

Interactive Storybooks with a Robot Companion

by

Hanna Lee

S.B., Massachusetts Institute of Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
May 24, 2018

Certified by
Cynthia Breazeal
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Interactive Storybooks with a Robot Companion

by

Hanna Lee

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2018, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The strong correlation between children’s early literacy skill and later academic (and economic) success has motivated much research into how children learn to read and what interventions aid the learning process. Research shows that children’s reading ability improves through shared reading exercises with parents, personalized feedback and curriculums, and reinforced associations between audio and visual representations of words. These findings, along with recent advances in technology, have prompted questions about the efficacy of modern educational systems, including digital books, online language learning programs, and robot tutoring systems. There is much interest in these technologies because they have the potential to provide interactivity, personalization, and scalability. At the same time, as with all new technologies, it is essential to consider the accessibility of such experiences, and take steps to make the technology easily usable and available. This thesis explores the concept of an easily authorable interactive storybook with a robot peer tutor, combining several previously studied ideas into one system. One major contribution of this work is a novel interactive reading system consisting of a storybook tablet app, a robot tutoring agent, and an online authoring interface that anyone can use to create stories for the app. Another major contribution is the design and implementation of a user study in which children and parents interact with the system, and the subsequent evaluation of the system across the dimensions of child learning, interaction design, and engagement. The results of the user study suggest that children improve their knowledge of the pronunciation and meaning of target words in the story through participating in the interactive reading experience with the robot, and that both parents and children find the experience engaging. The presentation of these results is followed by a discussion of technical limitations of the system and ways to improve the interaction for future deployment at a larger scale.

Thesis Supervisor: Cynthia Breazeal

Title: Associate Professor of Media Arts and Sciences

Acknowledgments

This thesis would not have been possible without many of the wonderful people in both my academic and personal lives. First, I would like to thank my advisor, Dr. Cynthia Breazeal, for her high level vision and guidance of my research direction, and strong leadership of the entire group. Next, I would like to thank our research scientist, Dr. Hae Won Park, for her support and patience, particularly when I was feeling doubtful of my ability to succeed. I benefited greatly from her design advice and technical assistance during the ideation and development processes of my project, and the entire group is lucky to benefit from the example she sets in kindness, leadership and work ethic.

There are several other members of the lab that I would like to mention. Sam was one of the first group members I worked with, and he has been a constant provider of good conversation, advice on grad school and career development, and help for all things Jibo-related. I have shared fun and insightful conversation with Nikhita, Sooyeon, Anastasia, Stefania, and Safinah, sometimes late into the night, and I am grateful them for brightening up my time in lab. All other members of the group, including Huili, Randi, Ishaan, Pedro, Jackie and Meng, have been supportive and friendly during my year here, and I will miss everyone in the years to come. And I owe a lot to Polly, our amazing administrator, whose passion, infectious smile, and generosity never failed to cheer me up. I will bring back lots of chocolate to share when I come back to visit the lab!

I would also like to thank my two undergraduate researchers, Hyemin Bang (MIT, 2019) and Hiya Vazirani (Wellesley, 2019), for their dedication and willingness to learn. They contributed to both design and development of various components of the project, as well as the careful annotation of user study video footage. They were a pleasure to work with, and taught me to be a better manager and leader.

In addition to the great community I had in the Personal Robots Group, I also want to acknowledge people in my life outside of the lab who have helped me along the way, and to whom I owe many thanks. I would first like to thank all of the members of the MIT Muses, my a cappella group, who gave me a safe and fun space to sing, be silly, and relax during the year. I would next like to thank my boyfriend, Larry, for all of the visits, phone calls, memes, laughter, and love. Larry, spending time with you cheers me up no matter how down I am, and makes all the good times even better. I am excited begin the next chapter of our

lives together in New York. Last but not least, I must thank my family - Mom, Dad, Mimi and Fafa. They always provide me perspective and support, and have helped me become more confident and strong. When I look over my shoulder to see how far I've come, I realize that I stand on the shoulders of giants, and owe many of my achievements to my parents' sacrifice and dedication. To my family - I hope I've made you proud, and I love you all very much.

This work was supported by the National Science Foundation (NSF) under Grants IIP-1717362 and IIS-1734443. Any opinions, findings and recommendations expressed in this document are those of the author and do not represent the views of the NSF.

Contents

1	Research Statement and Motivation	17
2	Related Work	21
2.1	Learning to Read	22
2.2	Digital Reading Systems	23
2.3	Story Authoring Tools	24
2.4	Robot Peers and Personalization for Children’s Literacy Education	25
3	Storybook Asset Management and Tablet App Development	27
3.1	Data Format to Represent Stories	28
3.2	Amazon Mechanical Turk for Object Labeling	30
3.3	Timestamp Alignment for Story Audio	30
3.4	Tablet Features and Design	34
3.4.1	Dynamic Asset Loading and Templatization	36
4	Authoring Interface	39
4.1	Design and Features	39
4.2	Stack	43
4.3	Created Stories Easily and Immediately Transferred to Tablet App	43
5	Reading Interaction with a Robot: Design and Implementation	47
5.1	System Overview	47
5.2	Robot Agent: Jibo	47
5.3	Communication Among Components	49
5.4	Two Modes of the Interaction	52
5.5	Controller and Finite State Machine	54

5.5.1	Evaluate Mode State Machine	54
5.5.2	Explore Mode State Machine	57
5.6	Deciding What Questions Jibo Should Ask	57
6	User Study	61
6.1	Study Overview	61
6.1.1	Participant Information	61
6.1.2	Study Stages	61
6.2	Study Protocol	63
6.2.1	Phase 1: Evaluate Mode	63
6.2.2	Phase 2: Authoring Interface	67
6.2.3	Phase 3: Explore Mode	68
6.2.4	Data Collected	69
7	Results and Discussion	71
7.1	Learning Objectives	71
7.1.1	Pretest to Posttest Improvement	71
7.1.2	Midstory Effect of Jibo Interventions	78
7.2	System and Interaction Design	83
7.3	Engagement	89
7.3.1	Authoring Interface Qualitative Discussion	89
7.3.2	Child Engagement and Enjoyment Qualitative Discussion	91
8	Conclusion and Future Work	95
A	Predetermined Questions for Jibo to Deliver in Evaluation Mode	97
B	Video Annotations for System Design Evaluation	99

List of Figures

3-1 An example of the StoryMetadata JSON format used to provide high level information about a story to the tablet before all assets for that story need to be downloaded. 28

3-2 An example of the Page JSON format used by the tablet to load an interactive page. The Page JSON file provides information about audio and image files, audio timestamps, and triggers (associations) between words and labeled objects in the image. 29

3-3 A screenshot of the interface used to collect data on locations of objects in the scanned storybook images. The workers are given the image and the associated storybook text, and asked to label important objects that appear in the provided text, and also a few important objects that do not appear in the text. The labeling is done by dragging the mouse to create a light green box around the object, and then entering a label in the text box that appears. This Turk task, combined with other processing scripts, would result in the the JSON file shown in Figure 3-2. In that JSON file, there are three sceneObjects, each one corresponding to a box labeled by a Turk worker. . . . 31

3-4	A simple example of the method used to obtain timestamps for audio files of people speaking the story text. The alignment algorithm provides the most accurate timestamps possible for the ground truth transcription by using a potentially faulty transcription and timestamps provided from Google Speech API. Often, the errors are more complicated, and cannot be solved by simply splitting up the first and last timestamps to accommodate extra words in the ground truth transcription. The alignment algorithm is a dynamic programming algorithm that finds the best alignment regardless of where the errors occur.	33
3-5	Various views of the storybook app, sized in this example for a Google Nexus tablet, 2048 by 1536 pixels. The tablet features a scrollable library view for story selection, a title page and the end page, and an interactive reader page.	35
4-1	A screenshot of the authoring interface web application on the title page. The title page allows users to upload an image, give a title, provide audio of a person speaking the title, and listing the target words of the story.	40
4-2	A screenshot of the authoring interface web application when creating a typical story page. Creating a page consists of uploading an image, writing text, adding questions and responses for the robot to deliver, recording audio, and labeling objects in the scene.	41
4-3	A simple diagram of the authoring interface stack. The authoring interface uses AngularJs and Node.js, as well as Express middleware for routing. The database is MongoDB.	43
4-4	A graphical depiction of how story assets flow from the user's computer all the way to the tablet app. Blue arrows indicate results of user events, such as when users press buttons, and black arrows depict the flow of assets and information. The top half figure shows how assets and other story information are pushed to Amazon cloud storage (S3), and the bottom half of figure shows how those assets are then downloaded to the tablet.	44

5-1	The three components of the reading interaction: the controller, tablet, and Jibo. The components communicate using ROS on a local network. The controller is finite state machine that sends messages over ROS via Rosbridge to control the reading interaction by sending commands and receiving inputs from Jibo and the tablet. Note that Jibo and the tablet do not communicate directly. The tablet makes API calls to Amazon S3 and SpeechACE in the cloud to download assets and upload recorded audio for evaluation.	48
5-2	Jibo robots, in both color varieties.	49
5-3	There are three ROS nodes, including the ROS core node, which is where the controller resides. All topics are either subscribed to or published to by the controller; in other words, the tablet and Jibo do not directly communicate with each other. Arrows indicate the flow of messages on the topic.	50
5-4	Left: ROS IP selection screen on tablet app. Right: ROS IP selection screen on Jibo.	51
5-5	This is an overview of the finite state machine that constitutes the high level logic of the controller. The states are color coded into four groups. The brown states are states that are shared by both explore and evaluate mode. The yellow states are those that only exist in explore mode. The green and blue states only exist in evaluate mode. The four blue states control the logic of showing sentences when a new page is loaded and prompting the child to read and providing help if the child asks for it. The four green states control the logic of Jibo asking questions at the end of a page, and handling the child's responses. The transitions are also color-coded. Blue arrows transitions that result from child input (either speech or tactile input on the tablet). Red arrows indicate transitions result from Jibo actions (such as completing a TTS command to deliver speech). Purple arrows indicate transitions that happen without any input, either due to default behavior or timeouts on child responses.	55
6-1	pretests and posttests were administered the same way - by having children read and explain words written on paper cards.	65

6-2	The researcher demonstrated the reading interaction for the child, using a story different than the evaluation story of that child, before the child started reading. This was done to familiarize the child with the process and provide an opportunity to clear up any questions or misunderstandings.	66
6-3	Images of children engaged in the reading task in evaluate mode.	66
6-4	A parent using the authoring interface during the user study.	68
6-5	The two sisters on the right both wanted to read and interact with the tablet during explore mode, and were so excited that they sometimes fought over whose turn it was. The children on the left span an age range of 5 years, but all of them were engaged in the reading activity.	69
7-1	At a glance: the pretest and posttest scores for knowledge (a) and pronunciation (b) of the target words, across all of the children.	73
7-2	Pretest and posttest scores for word pronunciation and knowledge, across different groupings of participants.	74
7-3	The normalized improvement in pre-to-post scores for target word pronunciation and knowledge, grouped by four different factors. Significant differences in knowledge improvement between children in different age groups and reading levels.	76
7-4	The normalized improvement for words that the child already knew how to pronounce (but did not know the meaning of) is higher than for words that the child did not know how to pronounce (or know the meaning of). The normalized improvement for words the child already knew how to pronounce is always 0/0 and is defined to be 1, but it is not plotted since it is not relevant to the analysis.	78
7-5	The pronunciation and knowledge normalized improvement for words that appeared in sentences the child asked for help reading vs words that the child did not ask about. Interestingly, when the child did not ask for help on a sentence containing a target word, the improvement is higher for both pronunciation and knowledge of that target word.	80

7-6	Children who did not ask for help to read sentences tended to exhibit greater improvement in pronunciation and knowledge. This is likely due to the fact that children who are more proficient readers are less likely to ask for help, and can learn from other Jibo interventions and exposure to the words.	81
7-7	The nine possible combinations of first and second utterances are divided into three groups. The blue shaded group represents improvement. The red shaded group represents no improvement or regression. The green shaded group represents perfect scores both times. These three groups represent the three categories of improvement used in the analysis.	82
7-8	More than half of the first-second utterance pairs were both correct pronunciations. Of the remaining pairs, there were about twice as many improvements as no-improvements.	83
7-9	The contingency tables for the analysis of the effect of Jibo interventions on improvement between first and second utterances of target words. The outcome category Improve only includes instances where the utterance score increased, while Improve* also includes instances where the utterance score was perfect both times the child pronounced the word. There is a clear trend suggesting that Jibo asking the child to pronounce words was the most effective, but not enough data exists for the results to be significant.	84
7-10	All system and interaction design annotations from the videos of children in the study. rows represent different types of annotations, and the columns represent the state during which those annotations occurred. The states are states of the controller's finite state machine.	86
7-11	Younger children exhibit more mistakes while using the system, and need more prompting from the researcher. These prompts are for when the child forgets to press the button to see the next sentence, or forgets to ask for help when stuck.	88
7-12	Jibo's ASR system has a harder time understanding younger children than older children.	88
7-13	The results of the agree/disagree questions given on the parent survey.	90
7-14	The results of the agree/disagree questions given on the child survey.	92

7-15 The results of questions on the child survey pertaining to evaluate and explore mode.	93
---	----

List of Tables

5.1	A description of each topic in the ROS network. Each topic is subscribed to or published to by the controller, since the tablet and Jibo do not communicate directly.	53
6.1	Information about the 17 child participants. There were three possible stories for evaluation mode. Some children did not complete evaluation mode because their reading level was too low for any of the stories to be a good fit for the interaction. The first couple of children were pilot testers and no pre and post test data was collected for them.	62
6.2	The target words for each of the three evaluation stories for the user study. Each word appears once in the story, unless otherwise indicated.	64
7.1	The target words for each of the three evaluation stories for the user study. The number in parenthesis indicates the number of occurrence of a word in a story.	78
7.2	Parents' general comments about the experience, collected from an online survey after the conclusion of the study.	91
7.3	Child responses to likes and dislikes about the reading experience.	93
A.1	The predetermined questions for The Hungry Toad.	97
A.2	The predetermined questions for Clifford and the Jet.	98
A.3	The predetermined questions for Henry's Happy Birthday.	98
B.1	The target words for each of the three evaluation stories for the user study. Each word appears once in the story, unless otherwise indicated.	99
B.2	Video annotations in the category of interaction design.	100

B.3	Video annotations in the category of child mistakes.	101
-----	--	-----

Chapter 1

Research Statement and Motivation

Early literacy is an area of great importance. The reading levels of children who enter kindergarten are diverse, and without intervention, these variations in initial reading ability intensify into larger achievement gaps. More specifically, analysis of data from the Early Childhood Longitudinal Study by West et al. shows that only 66 percent of first-time kindergartners are proficient in recognizing letters, and only 29 percent are proficient in understanding letter sound relationships at the beginning of words [33]. Furthermore, the variation in first-time kindergartner performance is high, and often attributed to factors outside the child's control, such as his or her socioeconomic status or parents' level of education. First-time kindergartners whose mothers obtained a Bachelor's degree or higher were over 5 times as likely to score in highest quartile on reading assessments than those whose mothers had less than a high school diploma [33]. Another analysis by Chatterji et al. found that gaps in first grade reading levels became clearly discernable in three major groups: African Americans, boys, and high-poverty children, and that these gaps were primarily associated with a lack of prior reading preparation in kindergarten and at home [7].

These gaps in reading abilities are magnified as students progress through grade levels. Longitudinal studies have shown that kindergarten and pre-kindergarten reading abilities are reliable predictors of academic achievement through elementary school and into high school. Foster et al. found significant differences in third grade literacy ability between students identified as having "low readiness" by a literacy assessment at the time of kindergarten entrance and those who had "average readiness" or "high readiness". For the task of text

comprehension, the deviation in scores between the “low readiness” group and the “high readiness” group increased from an effect size of 1.03 in first grade to 1.22 in third grade, while the effect size between the “average readiness” group and the “high readiness” group decreased from 1.52 to 1.42 [10]. This suggests that though some students catch up to their better prepared peers, many actually fall further behind. Even more concerning is the effect elementary school literacy level has on achievement in high school and beyond. Hernandez et al. reported that one in six children who do not read at a proficient level in third grade do not graduate from high school on time, a rate that is four times greater than that of proficient third grade readers [16]. Hernandez also claims that third grade is a critical time, where students begin to transition from “learning to read” to “reading to learn”, underscoring the importance of basic literacy skills for advancing education in other topics throughout schooling. Academic achievements in high school and beyond are in turn strongly correlated with expected income, with a 2015 study by Tamborini et al. showing that bachelors and graduate degree holders hold as much as \$1.5M more than those with only high school degrees [30].

Due to this strong correlation between early literacy levels and later success, there is a need for research to develop strategies for improving reading abilities in children starting from a young age. Existing research into shared reading activities, where children read directly alongside a peer, teacher, or parent, and receive personalized feedback, is promising. Other efforts, which leverage new technology, have seen mixed results. Smart boards [31], laptop programs [32], and E-book readers [23] are among those efforts that have in fact seen some success in engaging students and aiding their learning. However, one area that is still under researched is the potential robots have to improve learning. The well-supported idea that shared reading exercises promote early literacy skills, coupled with the rise of technology in classrooms and recent advances in social robotics, prompts questions about the role robots can play in education.

When discussing new technology for education, it’s essential to consider public perception and access to such technology. Although we live in a world driven by new gadgets, apps, and AI, many people do not understand or have access to the technology that has become so mainstream. Parents are often wary of their children spending too much time on mindless games online, setting rules to limit children’s screen time [17]. Parents are also wary of children playing with toys that the parents themselves cannot understand, and parents do

not want to feel left behind in the context of children’s development and education. This is why it is also important to create tools that empower parents and educators to set the curriculum and control the content presented on new technological platforms.

