Works: A Systems Perspective. *Infants and young children* 24, 1 (2016), 6–28.
12. B. Hart and T.R. Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H. Brookes Publishing Co.
13. Janet S. Hyde and Marcia C. Linn. 1988. Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin* 104 (1988), 53–69.
14. Margarita Kaushanskaya, Megan Gross, and Milijana Buac. 2013. Gender differences in child word learning. *Learning and Individual Differences* 27 (2013), 82–89.
15. Patricia K. Kuhl. 2007. Is speech learning gated by the social brain? *Developmental Science* 10, 1 (2007), 110–120.
16. S. H. Landry, K. E. Smith, P. R. Swank, and C. Guttentag. 2008. A Responsive Parenting Intervention: The Optimal Timing Across Early Childhood For Impacting Maternal Behaviors And Child Outcomes. *Developmental Psychology* 44, 5 (2008), 1335–1353.
17. Iolanda Leite and Jill Fain Lehman. 2016. The robot who knew too much: toward understanding the privacy/personalization trade-off in child-robot conversation. In *Proceedings of the 2016 conference on Interaction design and children*. ACM, 379–387.
18. Iolanda Leite, André Pereira, and Jill Fain Lehman. 2017. Persistent memory in repeated child-robot conversations. In *Proceedings of the 16th International Conference on Interaction Design and Children*.
19. Lesley Mandel Morrow. 1985. Retelling Stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal* 85, 5 (1985), 647–661.
20. Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots*.
21. Carole Peterson and Allyssa Mccabe. 2004. Echoing our parents: Parental influences on children's narration. In *Family stories and the life course: Across time and generations*. Lawrence Erlbaum, 27–54.
22. Catherine Plaisant, Allison Druin, Corinna Lathan, Kapil Dakhane, Kris Edwards, Jack Maxwell Vice, and Jaime Montemayor. 2000. A storytelling robot for pediatric rehabilitation. In *Proceedings of the fourth international ACM conference on Assistive technologies*. ACM, 50–55.
23. Janet C. Read and Stuart MacFarlane. 2006. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children*. ACM, 81–88.
24. Kimiko Ryokai, Cati Vaucelle, and Justine Cassell. 2003. Virtual peers as partners in storytelling and literacy learning. *Journal of computer assisted learning* 19, 2 (2003), 195–208.
25. Masanori Sugimoto, Toshitaka Ito, Tuan Ngoc Nguyen, and Shigenori Inagaki. 2009. GENTORO: A system for supporting children's storytelling using handheld projectors and a robot. In *Proceedings of the 8th International Conference on Interaction Design and Children*. 214–217.
26. Anuj Tewari and John Canny. 2014. What did Spot hide?: A question-answering game for preschool children. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. 1807–1816.
27. Jacqueline Kory Westlund and Cynthia Breazeal. 2015. The interplay of robot language level with children's language learning during storytelling. In *Proceedings of the 2015 ACM/IEEE international conference on Human-Robot Interaction (Extended Abstracts)*. ACM.
28. Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of LREC*.
29. Chun-Cheng Wu, Chih-Wei Chang, Baw-Jhiune Liu, and Gwo-Dong Chen. 2008. Improving vocabulary acquisition by designing a storytelling robot. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*. 498–500.