The primary goals of the thesis are to explore robot peer tutors as a method of delivering effective and engaging reading experiences to children to improve their vocabulary, and to make the process of story generation accessible to parents and educators. I also believe it is essential that the storybook platform that supports the reading experience is architected in a flexible way that can display a large number of different stories, so that we can deliver learning experiences to children of varying needs in the future. The research questions explored in this thesis are:

1. **Do children improve their vocabulary through interactive reading experiences with a social robot?**
2. **How can we make the technology platform supporting these reading experiences more inclusive and accessible for parents, educators, and the public?**

To tackle these questions, I developed a novel interactive reading experience with a robot companion. The storybooks read during the interaction are presented on a tablet app with features that invite the child to tap and swipe elements of the story. The robot teaches and evaluates children as they read, and engages them with questions and encouragement. I also developed an online authoring tool through which anyone can easily generate new stories that can be read in this interactive way.

The rest of this document is laid out as follows. Chapter 2 explores related work. Chapters 3 and 4 discuss the design of two fundamental components of the system - the tablet application that displays the storybooks, and the authoring interface that allows parents to create content. Chapter 5 details the design of the reading interaction, involving the tablet, a Jibo robot, and a controller. Chapter 6 describes the user study that was conducted to evaluate the reading interaction system, and Chapter 7 discusses the results of that study. Finally, Chapter 8 presents further limitations and suggestions for future work.

Chapter 2

Related Work

In this chapter I provide an overview of previous work done in four areas:

1. The theory of children’s language acquisition and coreading activities
2. Digital reading and evaluation systems
3. Interfaces for story creation
4. Robotic systems for education purposes

There has been a body of related work on the process of developing early literacy skills, the benefits of interactivity and dialogue in reading, and the effect of new technologies on education. Research shows that increased exposure to oral storytelling, and joint parent-child or peer-peer reading experiences are correlated with higher reading achievement. Research also shows that allowing books to come to life through digital media and exposing students to a high quantity of auditory and visual feedback is beneficial, and can provide better data to drive personalization models that tailor the reading experience to each child. Further along those lines, using personalized intelligent tutoring systems have proven to be effective in increasing student scores on certain problem solving tasks and language learning tasks. Finally, while robots are a relatively new factor in the education space, there has been some work on the effect of robots on children’s social and academic behavior, and this work suggests that social robots have the ability to keep children engaged in ways disembodied systems cannot.

2.1 Learning to Read

The development of literacy skills in young children is time sensitive and requires just-in-time intervention. In the 1960s through 1980s, Harvard education professor Jeanne Chall developed a now widely known theory on the six stages of reading development, eventually published in her book [4]. Chall describes the six stages as continuous and overlapping, with each building on the previous. The work in this document relates to the three stages leading up to Stage 2. These stages are summarized below:

- Prereading Stage (up to age 6): growth in use of spoken language, increased control of words and grammar, beginnings of understanding of sound structures
- Stage 1 (ages 6-7): learn alphabet, understand relationship between letters and sounds
- Stage 2 (ages 7-8): building on Stage 1 skills, learn to read words and stories, increased fluency in oral reading

Past these stages, children begin to learn new information through reading, analyze and synthesize what they read, and pass judgment on competing points of view.

Much work has been done to study various methods of improving children’s basic reading skills, i.e. those skills up through Stage 2. In the prereading stage, Neuman et al. [27] found that increased exposure to books and oral storytelling over the course of one school year, along with better training for teachers and caregivers, resulted in both higher average scores and greater pre-test to post-test improvements for tasks such as concepts of print and concepts of narrative. Studies have also found that shared reading exercises, in which students read together with a teacher or parent, were effective in increasing core vocabulary and understanding of what was being read. Whitehurst et al. [35] suggest a shared reading intervention that enhances what they describe as “emergent literacy” skills, which are skills such as vocabulary, knowledge of letters, linguistic awareness, and correspondence between sounds and words. This method is called *dialogic reading*, in which adult co-readers provide thoughtful questions and responses to children while reading to encourage children to react to and process the words they see.

Another study by Bus et al. [3] reports that the number of parent-child joint book reading experiences were correlated with reading achievement and emergent literacy skills. There is also research supporting the advantages of collaborative learning between peers or close-

aged siblings, in which there is no adult expert, and instead there is reciprocity in learning and the two parties mutually aid in each other’s development. For example, a study by Gregory et al. [13] found that through engaging with their older siblings, younger siblings were able to better translate purely academic learning into a more personal learning, since the older sibling could cast topics and new information in a way that was familiar to their shared environment and experiences. A 2017 Harvard study on peer reading in an elementary school classroom reported children with lower initial reading scores who interacted with high achieving peers demonstrated significant score improvements [9]. These findings suggest that a robot agent able to play the role of both expert and peer would be helpful in recreating the benefits found in parent-child and peer-peer interactions.

The idea that dialogic and shared reading interaction is core to children’s early literacy learning while encouraging multifaceted exchange of ideas among the readers is important, but it is difficult for all children to receive this support. The work of this thesis addresses these problems by proposing an interactive reading interface and a social robot reading companion to provide dialogic and collaborative reading experience for the early readers.

2.2 Digital Reading Systems

Interactive digital storybooks currently exist in several forms of multimedia. Such storybooks may be useful for increasing engagement, and the effects of digital storybooks vs print storybooks are promising, but the interactive features of digital storybooks have not yet been shown to achieve significant gains in children’s learning objectives.

In a study using the interactive story app known as the TinkRBook, researchers found that parents and children who read together using the TinkRBook engaged with each other more than those who read together using a normal print book [6]. Similarly, a study of preschoolers who read with either a digital storybook or a print storybook found that those who read from a digital storybook demonstrated greater word recognition score increases but not comprehension scores increases [36]. One interesting result from that same study also found that among children who used the digital book, those who had teacher guidance actually scored lower than those who explored the book independently. One possible explanation is that the digital book platform itself was useful for children, but adding in a teacher returned the children to a normal classroom state and they weren’t as engaged with

the book as a result. In slight contrast, a study by Kelley et al. on word learning with digital storybooks found that after three sessions with a digital storybook that could read aloud and let children touch elements on the screen, children demonstrated only minimal gains in vocabulary, approximately one word per child, suggesting that interactive components alone may not be sufficient, and that guided instruction in addition to interactivity is likely necessary for meaningful language learning with digital storybooks [21]. There are still many questions to be answered about what the correct balance should be between formal instruction and freedom to explore interactive features. Perhaps if the instruction comes from a peer and not a teacher, the students will stay engaged but also benefit from feedback.

2.3 Story Authoring Tools

The digital books described in the previous section are built with hard-coded graphics and interactable components, but are not scalable beyond the contents they pre-built. In the example of the TinkRBook, while the framework is exciting, it can only support stories authored by the app developer, as custom code is required for each story, and most people do not have the domain knowledge or the time to dedicate to a tedious authoring process.

Additionally, although there are examples of customizable digital storybook tools available online, such as StoryBird¹, Storyrobe² and Toontastic³ for iPad, that provide templates or sketches for users to fill in with their own pictures, the produced storybooks are very basic and exist online only. Some tools aim specifically to encourage children to author stories, but these tools are focused more on developing children’s animation skills and familiarity with technology, rather than the story content itself [20]. More advanced tools allow more artistic freedom or even automate some part of the story generation process. Champagnat et al. created an interactive storytelling system in which the author creates content for different stages in the protagonist’s journey, and underlying system automatically constructs high level plot structures [5]. Balet created a comprehensive authoring platform that supports 3D animation and rich interactions [1]. While these types of authoring tools are interesting and impressive in their own ways, one disadvantage they have compared to the system presented in this thesis is that they produce storybooks that stand alone, instead of producing story

¹StoryBird, <http://www.storybird.com>

²<https://storyrobe.wordpress.com/>

³<https://toontastic.withgoogle.com/>

experiences that also include interactions with a tutoring agent. More specifically, they do not provide any form of real-time assessment of the users' reading skills or feedback on the readers' performance. In fact, there are no existing authoring systems in the literature that support authoring stories for reading with a peer tutoring agent, so the system developed in this thesis is novel in that respect.

2.4 Robot Peers and Personalization for Children's Literacy Education

Recent research studying the effects of robot peers on children, particularly in education and learning outcomes, shows promising results. Breazeal et al. [2] found that children who were engaging with a robot during a learning task about animals treated the robots as sources of information and were able to retain information presented by the robot. They also found that children were more receptive to a responsive and animated robot than a stoic one. This is supported by results from Westlund et al. [34] who found that young children retain stories told to them by an expressive robot better than stories told by an unexpressive robot. Another study by Kanda et al. [19] found that when an English speaking robot was introduced in a Japanese school, there was a strong correlation between social interaction with the robot and improvement in English test scores, which they claim is due to high levels of enthusiasm for interacting with the robot. These findings highlight the importance of a robot peer that is actively engaged in the task to keep the child interested and paying attention.

There has been a body of work on how children engage with and react to robots in an educational context. Keren et al. found that kindergartners who engaged with a robot companion demonstrated higher performance on geometric thinking and metacognitive tasks [22]. Another study using the same robot platform also found that children enjoyed interacting with the robot and accepted its authority as a teacher [11]. There already exist several robot learning companions, including Dragonbot [24], Tega [25], Saya [14], and Asimo [28]. These robots provided abilities such as providing instruction, encouraging conversation, and exchanging stories. Of these, only Tega has been used in interactions involving other hardware, e.g. tablets. This motivates further research into how robot agents can be paired with other technology to create a more compelling and effective interaction.

One of the great advantages of tutoring and coreading activities over typical classroom learning is the benefit of personalization. Studies have shown that personalizing curriculum, feedback, task content and task difficulty for students results in greater learning success [8, 15]. For example, Heilman et al. [15] found that personalizing reading passages by tailoring them to the interests of the students increased students scores on post-tests. Additionally, a 2013 study suggests that guiding students in analyzing words and encouraging them to echo words that a peer or tutor has already spoken are both helpful tools for building children’s reading skills [29], suggesting that having a robot peer identify which words a student has pronounced incorrectly and then encouraging them to echo those words could be useful.

In the space of digital books and technology, personalization with e-book readers has proven useful in allowing researchers to track more information about the learning process in elementary schools, despite not leading to a significant difference in students’ reading accuracy [18]. One benefit of the system presented in this thesis is its ability to be deployed at scale with personalized feedback for each child, which could be leveraged for data collection to inform models about student learning patterns.

There has also been research involving personalization within interactions with robot agents. Gordon et al. showed that factoring in the student’s engagement as an augmentation to a Bayesian Knowledge Tracing algorithm improved learning objectives as well as child enjoyment [12]. Leyzberg et al. [26] showed that for a problem solving task, personalizing a robot tutor to deliver particular pre-created lessons to a student based on her strengths and weaknesses increased the students’ scores by one standard deviation.

These results are exciting, because children are able to learn from robots and robots can increase short-term engagement in educational tasks. However, in many of these works, the robot’s behavior is preset according to an experimental condition, and does not adapt based on input from the child. This thesis presents a system that uses a robot agent and also provides personalized evaluation in real time, combining ideas from many of the works cited above.

Chapter 3

Storybook Asset Management and Tablet App Development

One of the early tasks of the thesis was to design a method for transforming a static collection of audio, image and text files into a digital storybook. The story corpus contained scanned pdf images, text, and pre-recorded audio for approximately 80 stories. In the reading task with a robot, children and the robot experience the story together through an application on a shared tablet. This chapter presents the design and development of the tablet application, and the asset management pipeline used for creating and storing information about each story.

One of the main design goals was to create a platform that supported any number of storybooks, so long as the storybooks conformed to an agreed upon data format. This was inspired by studying the arduous process required to create a TinkRBook, a previous interactive storybook application, and realizing that the time and effort needed to produce a book on the TinkRBook platform limited the potential for large scale deployment. The app presented in this thesis was designed in a templated way, such that all stories would support the same types of interactions, but with different content. This is an important design goal because ultimately, the platform must scale in order to be useful in broader contexts outside of the lab, such as in schools or homes.

```
{
  "name": "the_hungry_toad",
  "humanName": "The Hungry Toad",
  "numPages": 15,
  "orientationString": "landscape",
  "targetWords": ["throat", "coat", "groaned", "toaster", "foam", "soap", "rowboat"]
}
```

Figure 3-1: An example of the StoryMetadata JSON format used to provide high level information about a story to the tablet before all assets for that story need to be downloaded.

The main interactions that the tablet app supports as a standalone app are:

- Associate objects in the image to words in the text
- Swipe a sentence to hear audio, with word highlighting for visual feedback
- Provide labels for objects in the image that do not appear in the text

These interactions were chosen because they bring together oral, graphical, and written representations of words. In addition to functioning as a standalone app, the tablet app can also be used as part of the robot reading interaction described in Chapter 5.

3.1 Data Format to Represent Stories

In the storybook app system, a story is represented as a collection of JSON files. There is one StoryMetadata JSON file per story, and a sequence of Page JSON files, each representing a single story page. The story assets, such as images and audio files, are stored in the cloud, and the Page JSON file for a particular page references the names of the necessary assets. The assets follow a canonical naming scheme, and are in a directory structure rooted at a known location in Amazon S3 cloud storage, making them easy to find. The Page JSON files follows a simple format that provides the minimal necessary information for the tablet application to generate the storybook dynamically. An example of one of these files is shown below.

As depicted in Figure 3-2, the JSON fields tell the storybook app where to download assets from, what questions the robot is allowed to ask (if the story is being read with a robot), where to label objects in the image, and what the timestamp of each word is in the provided audio file. Some of this information, such as the object labels and word timestamps, needed to be obtained via algorithms or human input. The next sections detail how the


```

{
  "text": "So Toad hopped down the road to see the doctor. \"A toad does not eat soap,\"
    said the doctor as she took it out. ",
  "storyImageFile": "the_hungry_toad_06",
  "audioFile": "the_hungry_toad_5",
  "timestamps": [
    {"start": "0.0", "wordIdx": "0", "end": "0.225"},
    {"start": "0.225", "wordIdx": "1", "end": "0.6"},
    {"start": "0.6", "wordIdx": "2", "end": "1.2"},
    {"start": "1.2", "wordIdx": "3", "end": "1.5"},
    {"start": "1.5", "wordIdx": "4", "end": "1.7"},
    {"start": "1.7", "wordIdx": "5", "end": "2.1"},
    {"start": "2.1", "wordIdx": "6", "end": "2.5"},
    {"start": "2.5", "wordIdx": "7", "end": "2.6"},
    {"start": "2.6", "wordIdx": "8", "end": "2.9"},
    {"start": "2.9", "wordIdx": "9", "end": "3.4"},
    {"start": "3.4", "wordIdx": "10", "end": "4.5"},
    {"start": "4.5", "wordIdx": "11", "end": "5.3"},
    {"start": "5.3", "wordIdx": "12", "end": "5.7"},
    {"start": "5.7", "wordIdx": "13", "end": "6.2"},
    {"start": "6.2", "wordIdx": "14", "end": "6.5"},
    {"start": "6.5", "wordIdx": "15", "end": "7.0"},
    {"start": "7.0", "wordIdx": "16", "end": "7.6"},
    {"start": "7.6", "wordIdx": "17", "end": "7.7"},
    {"start": "7.7", "wordIdx": "18", "end": "8.2"},
    {"start": "8.2", "wordIdx": "19", "end": "8.5"},
    {"start": "8.5", "wordIdx": "20", "end": "8.7"},
    {"start": "8.7", "wordIdx": "21", "end": "8.9"},
    {"start": "8.9", "wordIdx": "22", "end": "9.1"},
    {"start": "9.1", "wordIdx": "23", "end": "10.0"}
  ],
  "isTitle": false,
  "sceneObjects": [
    {"position": {"width": 161, "top": 70, "height": 149, "left": 11},
      "id": 0, "inText": true, "label": "toad"},
    {"position": {"width": 101, "top": 163, "height": 50, "left": 159},
      "id": 1, "inText": true, "label": "soap"},
    {"position": {"width": 234, "top": 15, "height": 206, "left": 142},
      "id": 2, "inText": true, "label": "doctor"}
  ],
  "triggers": [
    {"args": {"textId": 1, "timestamp": null, "sceneObjectId": 0}, "type": 0},
    {"args": {"textId": 11, "timestamp": null, "sceneObjectId": 0}, "type": 0},
    {"args": {"textId": 15, "timestamp": null, "sceneObjectId": 1}, "type": 0},
    {"args": {"textId": 9, "timestamp": null, "sceneObjectId": 2}, "type": 0},
    {"args": {"textId": 18, "timestamp": null, "sceneObjectId": 2}, "type": 0},
  ]
}

```

Figure 3-2: An example of the Page JSON format used by the tablet to load an interactive page. The Page JSON file provides information about audio and image files, audio timestamps, and triggers (associations) between words and labeled objects in the image.

labels and timestamps were initially obtained prior to the development of the authoring interface that now encapsulates the process.

3.2 Amazon Mechanical Turk for Object Labeling

Amazon Mechanical Turk (MTurk)¹ is a tool developers can use for crowdsourcing workers to complete simple human intelligence tasks. Examples of common tasks include audio transcriptions and object labeling. MTurk allows developers to provide their own HTML/Javascript templates for a custom task. The task I developed asked workers to draw bounding boxes around interesting objects in a given image, that corresponded to words that appeared in some given text. The task also required workers to mark at least two objects whose labels did not appear in the given text. The goal of the task was to collect labels for objects in the storybook images that stood out as important to the story. The images and text were drawn from pages of the storybooks in the corpus. An example of the interface is shown below.

Each storybook page was presented to three different MTurk workers. This data was then downloaded, aggregated, and incorporated into the Page JSON file for the story page that the image/text pair represented. The object labels and the story text for each page were used to generate pairs of object to word triggers, allowing flexibility on plurality and verb tense. For example, if an object were labeled "dog" and the word "dogs" appeared in the text, it counted as a valid match. These triggers were used in the tablet app to associate the words and objects; for example, tapping on a word causes the corresponding object to be highlighted.

After the development of the authoring interface, described in the next chapter, labeled objects were obtained via the user of the authoring interface, instead of by anonymous workers on MTurk. The MTurk task was mainly used to collect data to use as test data during the development of the storybook tablet app.

3.3 Timestamp Alignment for Story Audio

In order to support the feature "swipe a sentence to hear audio, with visual feedback on which word is being pronounced," it was necessary to generate per-word timestamps of the

¹Amazon Mechanical Turk, <http://www.mturk.com>

Label the important parts of the image (see 2 tasks below).

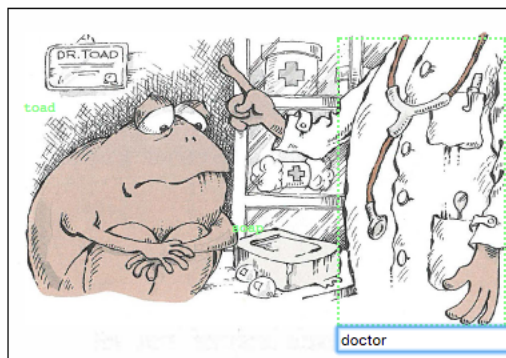
Read the following text and determine which keywords in the text correspond to objects in the image. Keywords can be nouns, adjectives, or verbs. **For each of these keyword-object pairs, use your mouse to draw a tight rectangle around the object then type a label for that object and press ENTER.**

For example, if the text contains the word "cat" and there is an image of a cat in the picture, put a box around the cat and enter the label "cat". If the word "swimming" appears and there is a picture of a person swimming, put a box around the person and label it "swimming".

Task 1: Label between 1 and 3 important objects with labels that appear in the green text.

Task 2: Label between 1 and 3 other relevant objects in the scene that do NOT appear in the green text.

The storybook scene and text are as follows:



So Toad hopped down the road to see the doctor. "A toad does not eat soap," said the doctor as she took it out.

Reset Submit

Figure 3-3: A screenshot of the interface used to collect data on locations of objects in the scanned storybook images. The workers are given the image and the associated storybook text, and asked to label important objects that appear in the provided text, and also a few important objects that do not appear in the text. The labeling is done by dragging the mouse to create a light green box around the object, and then entering a label in the text box that appears. This Turk task, combined with other processing scripts, would result in the the JSON file shown in Figure 3-2. In that JSON file, there are three sceneObjects, each one corresponding to a box labeled by a Turk worker.

audio files for each story. The story corpus contains the audio files for each page, as well as the plaintext that the audio files correspond to. A script then leverages the power of Google Speech API as well as an alignment algorithm in order to batch process pairs of audio and text files. This script first sends the audio file to Google Speech's recognize function to get Google's most accurate guess of the timestamped transcription of each audio file. Then, it modifies the timestamped transcription to better match the ground truth transcription. For the use case of matching timestamps to words, it mattered more that the Google-provided transcription had the same number of words as the ground truth transcription than that the Google transcription was actually correct, since the ground truth transcription was already known. If the Google transcription was the same length as the ground truth transcription, then the provided Google timestamps could be used with no modification.

In the more common case where the Google transcription was slightly wrong and was a different length than the ground truth transcription, it was necessary to run an alignment algorithm that I implemented to get timestamps for the ground truth transcription. The method used in this algorithm is strongly inspired by the well known minimum edit distance algorithm, with some augmentations. The problem is to turn a sequence of timestamp/word pairs into another sequence of timestamp/word pairs, where the output pairs are for the ground truth transcription and the input pairs are from the Google transcription.

A simple example is shown in Figure 3-4. The input transcription has 4 timestamps and the ground truth transcription needs 6. The intuitive way to split up the timestamps should be based on a best match between the words. In this example, the optimal match is to split "apart" into "and bark" and split the timestamp in half as well, and to split "Dogs" into "The dog". Sometimes, the mistakes occur in the middle of a transcription instead of at the ends, and often, the error is that the Google transcription is longer than the ground truth transcription, meaning timestamps need to be combined instead of separated. The dynamic programming algorithm handles all of these cases. In the original edit distance algorithm that inspired the alignment algorithm, the cost of a particular alignment is the sum of the number of mistakes, where a mistake is an imperfect match between two characters. The timestamp alignment algorithm augments this technique by assigning nonbinary cost to each mistake, so that matching "dogs" to "dog" accumulates a lower cost than matching "dogs" to "will". The cost of a mistake is determined via a word similarity metric that is calculated using a stemmer, homonym detection, and hamming distance. If two words have the same

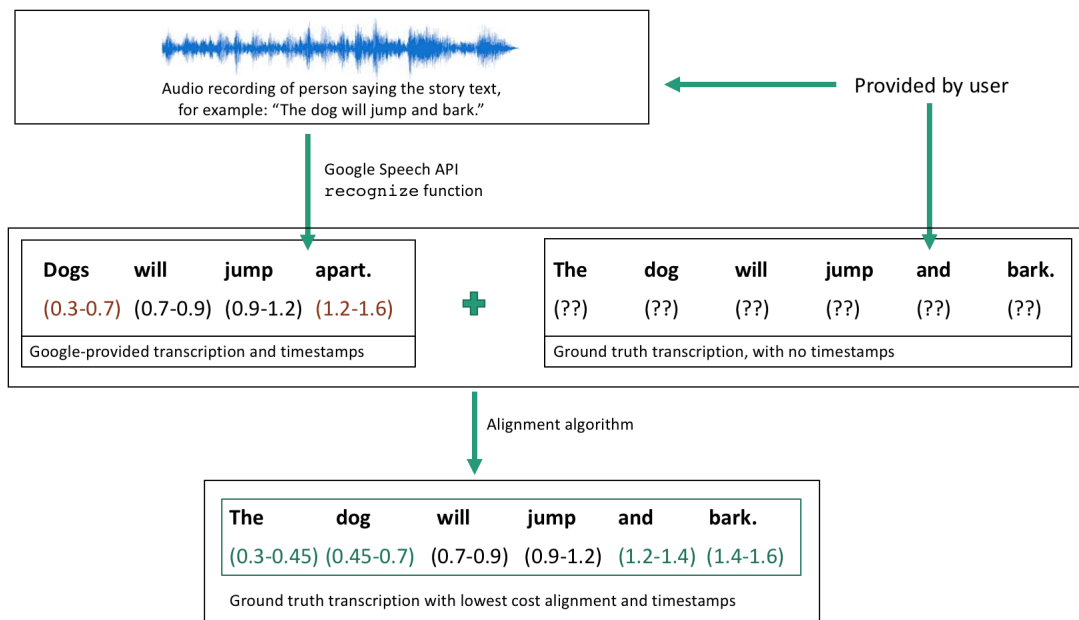


Figure 3-4: A simple example of the method used to obtain timestamps for audio files of people speaking the story text. The alignment algorithm provides the most accurate timestamps possible for the ground truth transcription by using a potentially faulty transcription and timestamps provided from Google Speech API. Often, the errors are more complicated, and cannot be solved by simply splitting up the first and last timestamps to accommodate extra words in the ground truth transcription. The alignment algorithm is a dynamic programming algorithm that finds the best alignment regardless of where the errors occur.

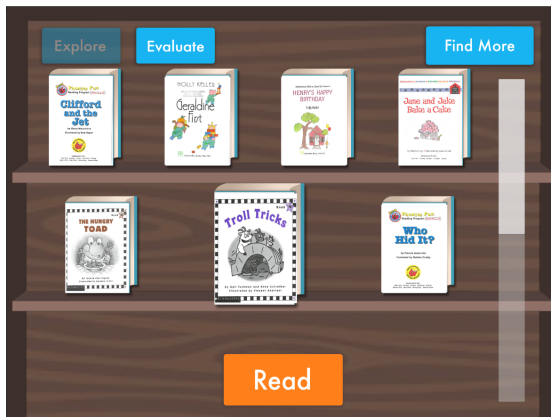
stem, the cost is considered to be 0. If two words are homonyms, the cost is also considered to be 0. If neither of those conditions is met, the cost is the Hamming distance between the two words. Using this word-word cost function, the algorithm determines a sequence of operations that transforms the Google transcription into the ground truth transcription word by word. These operations are one of Keep, Delete, Insert. The algorithm then uses this sequence of operations to re-proportion the timestamps accordingly.

3.4 Tablet Features and Design

The design of the storybook tablet app emulates an open printed book, and uses bright colors with high contrast to appeal to children. The app's main view is a familiar library view, with each of the story's title page visible on each library book. This mimics the intuitive and common main view of other reading apps such as iBook on the iPhone. The navigation is intuitive and the number of buttons is minimal to avoid distraction and confusion while reading. In the reader view, the text is displayed with one sentence per line, making it easy for children to process the words. This was achieved through a recursive placement algorithm that is run every time the child navigates to a new page.

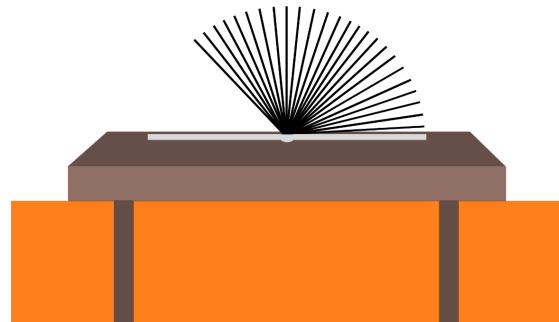
The tablet application was developed using the Unity game framework, in C#. The app is intended to be deployed on Android tablets. Different views of the app are shown in Figure 3-5. Children can select a story from the library view, and progress through the pages of the book with easy navigation buttons. Each page consists of a layout containing an image and text. There are latent bounding boxes around regions in the image that correspond to keywords in the text. These boxes are not visible until triggered. Each word of the text is overlaid on a button, and clicking on a word in the text will trigger the temporary appearance of the bounding box around the corresponding object in the image. The purpose of this feature is to help children associate images with words. The child can swipe on a particular sentence to autoplay the audio for that sentence, and can press the play/pause button to start the audio from the beginning of the story. As the audio is playing, the words that are being spoken are highlighted, so that the child can associate the spoken words with the written words.

Features Used Only in Reading Interaction Everything described so far can be done on the tablet as a standalone app. Another important feature of the tablet app is its ability



(a) Library view with story selected

Loading...



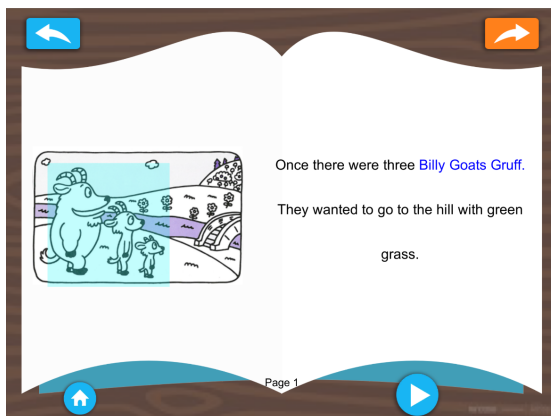
(b) Loading screen featuring a book's pages flipped open



(c) Title page when story assets loaded



(d) Audio autoplaying with words highlighting as they are spoken



(e) Tapping on the image highlights objects and associated words



(f) "The End" page with confetti to congratulate child on finishing

Figure 3-5: Various views of the storybook app, sized in this example for a Google Nexus tablet, 2048 by 1536 pixels. The tablet features a scrollable library view for story selection, a title page and the end page, and an interactive reader page.

to record audio that the child has spoken, and process it for pronunciation evaluation. The evaluation is done via a third party service called SpeechACE². The tablet records audio snippets and sends them to SpeechACE along with the expected text of the snippet. For this reason, speech evaluation is done on a per sentence basis, so that the expected text is known. In evaluate mode, the recording is always active when sentences are shown to the child one by one, and when the child presses a button to see the next sentence, that button press also functions as a notification that the child is done reading the current sentence, so the recording is spliced at that time and streamed to SpeechACE. SpeechACE returns an HTTP response with a JSON body, and this JSON body provides pronunciation scores at the sentence, word, and phoneme level. The tablet app also runs a web client so that it can communicate with other components of the reading interaction system over a ROS network. The tablet app registers a set of message handlers that push tasks onto a main task queue in response to receiving commands over ROS. This is explained further in Chapter 5.

3.4.1 Dynamic Asset Loading and Templatization

As mentioned previously, a key feature of the tablet app is that it is a templated framework, with no code specific to any particular story. This is a huge advantage over previous systems such as the TinkRBook, where every story requires custom control logic. The framework can load an interactive scene for any page that is described in the Page JSON format, turning the problem of creating interactive storybooks from one of hardcoding behaviors for each page to one of generating JSON that the app automatically processes and uses to create the scene. This behavior is made possible by the creation of a page template with the necessary page elements, such as a location for the image and the text, as well as navigation buttons. There are also templates for common storybook features such as tappable text, tappable regions to overlay on top of the image, and rows of text that supported swipe-to-play-audio. On page navigation events, the templates are automatically loaded with the information for the new page.

On app load, the app uses an Amazon S3 client to download the StoryMetadata and title page of each story to display in the library. The rest of the story assets are not downloaded until they are needed, which is when a user selects a story to read. At that point, the S3 client downloads all of the Page JSON files and for each page also downloads the necessary image

²SpeechACE: <https://www.speechace.com/>

and audio assets, which are stored locally. As the user progresses through the story, the template is loaded with the appropriate assets for the current page of the story. Connections between words and objects are also generated based on the information in the Page JSON files. Changes to the JSON files are reflected in the app upon restart, without any software update to the app itself. This is advantageous because the system is agnostic to the method through which the content of the stories is generated.

The next chapter discusses the authoring interface, which provides a way for anyone to create and upload content that can be read in the storybook tablet app.

Chapter 4

Authoring Interface

One major contribution to the novelty of the interactive storybooks presented in this thesis is the accompanying authoring interface that enables parents, educators, or anyone else to create stories that are readable on the tablet app described in the previous chapter. This chapter explains the technical design, features, and layout of the authoring interface, as well as how it fits into the larger system.

A common barrier to the adoption of new technological platforms is a lack of content. This is why technology companies such as Apple and Google both urge developers to create apps for iOS and Android respectively, and why home agents such as Google Home, Jibo, and Amazon Echo all support user-created "skills". Along a similar vein, in order to increase the appeal and usability of the interactive storybook platform, and democratize the platform for use by everyone, it is important to empower users to generate their own content. This is beneficial because parents can create stories with themes and characters that they know will appeal to their children, and educators can have control over the reading level, target words, and questions that appear in the storybook interaction, to better match the level of students.

4.1 Design and Features

The authoring interface is a web application. One design goal was to make the app intuitive to users, so the interface looks and feels like a simplified version of presentation generating software like Microsoft PowerPoint, which many users are already familiar with. Figures 4-1 and 4-2 show the design of the interface.

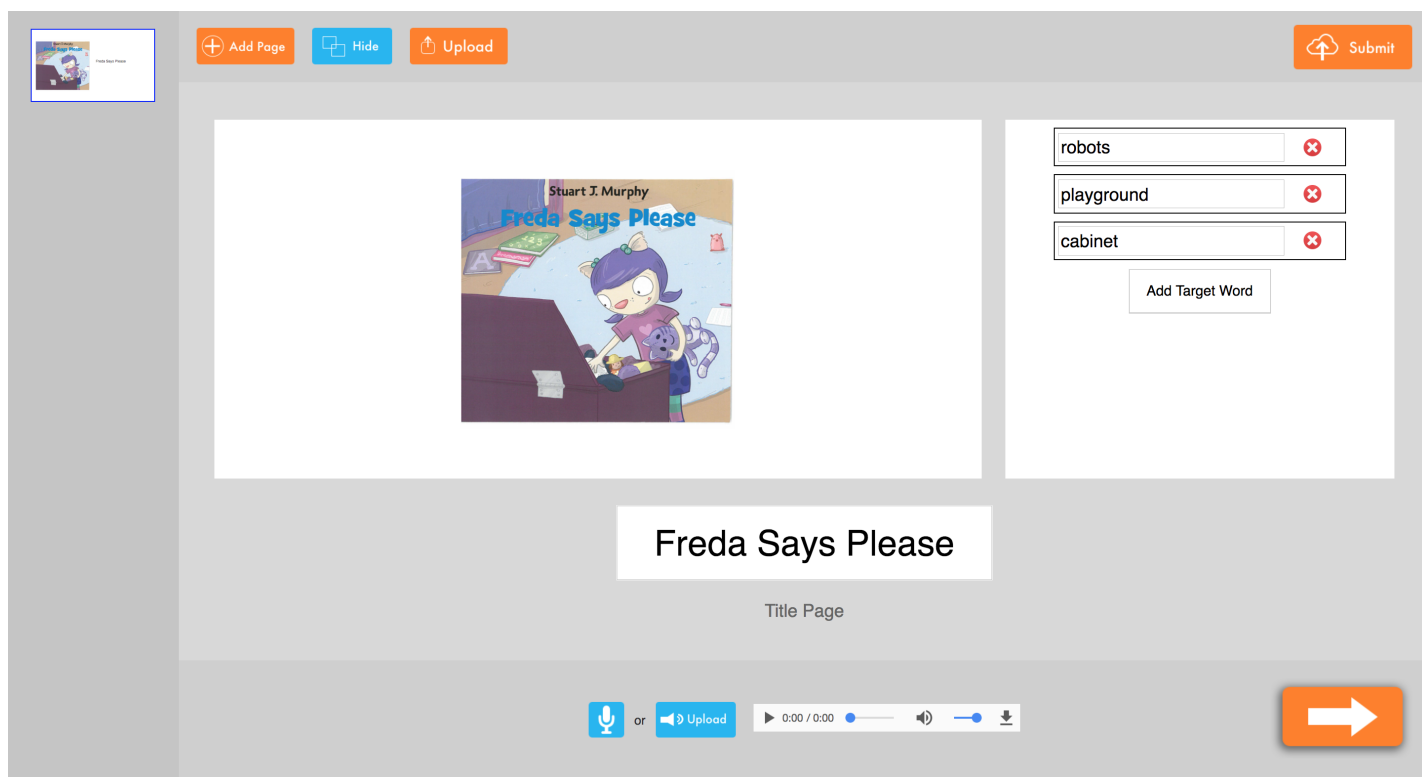
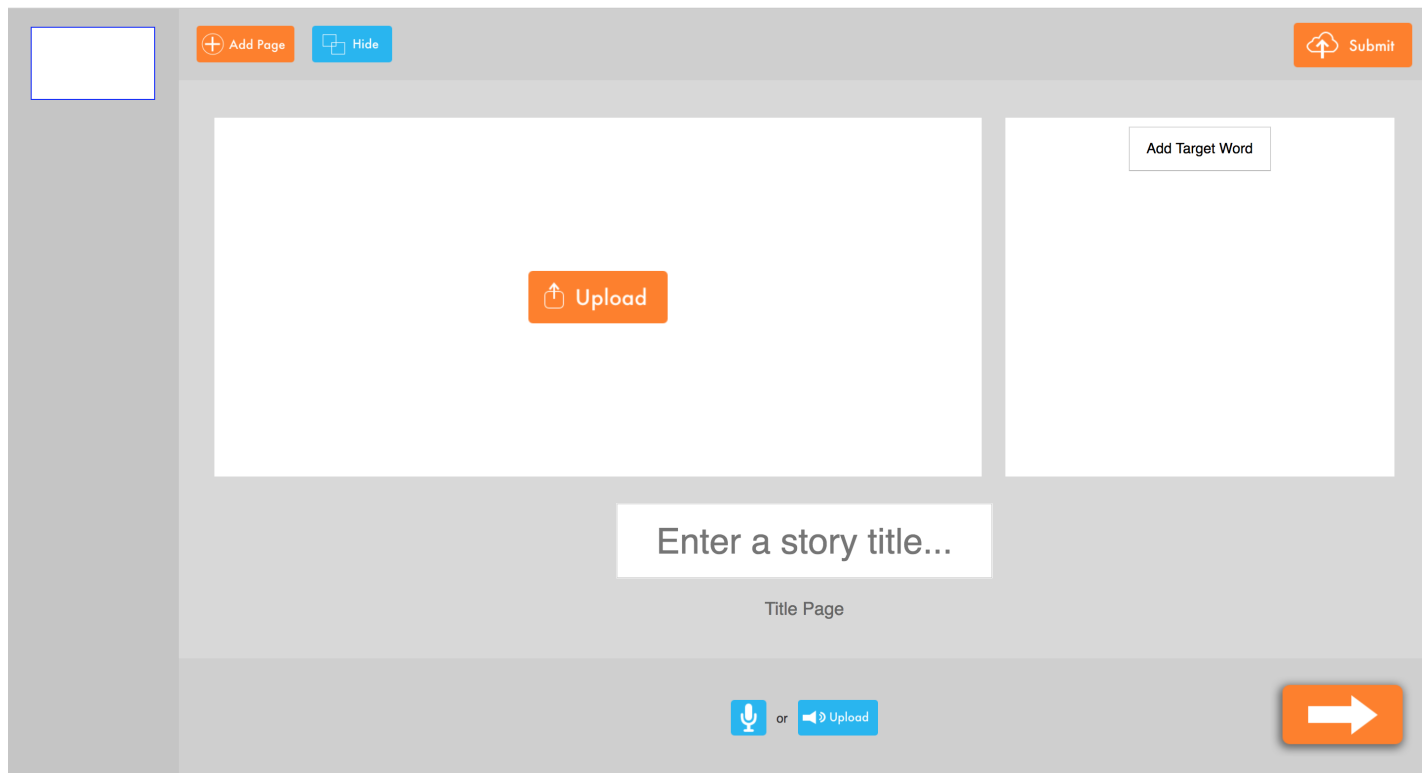


Figure 4-1: A screenshot of the authoring interface web application on the title page. The title page allows users to upload an image, give a title, provide audio of a person speaking the title, and listing the target words of the story.

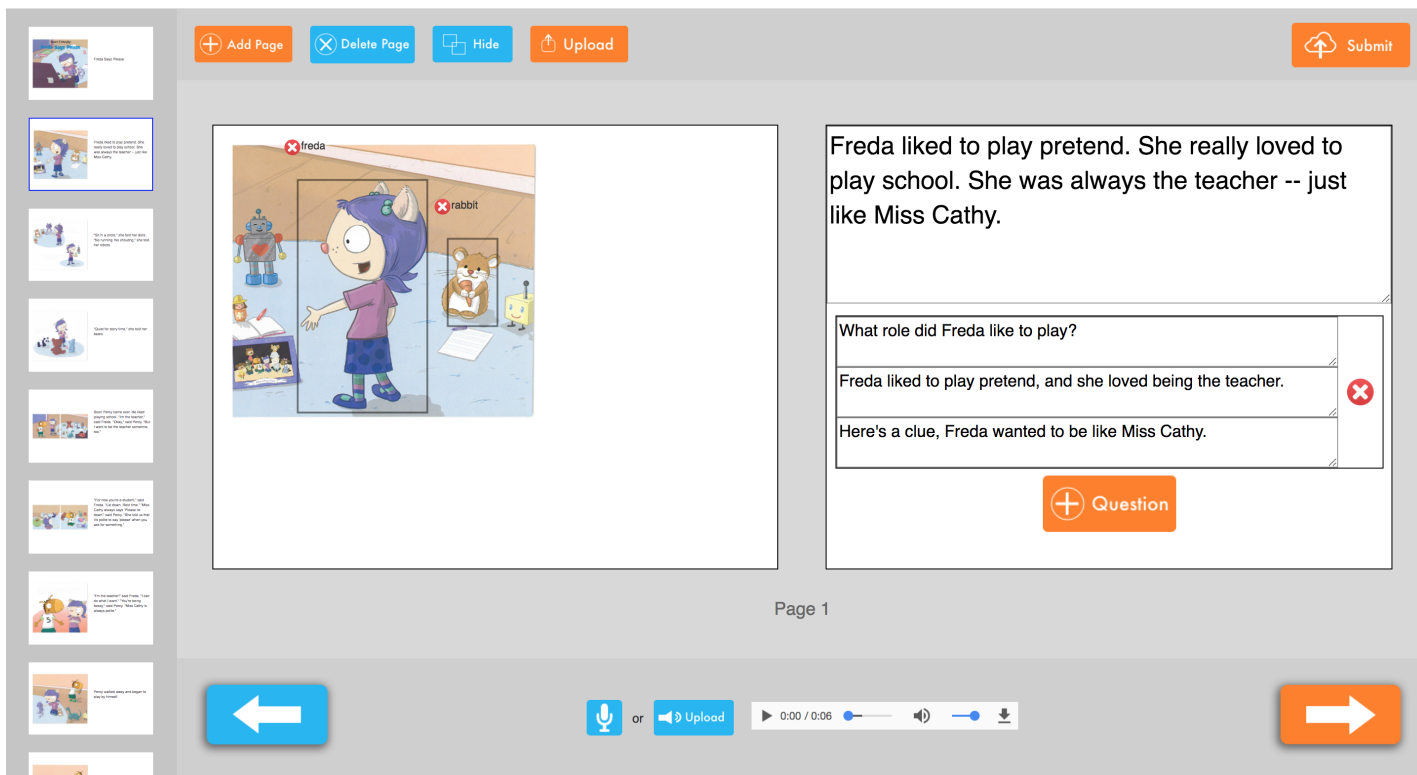
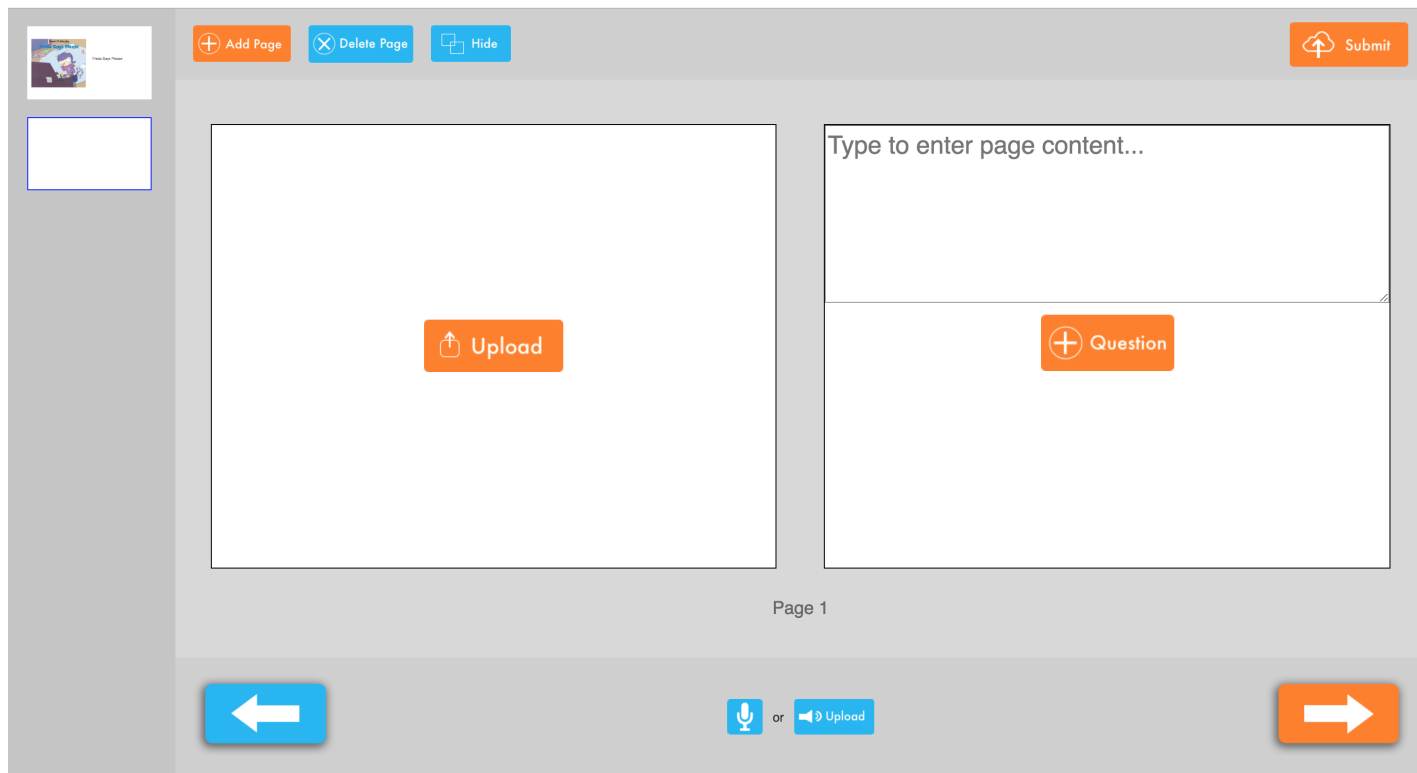


Figure 4-2: A screenshot of the authoring interface web application when creating a typical story page. Creating a page consists of uploading an image, writing text, adding questions and responses for the robot to deliver, recording audio, and labeling objects in the scene.

The authoring interface supports a variety of features. These are described below:

1. Image upload - Click the upload button to upload an image, and re-upload a different image later if desired.
2. Compose story text.
3. Indicate target words - On the title page, fill in the target words for the story, so that the reading interaction controller can emphasize them.
4. Audio record - Simply press the record button to record audio, then press again to stop recording. Or, upload a prerecorded audio file.
5. Audio playback - Listen to the recorded or uploaded audio. If it is unsatisfactory, record or upload again.
6. Object labeling - Drag bounding boxes around objects of interest and provide a label. The labels can appear in the story text or can be outside of the story text.
7. Trigger detection - Objects are automatically associated to words in the text if the labels appear in the text.
8. Enter prompts for Jibo to deliver during the reading interaction for this story. Each prompt is a question, response and optional hint.
9. Automatic timestamping of audio - The timestamping occurs using the same algorithm from Section 3.3
10. Easy integration with tablet app - Press a single button, "Submit," when finished, and the story is immediately available on the tablet app.

These features can be seen in action in Figure 4-2. A discussion of the usability of the authoring interface, including which features users found interesting and easy to use vs. uninteresting and difficult to use, is presented later in Chapter 7.

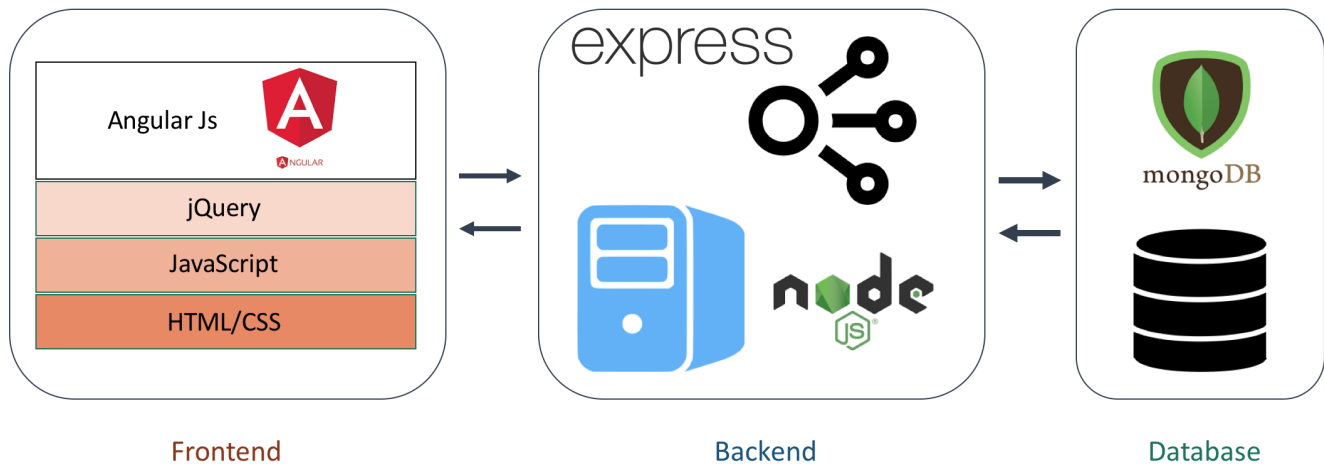


Figure 4-3: A simple diagram of the authoring interface stack. The authoring interface uses AngularJs and Node.js, as well as Express middleware for routing. The database is MongoDB.

4.2 Stack

From a technical perspective, the authoring interface is a single-view web app whose server is written with Node.js¹ and whose frontend client is a browser app written using AngularJs². Image and audio files that the user provides are stored locally on the server, and then pushed to Amazon S3 when the user is done creating the story. Figure 4-3 shows the system diagram and stack for the authoring interface.

4.3 Created Stories Easily and Immediately Transferred to Tablet App

One advantage of this system is that it requires no programming knowledge from the user to transfer a story from the authoring interface to the storybook app. Users can simply click "Submit" on the authoring interface, and "Find More" on the tablet app, and the new book will be downloaded to the tablet. This is described graphically in Figure 4-4. Blue arrows indicate results of user events, and black arrows depict the flow of assets throughout the system, from the user's computer storage, to the client, to the server, to the cloud, and finally to the tablet.

¹Node.js: <https://nodejs.org/en/>

²AngularJs: <https://angular.io/>

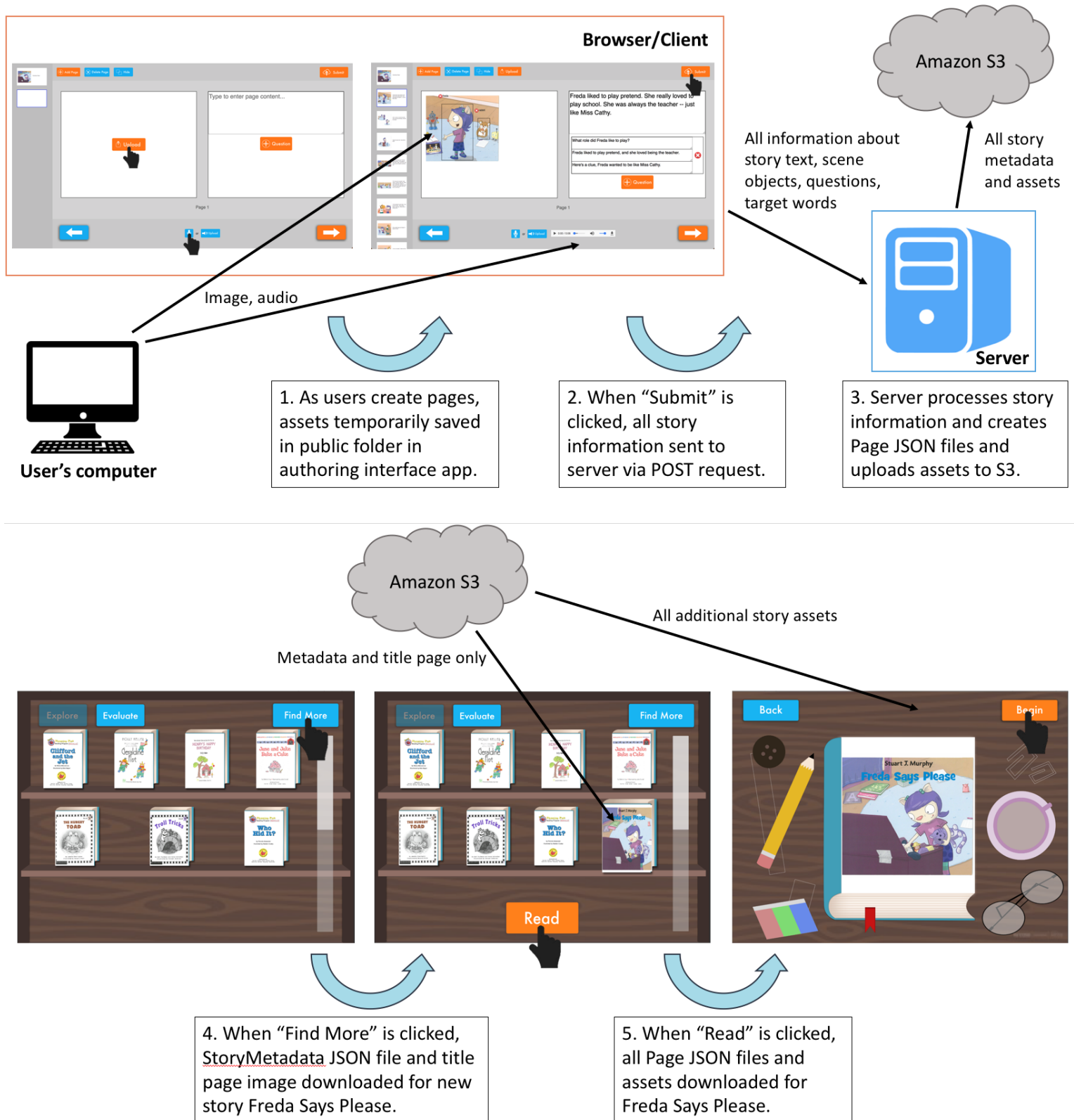


Figure 4-4: A graphical depiction of how story assets flow from the user's computer all the way to the tablet app. Blue arrows indicate results of user events, such as when users press buttons, and black arrows depict the flow of assets and information. The top half figure shows how assets and other story information are pushed to Amazon cloud storage (S3), and the bottom half of figure shows how those assets are then downloaded to the tablet.

More specifically, when the user presses "Submit", the client sends a POST request to the server with the information for each page that the user created, along with the target words for the story. The server processes this information, generating timestamps, uploading audio and image assets to S3 following a standard naming pattern, and generating triggers (which are associations between scene objects and words in the text). The server also creates JSON files representing each page, and uploads them to S3. Finally, the server uploads a StoryMetadata JSON file to S3. Later, when the "Find More" button is pressed on the tablet app, the tablet app searches the known S3 directory for new StoryMetadata files, and downloads them. These stories then appear in the story library and can be selected for reading. Recall that only when a story is selected for reading are all of its assets downloaded.

The authoring interface and tablet app together form a working system for simple story authoring and interactive display. The next chapters focus on the process of developing and testing an interactive reading experience with a robot, using the tablet app as a shared medium between the robot and the child.

Chapter 5

Reading Interaction with a Robot: Design and Implementation

5.1 System Overview

With a platform in place for creating stories and displaying them on the tablet app, the next major step was to create an interaction between a child, the tablet app, and Jibo, that would engage, teach and evaluate the child. This chapter presents the technical system design of the reading interaction.

There are three large components that comprise the reading interaction system. These are: the tablet application, the Jibo robot, and a controller program. The controller is a Python script that runs on a separate machine. In the interaction setup used in user studies, the controller runs in an Ubuntu 16.04 virtual machine hosted on a MacBook, and the storybook application runs on a Huawei MediaPad M3 tablet. The three components of the system communicate over ROS, the Robot Operating System ¹, which serves as a publish/subscribe message passing layer. A simple diagram of the interaction setup is given in Figure 5-1.

5.2 Robot Agent: Jibo

The robot used in this interaction is Jibo, an 11 inch tall robot with a stationary base that can display a variety of emotions and gestures with his swiveling body and expressive

¹ROS: <http://www.ros.org/>

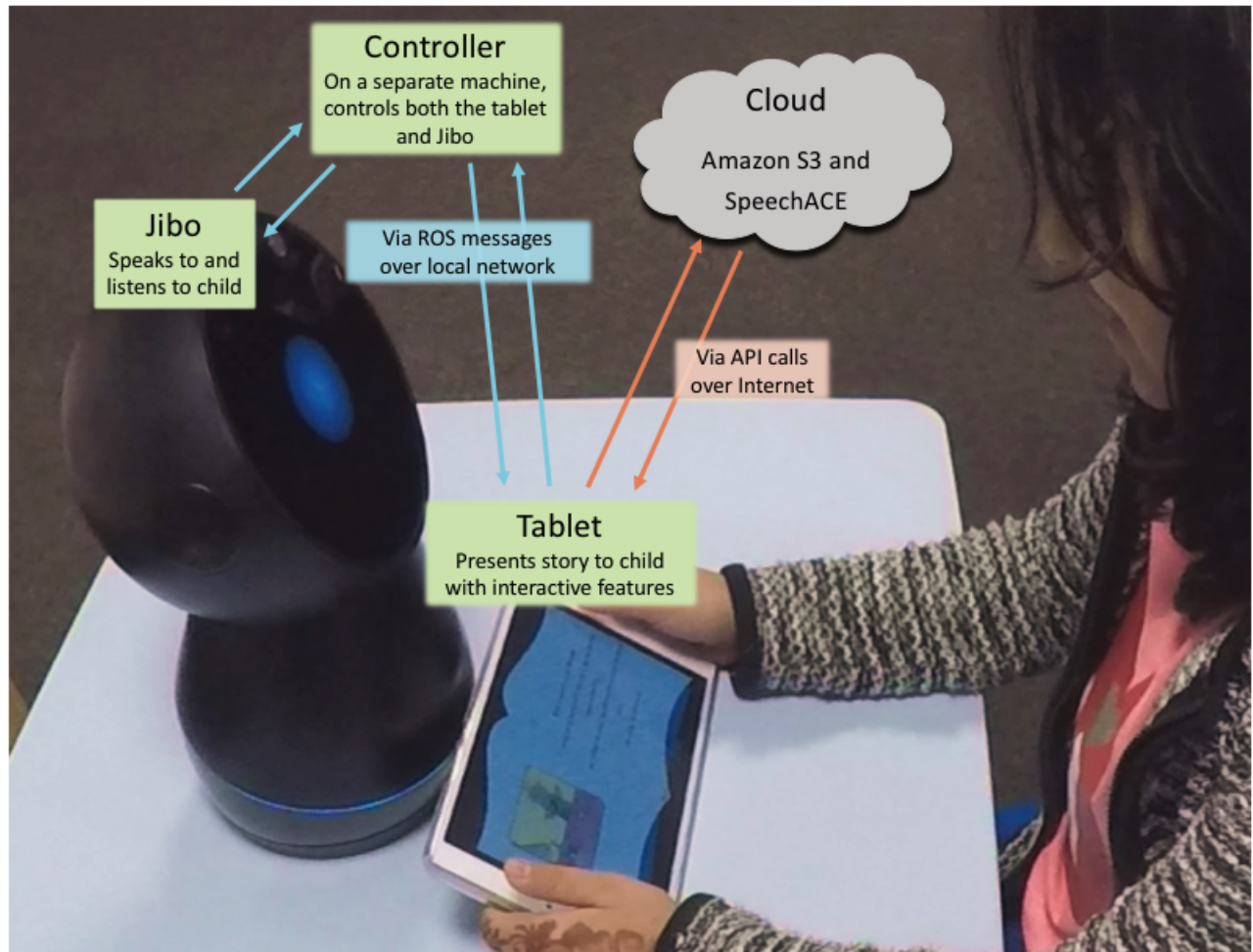


Figure 5-1: The three components of the reading interaction: the controller, tablet, and Jibo. The components communicate using ROS on a local network. The controller is finite state machine that sends messages over ROS via Rosbridge to control the reading interaction by sending commands and receiving inputs from Jibo and the tablet. Note that Jibo and the tablet do not communicate directly. The tablet makes API calls to Amazon S3 and SpeechACE in the cloud to download assets and upload recorded audio for evaluation.



Figure 5-2: Jibo robots, in both color varieties.

screen that takes the place of a face. Jibo is connected to the Internet and provides useful features such as automatic speech recognition (ASR), face recognition, an animation engine, and easily controllable text-to-speech (TTS). Jibo is also programmable with custom skills developers can write. In the interaction setup, Jibo is used primarily for ASR, TTS and animation. Our lab has written a custom skill for Jibo that turns Jibo into a ROS node, so that it can communicate with other components as described in the diagram above. More specifically, Jibo subscribes to and publishes messages over a simple web socket so that it can report information such as ASR results, and respond to commands to deliver speech or play animations.

5.3 Communication Among Components

The system is set up over a ROS network. The network consists of nodes and topics. Each of the three major components is a node in the network, and each type of message corresponds to a topic. An overview of the nodes and the topics each component subscribes and publishes to can be found in Figure 5-3.

The ROS core node runs on a separate machine from Jibo and the tablet, but on the same local network. Since neither the tablet nor Jibo is a native ROS program, the machine running ROS core also runs Rosbridge, which exposes a JSON API to ROS functionality via a websocket server. Both the tablet app and Jibo use web clients to send and receive messages via the Rosbridge to the ROS core. In order to support this functionality on Jibo,

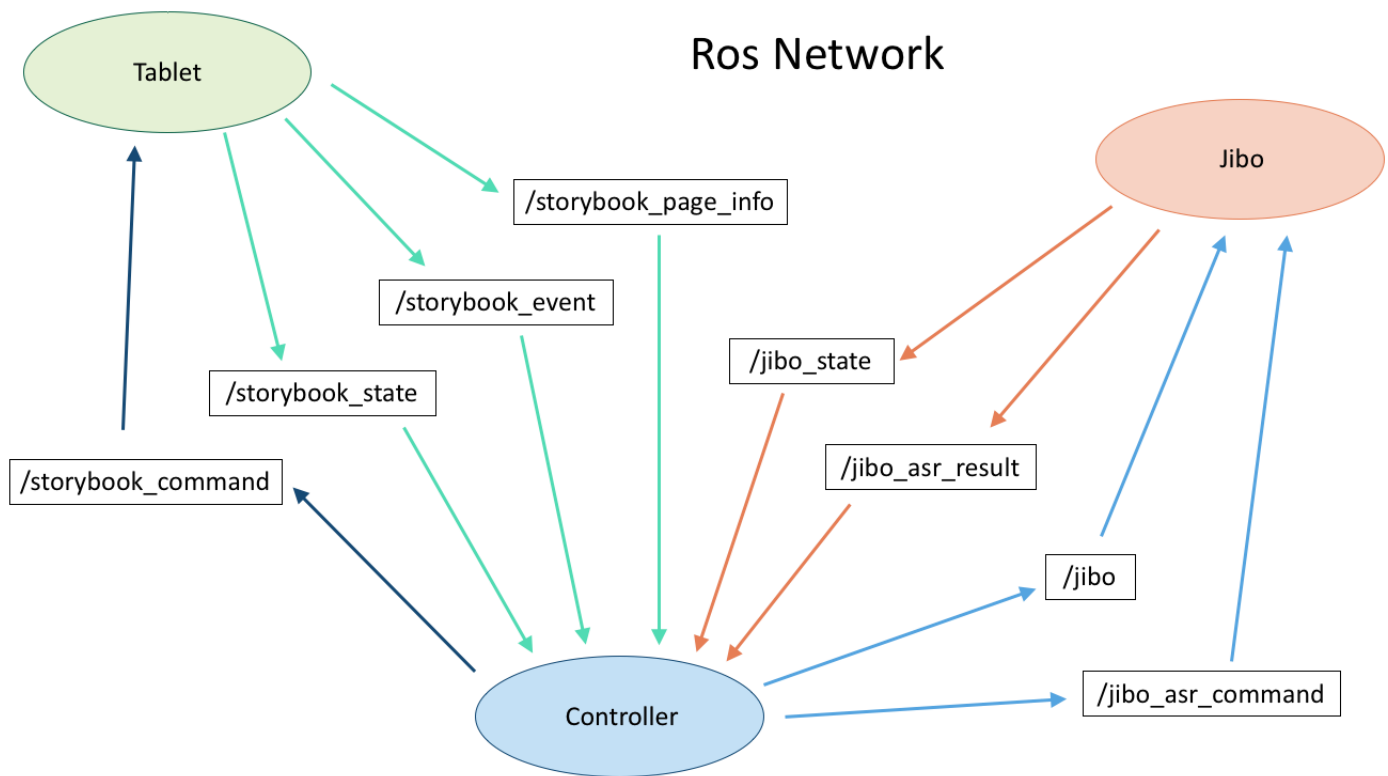


Figure 5-3: There are three ROS nodes, including the ROS core node, which is where the controller resides. All topics are either subscribed to or published to by the controller; in other words, the tablet and Jibo do not directly communicate with each other. Arrows indicate the flow of messages on the topic.

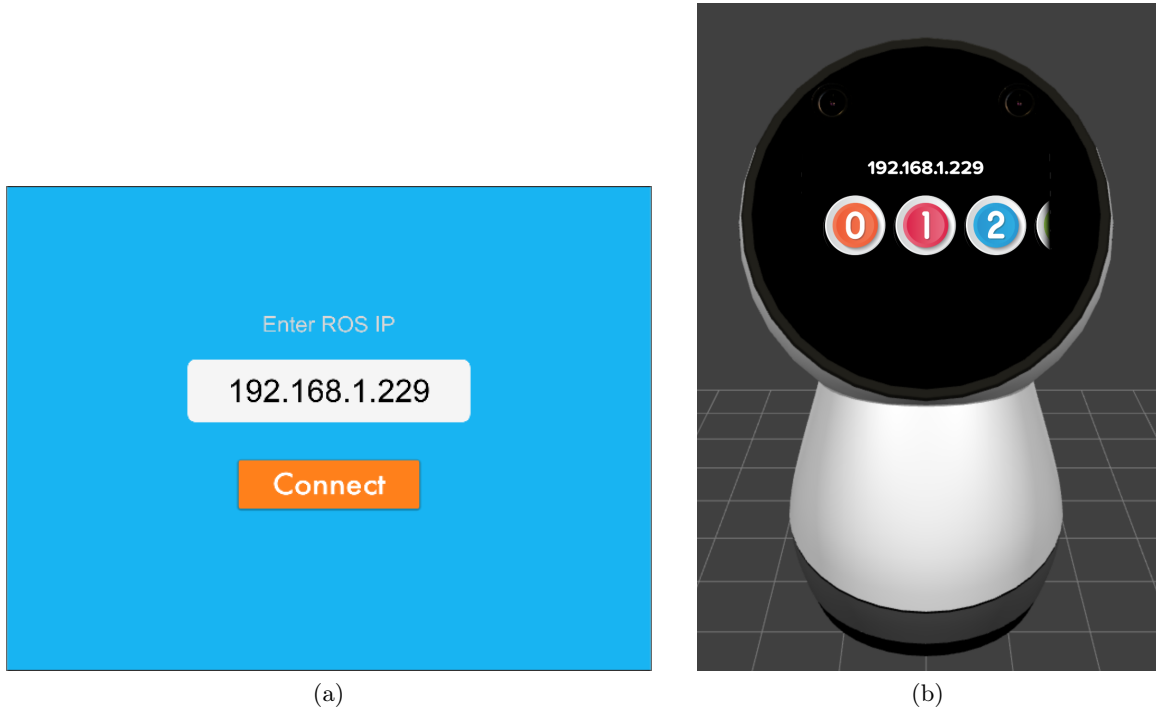


Figure 5-4: Left: ROS IP selection screen on tablet app. Right: ROS IP selection screen on Jibo.

we wrote a custom skill that turns Jibo into a ROS node by subscribing and publishing to topics and opening a web socket connection to the ROS core. The tablet also needs to connect to the ROS core. The ROS core has an IP address on the local network, usually something like "192.168.1.229," which the user needs to enter on either the tablet start screen or the Jibo start screen. Once the ROS IP address has been entered once, the system remembers it so the user does not need to enter it again. The user instead needs to confirm via a single button press that the ROS core IP address is still the same every time the tablet app or Jibo is restarted.

The decision to set up the ROS network over a local network instead of over the Internet was made to reduce latency reduce the chance of network failures. However, it is important to note that there is nothing preventing the system from being adapted to have the controller reside in the cloud and communicate with tablets in distant locations. This flexibility could prove useful in future deployments of the interaction.

Every topic has the controller as either a publisher or a subscriber. The tablet app and Jibo do not communicate directly with each other. The controller spins up a separate thread to handle each of the topics it subscribes to, and forces serialization of message

handling by having each separate message thread push tasks onto a main task queue, which is pulled from in the main thread. There are no concurrency issues during the course of the interaction due to the structure of the finite state machine. In other words, the controller's behavior is extremely restricted, and it will only respond to certain messages depending on its state, so messages that are delayed or arrive late can be ignored. The messages sent from the controller are commands, which the other components will respond to. For example, the controller might send text for Jibo to say, such as "Can you pronounce the blue word for me?", while at the same time sending a command to the storybook app to highlight a particular word using blue font color. Messages sent to the controller are status updates, event reports, and information logs, which the controller uses to manage the state of the system and make decisions. For example, Jibo might report whether or not it has completed delivering the requested speech, and the tablet might report that the child has moved on to the next page of the story. A chart of all of the messages along with a brief description of each is presented in Table 5.1.

5.4 Two Modes of the Interaction

During the interaction, a story can be read in two modes, either the evaluate mode or the explore mode. In the evaluate mode, Jibo is leading the interaction. On each page, the sentences are shown one by one, and the child must read each sentence and tap a button to move on to the next sentence. If the child gets stuck, she can ask Jibo for help, and Jibo will read the sentence then ask her to repeat it. The audio stream for each sentence is recorded and sent asynchronously to a service called SpeechACE for evaluation. The results of the evaluation are then used to inform Jibo's possible actions. At the end of each page, Jibo will ask the child questions, or simply move on to the next page. This question asking mechanism is a core part of the learning experience, and is explained in more detail in a section below. This process repeats for each page of the story. The goal of this mode is to teach the child new words and evaluate their reading.

In explore mode, the child is free to read the story at her own pace, with little interruption or comment from Jibo. The child can interact with the features on the tablet, and can tap on words or objects on the page to ask Jibo what they are. The explore mode is primarily for children to listen to the audio, clear up confusion about words they don't know, and

Topic	Publisher	Subscriber	Description
storybook_command	Controller	Tablet	Command tablet to perform actions like highlight certain words or begin and end recording.
storybook_state	Tablet	Controller	Published by tablet constantly at 5Hz to report state information like which sentence the tablet is evaluating.
storybook_event	Tablet	Controller	Published by tablet to inform controller about user actions like taps, swipes, story selection, story navigation.
storybook_page_info	Tablet	Controller	Published when the storybook changes pages, to inform controller what's on the current page.
jibo	Controller	Jibo	Command Jibo to say things via TTS or to play animations.
jibo_state	Jibo	Controller	Published by Jibo constantly at 10Hz to report state information like whether Jibo is speaking or playing an animation.
jibo_asr_command	Controller	Jibo	Command Jibo to start and stop listening for and publishing ASR results.
jibo_asr_result	Jibo	Controller	The results of Jibo's ASR, including transcription and confidence.

Table 5.1: A description of each topic in the ROS network. Each topic is subscribed to or published to by the controller, since the tablet and Jibo do not communicate directly.

be free from any pressure to answer questions. The child can navigate both forwards and backwards through the story (whereas in evaluate mode the story always progresses forward and does not return to previous pages). This mode is suitable for very young children, since even if they cannot read well, they can listen to the story and learn to associate spoken words with written words, through the time-stamped text highlighting.

5.5 Controller and Finite State Machine

The bulk of decision making and logic is contained within the controller. This controller lives on the same Ubuntu virtual machine that runs the ROS core node. The controller is responsible for managing the state of the system, tracking the child's progress, making decisions about what actions Jibo should take, and orchestrating the activation of UI elements on the tablet. The controller is a finite state machine, written using the Python transitions library ². The controller is itself a ROS node, so it communicates with the other components by publishing and subscribing to topics. The state machine structure, comprised of its states and transitions, is shown in Figure 5-5. What is not depicted in the figure is the actions that are triggered on every transition. The following two subsections describe a typical interaction in both evaluate mode and explore mode, providing a more sequential explanation of the state transitions.

5.5.1 Evaluate Mode State Machine

Once the tablet switches to evaluate mode, the state machine controller's state becomes "BEGIN_EVALUATE". After a story has been selected and the title page is shown, Jibo greets the child and begins engaging the child in the reading interaction. Jibo says "I'm so excited. This is my favorite storybook, will you read it to me?". After Jibo finishes speaking, the controller changes state to "WAITING_FOR_NEXT_PAGE", and the controller commands the tablet to navigate to the next page. When this new page's information is relayed back to the controller, the controller instructs the table to show the first sentence and begin audio recording, and changes state to "WAITING_FOR_CHILD_AUDIO". The child is then expected to read the sentence to the best of her ability. After the child is done reading, she presses the "next" button, which stops the audio recording, and informs the

²Python transitions library: <https://github.com/pytransitions/transitions>

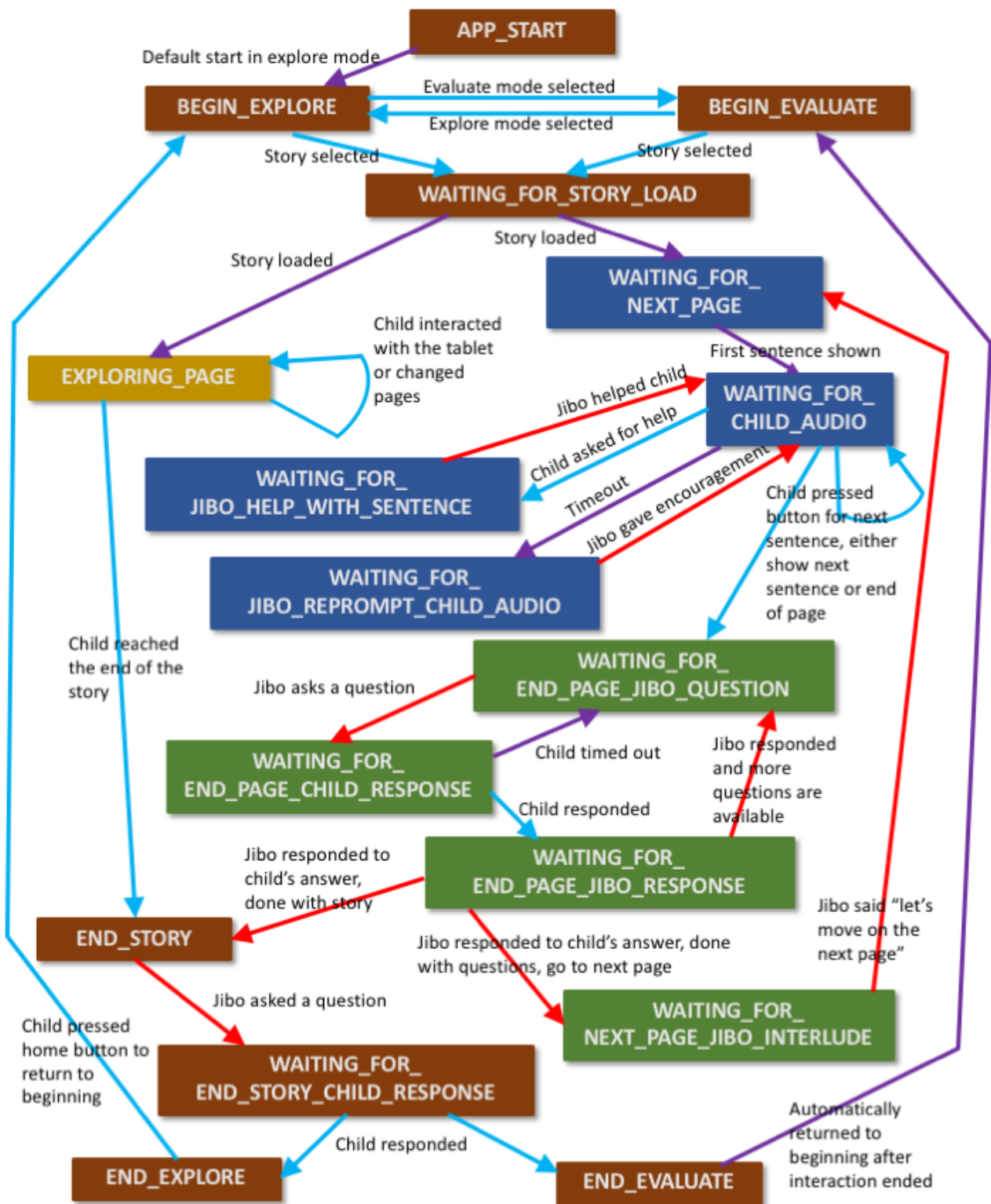


Figure 5-5: This is an overview of the finite state machine that constitutes the high level logic of the controller. The states are color coded into four groups. The brown states are states that are shared by both explore and evaluate mode. The yellow states are those that only exist in explore mode. The green and blue states only exist in evaluate mode. The four blue states control the logic of showing sentences when a new page is loaded and prompting the child to read and providing help if the child asks for it. The four green states control the logic of Jibo asking questions at the end of a page, and handling the child's responses. The transitions are also color-coded. Blue arrows transitions that result from child input (either speech or tactile input on the tablet). Red arrows indicate transitions result from Jibo actions (such as completing a TTS command to deliver speech). Purple arrows indicate transitions that happen without any input, either due to default behavior or timeouts on child responses.

controller. While the controller is in the "WAITING_FOR_CHILD_AUDIO" state, if the child doesn't start speaking for a long (configurable) time, e.g. 15 seconds, Jibo will encourage the child to try to read the sentence. If the child wants help with the sentence, she can say "I need help" or "Can you read it to me?" and then the controller will change state for "WAITING_FOR_JIBO_HELP_WITH_SENTENCE" and Jibo will read the sentence and ask the child to repeat it. After "next" is pressed, the tablet sends the current audio snippet to SpeechACE, and then the controller either instructs the storybook to show the next sentence if one is available (and repeat the audio recording process), or provides an opportunity for Jibo to ask a question if it's the end of the page. When the SpeechACE results come back to the tablet, the tablet relays them to the controller.

In the case where the child has reached the end of a page, the controller switches into "WAITING_FOR_END_PAGE_JIBO_QUESTION" and Jibo has the opportunity to ask questions. After Jibo's prompt has been delivered and the storybook has completed any requested highlighting of text or objects on the page, the controller switches to "WAITING_FOR_END_PAGE_CHILD_RESPONSE". At this point, a valid action in response to the question will trigger the system to switch back to "WAITING_FOR_NEXT_PAGE" state and instruct the tablet to move to the next page, which will send the system back to the state progression described in previous paragraph. An example of a valid action is the child tapping on a word in the text in response to being asked "Can you tap the word 'donned' for me?", or the child speaking in response to being asked "Can you explain what a 'spatula' is?". Actions that are not valid based on the type of question, such as if the child taps on an object when asked to tap on a word, are simply ignored.

When the storybook is on the final page, instead of transitioning to "WAITING_FOR_NEXT_PAGE" after the "next page" button is clicked, the controller will transition to "END_STORY", where it will tell Jibo to say some concluding words about the story, ask another question, and then end the interaction. There are a predetermined set of concluding sentences Jibo can deliver for each story, such as "That was a great story, I'm so happy the toad learned his lesson!" or "Wow, I wish my birthday could be as fun as Henry's. Do you think Henry's birthday party seemed fun?". After the question is asked, the controller switches to "WAITING_FOR_END_STORY_CHILD_RESPONSE", and after the child responds, Jibo gives some positive acknowledgement and then the controller commands the storybook to return to the library view, ending the reading interaction.

5.5.2 Explore Mode State Machine

The state machine for explore mode is much simpler. When explore mode is selected on the tablet app, the controller switches to the "BEGIN_EXPLORE" state. After a story is selected, the controller changes state to "EXPLORING_PAGE". When the child navigates to a different page, the controller updates its knowledge of the current page but does not change state; it stays in "EXPLORING_PAGE". The child can swipe sentences to play audio, and can tap on words or objects. When a tap is received, the controller sends a message to Jibo telling Jibo to pronounce the word or object label, to teach the child. The child is free to explore the page, relistening to audio, attempting to read the text, and tapping on objects or words that are interesting or unfamiliar. At the end of the story, the behavior is the same as in evaluate mode, and the controller transitions to "END_STORY".

5.6 Deciding What Questions Jibo Should Ask

A major component of the controller is a student model that stores pronunciation data streamed in from SpeechACE results, and then is able to generate questions and queue them up for the controller to tell Jibo to deliver. The questions fall into one of five categories.

1. Single word pronunciation
2. Full sentence pronunciation
3. Tap on a word
4. Tap on an object in the picture
5. Open ended question

The student model is a module in the controller that provides questions drawn from predetermined questions and auto-generated questions. The predetermined questions are open ended questions that ask the child to explain a particular target word (e.g. "Do you know what helium is? Can you explain it to me?") or provide a response to a reading comprehension task (e.g. "How do you think the doctor feels right now?"). These questions come from the author of the story, via the authoring interface. Jibo's responses to the child's answers to these questions are also predetermined, and will often include the correct answer,

so that if the child was correct, Jibo’s response reinforces the child’s knowledge, and if the child was wrong, Jibo can correct the child.

The auto-generated questions are created based on the child’s pronunciation attempts while reading. Every time the child speaks a sentence, the audio is recorded and sent to a service called SpeechACE along with the intended sentence text. SpeechACE returns a JSON response with pronunciation quality scores for the entire sentence, and quality scores for each word, and for each phoneme within each word. The tablet receives this JSON response and sends it to the controller over ROS. A full sample SpeechACE response can be found in their API documentation. The student model receives this information and keeps track of all pronunciation scores throughout the entire interaction. When the student model selects which word to ask a question about, it searches among the target words on the current page to find the one with the lowest history of scores. Once the student model selects a word, it either asks the child to pronounce the word or tap on the word, with equal probability. If there are no words with any scores, it selects a target word on the page at random. If there are no target words on the page, it looks at the available scene objects and asks the child to tap on one of them. The student model also produces questions that require the child to reread an entire sentence if SpeechACE reports that the child performed below a certain threshold on the entire sentence. This threshold was chosen as 0.3 out of 1 after empirical testing. When the child answers a question requiring her to pronounce a word, tap a word, or tap an object, Jibo can evaluate the correctness of the child’s answer and respond accordingly, in contrast to the open ended questions in which Jibo’s response is predetermined.

The framework for actually delivering questions and evaluating responses is implemented in a way that makes it easy to incorporate more question types. Each type of question is implemented as a class that extends a base `EndPageQuestion` class, and implementing a new question type involves implementing methods for asking the question, providing a hint, checking if a provided response is correct, and responding to the child. The timing of question asking, detection of when the child has attempted a response, and timing of the delivery of hints and responses is all centralized in the controller state machine logic and does not have to be reasoned about when adding a new question type.

One interesting problem encountered with question generation is that because the SpeechACE results come in asynchronously, the student model often doesn’t have enough information to

generate all the questions immediately when the child ends the page. The solution to this is to split the question generation mechanism into two phases whenever possible. If there are predetermined questions that Jibo can deliver, those questions are delivered first and then after that question is finished, the student model tries to generate questions again to see if new SpeechACE results result in new questions being generated. If there are no predetermined questions for Jibo to deliver, then the student model first generates questions based on the results it has seen, and then when the SpeechACE results from the latest sentence are received, the student model attempts to generate questions given those new results, making sure not to duplicate any questions that were already delivered on the current page.

At the end of the story, the question that Jibo asks the child is a generic question, such as "What was your favorite part of the story?". Jibo then comments positively on the child's response, whatever it may be. This ends the reading interaction for that story, and the app returns to the library view. The child can then select another story to read in either evaluate or explore mode.

This concludes the description of the reading interaction. The next chapter dives into the user study in which children participated in reading interactions with Jibo and parents created stories with the authoring tool.

Chapter 6

User Study

6.1 Study Overview

6.1.1 Participant Information

To evaluate the entire system and reading interaction, a user study was conducted with children and parents from the local Boston area. The target participants in this study were parent-child pairs. The desired child demographic is early reader, which translates to ages 5-8. There is no restriction on the parent demographic. The parent and child participated in different consecutive stages of the study.

Some participants were recruited from a lab mailing list and word of mouth, while others were recruited from two local schools. The parent-child pairs recruited from the mailing list underwent the study in our lab space, while the children recruited from schools underwent the study in a space at the school, without the parent component. A breakdown of all 17 child participants and the 7 parents is below. The first 2 children were pilot testers, and their data is not used. Several participants had a very low reading ability, rendering them unable to complete all portions of the study. These constraints have effects on which data can be used for certain analyses, and this is further elaborated on in the discussion of results in Chapter 7.

6.1.2 Study Stages

The purpose of this qualitative study is to evaluate the interactive storybook with respect to child learning outcomes and child engagement, and to evaluate the accompanying authoring

Id	Gender	Age	In Lab	Evaluation Story	Notes
0	female	7	yes	Henry's Happy Birthday	pilot test
1	female	7	yes	Henry's Happy Birthday	pilot test
2	male	8	yes	Henry's Happy Birthday	
3	female	6	yes	Clifford and the Jet	
4	female	5	yes	The Hungry Toad	
5	female	5	yes	Henry's Happy Birthday	did not complete
6	male	6	yes	Clifford and the Jet	
7	female	6	yes	Clifford and the Jet	
8	male	8	yes	The Hungry Toad	
9	female	6	yes	Clifford and the Jet	
10	male	5	no	N/A	did not complete
11	female	6	no	N/A	did not complete
12	female	6	no	Henry's Happy Birthday	
13	female	8	no	Henry's Happy Birthday	
14	male	7	no	Henry's Happy Birthday	
15	male	7	no	Henry's Happy Birthday	
16	female	7	no	Henry's Happy Birthday	did not complete

Table 6.1: Information about the 17 child participants. There were three possible stories for evaluation mode. Some children did not complete evaluation mode because their reading level was too low for any of the stories to be a good fit for the interaction. The first couple of children were pilot testers and no pre and post test data was collected for them.

interface with respect to usability. The storybook takes the form of a tablet app, run on a Huawei MediaPad M3 2016 tablet¹ for the studies, and can be operated in two modes: evaluate mode and explore mode. The study consisted of these three stages:

- Stage 1: The child participated in an interaction with the storybook (in evaluate mode) and a Jibo robot.
- Stage 2: The parent participated in a story creation experience using the authoring interface.
- Stage 3: The child participated in a second storybook interaction (in explore mode) that used the story created by the parent.

6.2 Study Protocol

For studies conducted in lab, the parent and child were greeted in the lobby and escorted to the lab space. The researcher gave an overview of what the study would entail and how long it would take, and then parent was asked to fill out a consent form. The consent form grants the right to use any collected video and audio data for this research study and potentially in related publications and academic databases. The child was told to sit at a small table in an open lab area. The parent was told to either sit on a couch behind the child or move to another room, so as to avoid distracting the child. Jibo sat on the small table across from the child, and the tablet was placed between Jibo and the child on the table. The researcher sat to the side of the child, about 3 feet away, with a laptop to monitor progress and initiate the study interaction. The components of the system were connected to a MiFi hotspot.

6.2.1 Phase 1: Evaluate Mode

Pre Test and Story Selection

Each child underwent Stage 1 of the study using one of three stories. From most to least difficult, the stories used were Henry’s Happy Birthday, The Hungry Toad, and Clifford and the Jet. Each story had a set of preselected target words, which are shown in Figure 6.2. The stories were provided by education expert Maryanne Wolf, the Director of the Center

¹Huawei M3: <https://consumer.huawei.com/us/tablets/mediapad-m3-8/>

Clifford and the Jet	The Hungry Toad	Henry's Happy Birthday
jet (x2)	soap (x4)	donned (x2)
fog (x2)	throat (x3)	spatula (x2)
cab (x2)	toaster (x3)	unruly (x2)
jog	coat (x3)	zigzag (x2)
	foam	bunting (x2)
	rowboat	helium
	groaned	donkey
		raft
		siren
		miserable
		kite
		plain

Table 6.2: The target words for each of the three evaluation stories for the user study. Each word appears once in the story, unless otherwise indicated.

for Reading and Language Research at Tufts University. The full story texts cannot be released since they are copyrighted.

The pretest consisted of the researcher presenting the child with some paper cards, and asking the child to read the word on each card and explain the meaning of the word as best as possible. The story initially chosen for each child was based on the child's age and grade level, with 5-6 year olds being presented with the simplest story, 7 year olds being presented with the mid-level story, and 8 year olds being presented with the more difficult story. However, if the child scored particularly high or particularly low on the pretest for the assigned story, the researcher switched stories to better match the level of the child. More specifically, if the child knew the pronunciation and meaning of more than 90% of the words already, a harder story was chosen if one was available, and if the child could pronounce fewer than 25% of the words, an easier story was chosen if one was available.

Demonstration of Reading Interaction

After the child completed the pretest, the researcher gave a demonstration of how the reading interaction would proceed. The researcher chose a story different from the one being used for this child. The researcher demonstrated reading each sentence, pressing the button when ready for the next sentence, asking for help on a sentence, answering questions, and saying "I don't know" in response to questions. The researcher also had the child try answering a question from Jibo. After the demonstration, the researcher asked the child if she had



Figure 6-1: pretests and posttests were administered the same way - by having children read and explain words written on paper cards.

any questions about the interaction, and whether she was ready to read her own story. The researcher answered the child’s questions and cleared up any confusion, and then restarted the reading interaction using the story intended for the child.

Reading Interaction

The researcher observed the child as the child proceeded through the reading interaction in evaluate mode. The child would sometimes attempt to converse with the researcher, and the researcher would respond briefly and then tell the child to direct her attention back to the Jibo and the reading task. The researcher intervened if there were technical errors and the system needed to be adjusted or restarted. The researcher also intervened by giving instructions or suggestions if the child was stuck and truly didn’t know how to proceed.

Jibo asked questions to the child, and as described in Section 5.6, these questions were a combination of predetermined open ended questions, and realtime auto-generated questions. The predetermined questions focused on the target words of the story or on understanding character emotions during the story. The predetermined questions for each story are listed in Appendix A.

Dealing with Technical Failures There were times when the system became frozen and intervention was necessary. To support this, a feature was added to save the state

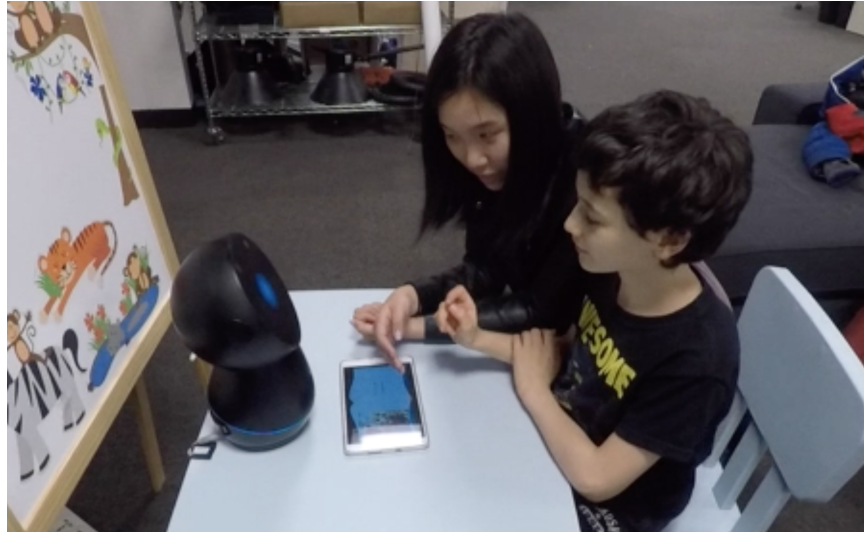


Figure 6-2: The researcher demonstrated the reading interaction for the child, using a story different than the evaluation story of that child, before the child started reading. This was done to familiarize the child with the process and provide an opportunity to clear up any questions or misunderstandings.



(a)



(b)

Figure 6-3: Images of children engaged in the reading task in evaluate mode.

of the controller whenever the controller is shut down. The controller will try to reload the system from that previous saved state on startup, and command the tablet app to pick up where it left off in a story if the user decides to continue the same story that was read in the previous session of the app. Another feature to improve robustness of the system is the retry of essential messages such as TTS commands until they are received. Jibo publishes state messages constantly, and this state includes any audio that Jibo is delivering, so it is possible to detect when Jibo has begun to execute a TTS command. This retry was extremely important when failures were common, particularly in schools where networks were less reliable. There were never any system failures that could not be resolved by restarting the app, but this was done sparingly, due to the delay and the increasing frustration of the child. Often, the solution was just to wait slow messages to arrive, or to retry speech to trigger Jibo's ASR.

Post Test

After the child completed the reading interaction, the researcher congratulated the child on finishing and offered a high five, before continuing into the posttest. The posttest was conducted identically to the pretest, using the same words and giving the child the same tasks, which were to pronounce the words and explain the meaning of the words as best as possible. After this, the child was given a break and allowed to play with coloring sheets, card games, and other puzzle toys in the lab while the researcher brought the parent to a quiet space for the second stage of the study.

6.2.2 Phase 2: Authoring Interface

Stage 2 of the study involved participation from the parent. The goal of this stage of the study was to evaluate the UI of the authoring interface and to observe which objects the parents choose to label in the story. Prior to the study, the researcher prepared an outline of a story by uploading images and typing text of the first 10 pages of the story. The story was the same for all children, and was chosen to be of medium difficulty. The story is called Freda Says Please. The researcher demonstrated how to record audio, by recording the title page. The researcher also demonstrated how to playback the audio and re-record it if it is not satisfactory. Then the researcher demonstrated how to label an object, and explained that labels could be words that either did or did not appear in the text.



Figure 6-4: A parent using the authoring interface during the user study.

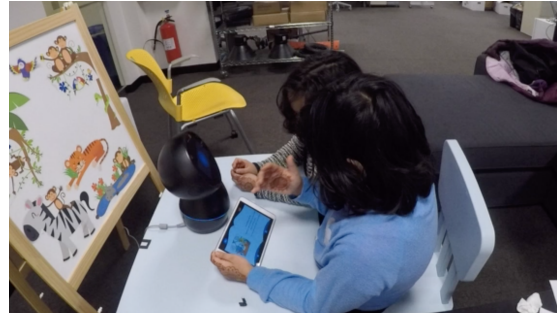
After the demonstration, the researcher left the parent alone to record the rest of the pages and label any objects they wished to. When the parent was done, the researcher guided them through pressing the submit button to upload the story, asked them about any technical problems they encountered and then invited them to watch the child read the story in the next stage of the study.

6.2.3 Phase 3: Explore Mode

The child was asked to come back to sit at the table with Jibo, and was told that she was about to read a story that her parent had just helped create. The researcher demonstrated how to read a story in explore mode, using a story different than the one the parent had made. The researcher showed the child how to listen to the audio of a page, swipe to listen to just one sentence, tap on words to hear Jibo pronounce them, and tap on objects to hear Jibo say them. The researcher then returned to the library view of the tablet app and pressed "Find More" to download the newly uploaded story to the tablet. Then the researcher handed the tablet to the child and encouraged the child to read through the story at her own pace. The parent was allowed to watch from behind the child. Several of the families who came into lab for the study had multiple children, and sometimes siblings wanted to read the story in explore mode together. This is depicted in Figure 6-5.



(a)



(b)

Figure 6-5: The two sisters on the right both wanted to read and interact with the tablet during explore mode, and were so excited that they sometimes fought over whose turn it was. The children on the left span an age range of 5 years, but all of them were engaged in the reading activity.

6.2.4 Data Collected

The data collected from each study consisted of videos of the pretest and posttest of the child, and videos of the reading interaction itself. For subjects who participated in the study in the lab, there were also survey results, which consisted of a combination of questions asking participants to agree or disagree with certain statements, and questions that were more open ended . Additionally, the content authored by the parent, and some videos of the parent using the authoring interface, are available. Videos for all parts of the study were recorded using a GoPro, or in cases where the GoPro failed, an iPhone8 camera (this occurred only once, due to a full SD card).

Chapter 7

Results and Discussion

Reflecting back on the initial research questions posed in Chapter 1, this chapter will discuss the findings of the user study. The participants' learning outcomes, child-robot reading interaction, and storybook framework were evaluated across three dimensions:

1. Learning objectives - During and after the interaction, did the child's knowledge of the pronunciation and meaning of target words increase? For which groups of children were the improvements largest?
2. System and interaction design - Which parts of the interaction triggered positive (pride) or negative (frustration, disinterest, confusion) responses in the child? What are the components of the system that require rethinking or revising?
3. Engagement - What qualitative feedback did parents have about the usability and usefulness of the authoring tool, and what qualitative feedback did children have about the reading experience?

7.1 Learning Objectives

7.1.1 Pretest to Posttest Improvement

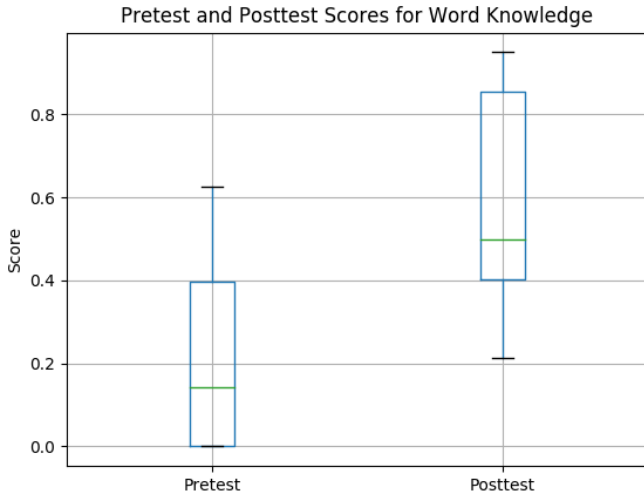
The first analysis performed compared the pretest and posttest results of each child, independently for pronunciation of each target word and knowledge of the meaning of each target word. Then the pre-to-post improvement between children in various groupings was compared, such as by age, gender, and initial reading level.

Of the 17 child participants, 12 engaged in the reading interaction in evaluate mode and are used in this analysis. The other four children had a reading ability below the baseline required to participate. One of the 12 children had incomplete posttest data, and therefore excluded from the analysis. That left 11 children whose data are used. One of these 11 children experienced technical problems or lack of interest that caused the study to terminate prematurely, and therefore for that child, only the target words that were reached by the time of termination are used in the analysis.

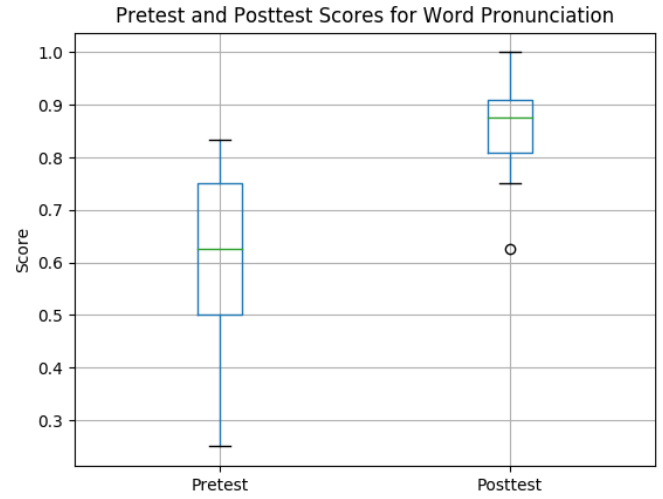
A within-subject analysis of the average raw improvement in pronunciation and knowledge across the entire population are shown in Figure 7-1. The pretest and posttest scores for each participant in either pronunciation or knowledge of target words is calculated as the number of points the participant received divided by the maximum number of points. For each word, the participant could receive a score of 0, 1, or 2 for both pronunciation and knowledge. 0 indicates the child provided no response or was entirely incorrect; 1 indicates that the child made an error on one phoneme for pronunciation (e.g. bat vs. boat) or could categorize the word (e.g. "an animal") but didn't provide a good explanation; 2 indicates that the child pronounced the word correctly or gave an explanation of the word that was satisfactory. For word pronunciation, a Wilcoxon signed-rank test shows that there was a significant increase in pretest to posttest scores (pre: $\mu = 0.58279, \sigma = 0.19633$; post: $\mu = 0.85617, \sigma = 0.10856$; $p = 0.00333$). For word knowledge, the same test also shows a significant increase in pretest to posttest scores (pre: $\mu = 0.22662, \sigma = 0.24962$; post: $\mu = 0.59178, \sigma = 0.27107$; $p = 0.00331$). Both of these results are significant, demonstrating that a single encounter with the interactive reading experience with a robot does indeed improve children's vocabulary for target words in the story that is read. The improvement in word pronunciation is more evident than the increase in word knowledge.

A between-subject analysis of the the raw difference between pretest and posttest scores for knowledge and pronunciation, grouped by four different categories, are shown in Figure 7-2. These four categories are *age group* (5-6 year olds were given the label "young" and 7-8 year olds were given the label "old"), *gender* (male or female), *reading level* ("children who read the Clifford story or the Toad story were given the label "beginner" and children who read the Henry story were given the label "advanced"), and *location* (either in lab or in school). For this analysis, a Mann-Whitney U test was used.

Between different age groups, the raw improvement of pronunciation scores was higher for



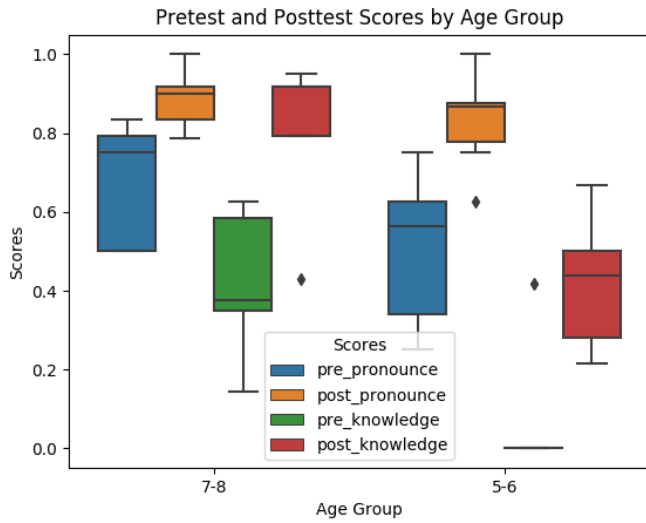
(a) Knowledge Raw Pretest and Posttest



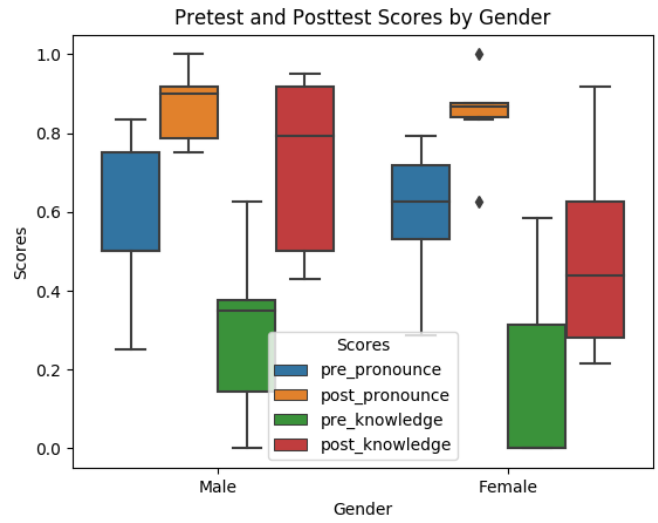
(b) Pronunciation Raw Pretest and Posttest

Figure 7-1: At a glance: the pretest and posttest scores for knowledge (a) and pronunciation (b) of the target words, across all of the children.

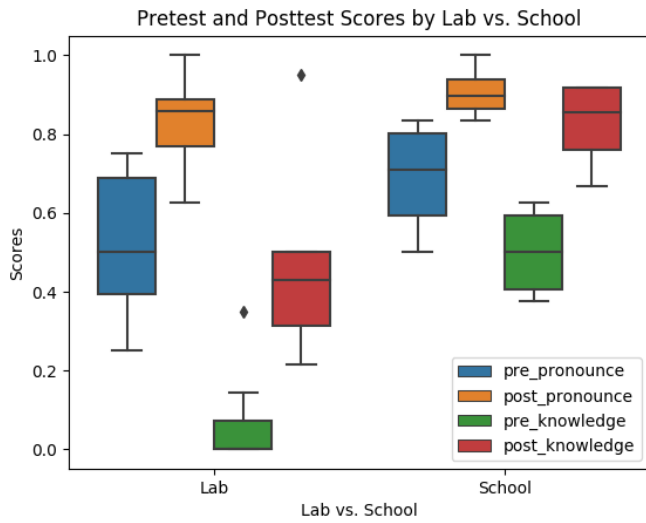
young participants ("young": $\mu = 0.32440, \sigma = 0.18946$; "old": $\mu = 0.21214, \sigma = 0.14338$; $p = 0.26093$), but the raw improvement of knowledge scores was slightly higher for old participants ("young": $\mu = 0.34821, \sigma = 0.12963$; "old": $\mu = 0.38548, \sigma = 0.13083$; $p = 0.26093$). Between participants of different genders, the raw improvement in pronunciation scores was higher in males than females ("female": $\mu = 0.24802, \sigma = 0.19676$; "male": $\mu = 0.30381, \sigma = 0.15331$; $p = 0.20512$). The raw improvement in knowledge scores was also higher in males compared to females ("female": $\mu = 0.32044, \sigma = 0.10637$; "male": $\mu = 0.41881, \sigma = 0.13538$; $p = 0.10011$). Participants who partook in the study in lab had higher raw improvement in pronunciation scores compared to students who partook in the study in school classrooms ("lab": $\mu = 0.30459, \sigma = 0.18399$; "school": $\mu = 0.21875, \sigma = 0.15729$; $p = 0.31792$). The same trend is observed in knowledge scores ("lab": $\mu = 0.38929, \sigma = 0.14712$; "school": $\mu = 0.32292, \sigma = 0.07116$; $p = 0.31792$). Participants with a lower reading level improved more in pronunciation compared to participants with a higher reading level ("advanced": $\mu = 0.20500, \sigma = 0.13964$; "beginner": $\mu = 0.33036, \sigma = 0.18720$; $p = 0.20512$). However, participants with a higher reading level experienced a slightly higher improvement in knowledge scores compared to those with a lower reading level ("advanced": $\mu = 0.37833, \sigma = 0.13840$; "beginner": $\mu = 0.35417, \sigma = 0.12496$; $p = 0.39186$).



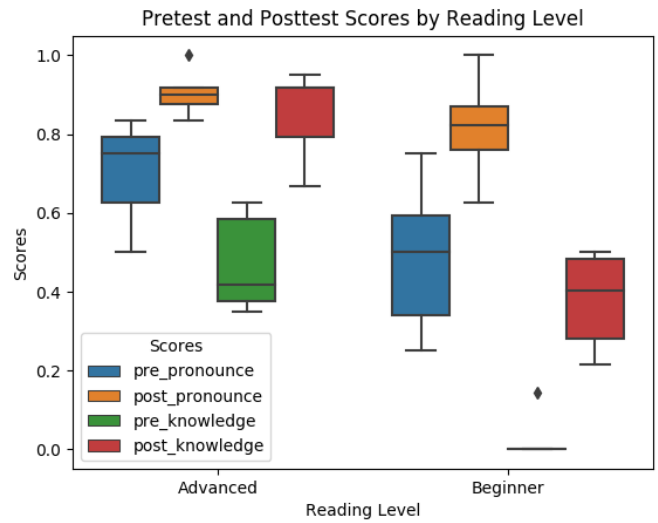
(a) Age Group



(b) Gender



(c) Lab vs. School



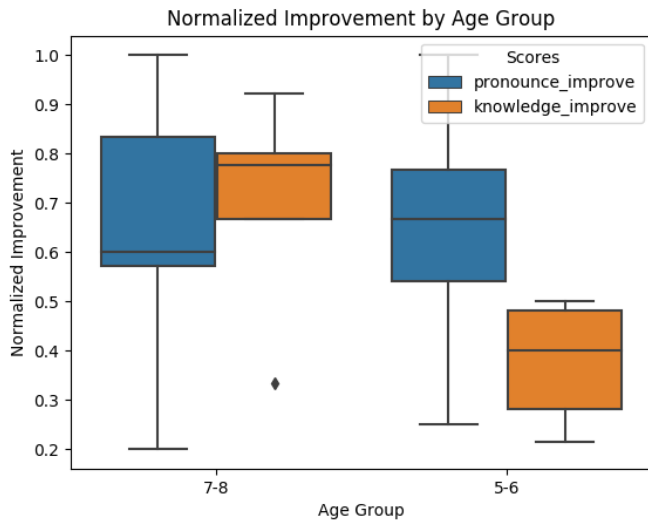
(d) Reading Level

Figure 7-2: Pretest and posttest scores for word pronunciation and knowledge, across different groupings of participants.

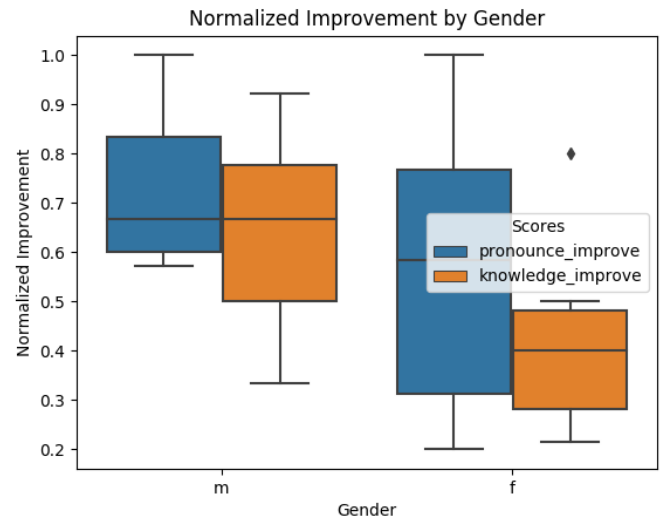
The raw pre-to-post improvement in scores does not tell the whole story, since children who began with a high pretest score have less opportunity to improve their score than those who began with a low pretest score. To further explore the effect of the four different factors on participant's learning, another metric was used. This metric is the normalized pre-to-post improvement, and it was calculated as the raw pre-to-post improvement divided by the maximum improvement the child could achieve based on the pretest result. As mentioned previously, for each word in either a pretest or posttest, the child could receive a score of 0, 1 or 2 for pronunciation and knowledge, so if there were w words, then the max score for either pronunciation or knowledge would be $m = 2w$. If a child's pretest score was p and her posttest score was q , then the normalized measure of improvement is $\frac{q-p}{m-p}$, where m is the max score. A Mann-Whitney U test was used for the following unpaired analyses.

For target word pronunciation, none of the four factors had a significant impact on the normalized improvement in pre-to-post scores. These results are shown in Figure 7-3. Although results were not significant, there were observable trends. A Mann-Whitney U test shows that older participants had very similar normalized improvement in pronunciation scores compared to younger participants ("young": $\mu = 0.64722, \sigma = 0.25613$; "old": $\mu = 0.64095, \sigma = 0.30285$; $p = 0.5$). This is in contrast to a more pronounced difference in the raw improvements. This suggests means that although younger children learned to pronounce a higher number of words, both old and young students learned to pronounce the same percentage of target words that they did not know on the pretest. Male participants had a higher normalized improvement in pronunciation scores than female participants ("female": $\mu = 0.56944, \sigma = 0.31348$; "male": $\mu = 0.73429, \sigma = 0.17996$; $p = 0.20459$), similar to the trend seen previously with raw improvement. Participants in schools had higher normalized improvement than participants in lab ("lab": $\mu = 0.62687, \sigma = 0.23548$; "school": $\mu = 0.67500, \sigma = 0.34467$; $p = 0.31753$), which is opposite the trend seen with raw improvement. And participants with a higher reading level had slightly higher normalized improvement than participants with a lower reading level ("advanced": $\mu = 0.66000, \sigma = 0.30037$; "beginner": $\mu = 0.63135, \sigma = 0.25763$; $p = 0.39161$), following the trend seen with raw improvement.

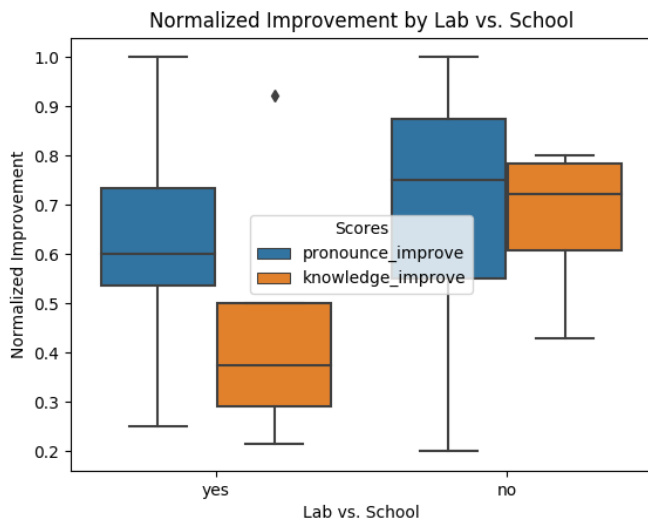
For word knowledge, two factors did not have a significant impact, while the other two factors did. These results are shown in Figure 7-3. Factors that did not create significant difference were gender and location (lab vs. school). For gender, a Mann-Whitney U test



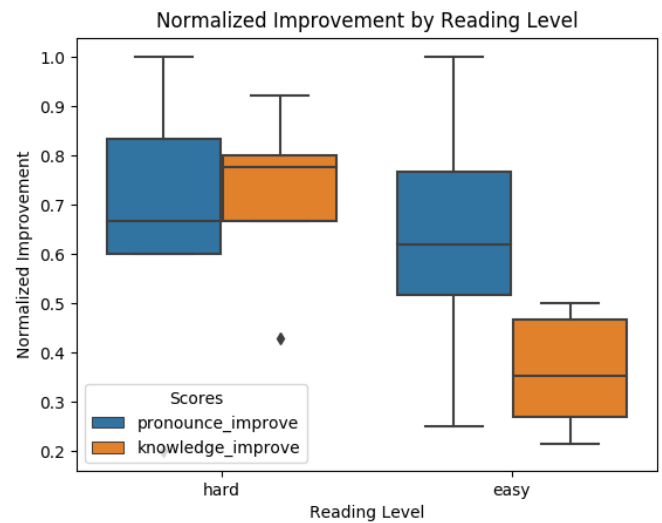
(a) Normalized Improvement by Age Group



(b) Normalized Improvement by Gender



(c) Normalized Improvement by Lab vs. School



(d) Normalized Improvement by Reading Level

Figure 7-3: The normalized improvement in pre-to-post scores for target word pronunciation and knowledge, grouped by four different factors. Significant differences in knowledge improvement between children in different age groups and reading levels.

shows that male participants had a slightly higher normalized improvement ("female": $\mu = 0.42798, \sigma = 0.21153$; "male": $\mu = 0.64017, \sigma = 0.23101$; $p = 0.10011$), and for location, participants in schools had a much higher normalized improvement than participants in lab ("lab": $\mu = 0.44224, \sigma = 0.23916$; "school": $\mu = 0.66825, \sigma = 0.17010$; $p = 0.07771$).

There were significant differences with respect to the effect of age and reading level on normalized improvement of knowledge of target word meanings. For age, a Mann-Whitney U test shows that children who are older achieved greater improvement in knowledge of target words ("old": $\mu = 0.70017, \sigma = 0.22437$; "young": $\mu = 0.37798, \sigma = 0.12290$; $p = 0.02734$). For reading level, the same test shows that the advanced readers achieved greater improvement in knowledge of target words than those who in the beginner level, ("advanced": $\mu = 0.71922, \sigma = 0.18625$; "beginner": $\mu = 0.36210, \sigma = 0.12120$; $p = 0.01109$). These results are also shown in Figure 7-3.

Older children and more advanced readers showed a higher improvement in their knowledge of target words, perhaps since they already are able to grasp the pronunciation. As shown in Figure 7-2, these children had much higher pretest scores for pronunciation of words. Younger children are likely still struggling with reading the word, and so do not pay as much attention to their meanings. As shown in Figure 7-3, the pronunciation gains from both groups are somewhat similar. It seems that both younger and older children are able to improve their pronunciation of words, but younger children have a more difficult time grasping concepts and meanings of words. To dive further into this, an analysis was done to compare improvement on words that children didn't know how to pronounce or explain with improvement on words that children knew how to pronounce but not how to explain. Of the 88 pre and post test samples (per word per child), 46 of them were words the child did not know how to pronounce or explain ("neither"), 18 were words that the child knew how to pronounce but not explain ("pronounce_only"), and the remaining 24 were words that the child knew how to pronounce and explain. The normalized improvement for knowledge learning was indeed higher for words that the child already knew how to pronounce. This result was not significant under a Mann-Whitney U test, but the trend is clear ("pronounce_only": $\mu = 0.63889, \sigma = 0.44740$; "neither": $\mu = 0.51087, \sigma = 0.48864$; $p = 0.17818$). This result is shown in Figure 7-4.

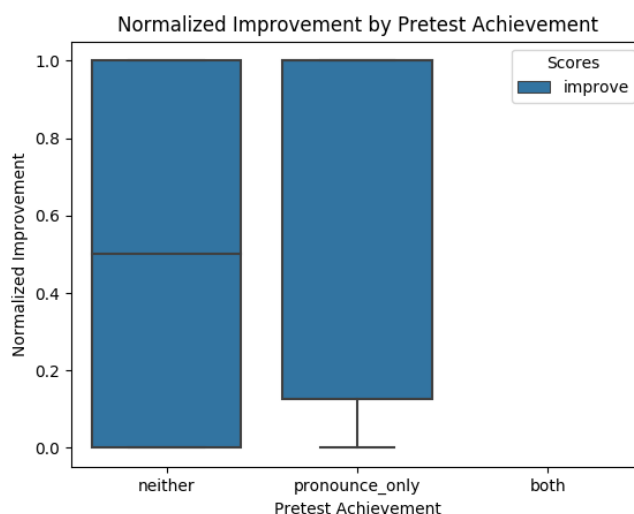


Figure 7-4: The normalized improvement for words that the child already knew how to pronounce (but did not know the meaning of) is higher than for words that the child did not know how to pronounce (or know the meaning of). The normalized improvement for words the child already knew how to pronounce is always 0/0 and is defined to be 1, but it is not plotted since it is not relevant to the analysis.

Clifford and the Jet	The Hungry Toad	Henry’s Happy Birthday
jet (2)	soap (4)	donned (2)
fog (2)	throat (3)	spatula (2)
cab (2)	toaster (3)	unruly (2)
	coat (3)	zigzag (2)
		bunting (2)

Table 7.1: The target words for each of the three evaluation stories for the user study. The number in parenthesis indicates the number of occurrence of a word in a story.

7.1.2 Midstory Effect of Jibo Interventions

Video Annotations

To evaluate whether the interventions from Jibo were significant in achieving those objectives, the video data collected in the studies was annotated. The purpose of the annotations was to note what questions Jibo asked the child, note when the child asked Jibo for help on difficult sentences, and to mark the child’s pronunciation score for target words that appeared in the story more than once. For each of the stories, the target words that appeared more than once are given in Table 7.1.

An annotator coded the videos of the reading interaction for the following features:

1. Pronunciation scores when children say one of the duplicated target words during the story
2. Jibo interventions (including the type of intervention and the relevant target word if there is one)
3. Child asking for help on a sentence, and any target words that appear in that sentence

These annotations enabled the analyses of the effect of Jibo's actions on the child's learning outcome, which are presented in the next two sections.

Effect of Asking for Help

The first analysis done using video annotation data was to see if there was a correlation between the number of times a child asked for help and how much their pre-to-post scores improved.

To perform this analysis, the video for each interaction was marked with which target words appeared in a sentence the child asked for help on. Then all pretest and posttest word utterances were divided into two groups, either words that were asked about for help by the child, or words that were not asked about. This resulted in a total of 88 data points, of which 13 represented words the child had asked about, and 75 represented words the child did not ask about. An analysis was performed to compare the normalized improvement in pronunciation and knowledge between these two sets of words. The results were significant and unexpected, and are shown in Figure 7-5. A Mann-Whitney U test shows that the normalized improvement in pronunciation scores from pretest to posttest is lower in words that appeared in sentences that the child asked Jibo to read ("help: $\mu = 0.50000, \sigma = 0.47140$; "no_help": $\mu = 0.68055, \sigma = 0.44965$; $p = 0.01267$). There is a similar trend for knowledge improvement. The same test also shows that knowledge improvement is significantly higher for words that did not appear in sentences that the child asked for help on ("help: $\mu = 0.29167, \sigma = 0.45017$; "no_help": $\mu = 0.60577, \sigma = 0.46796$; $p = 0.00479$). It has already been shown that children who are older and are familiar with more words exhibit a larger improvement in knowledge. It's possible that the reason that there was a larger improvement in knowledge for words that the child did not ask about is that those children already had a higher base ability to pronounce the words and didn't need help. Indeed, the pretest scores of words that were not asked about were higher than the scores of words

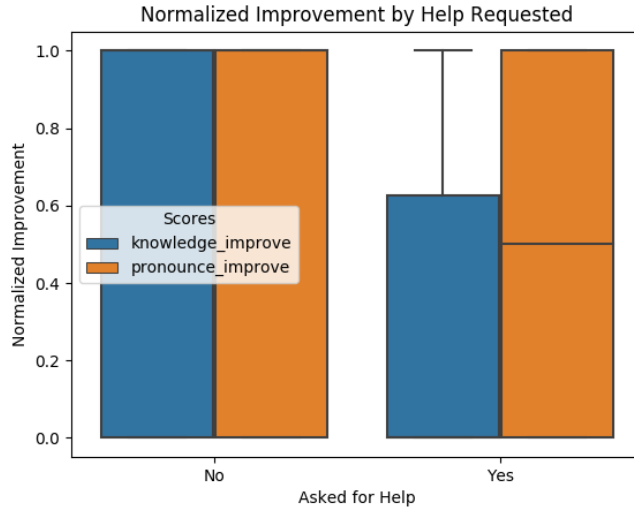


Figure 7-5: The pronunciation and knowledge normalized improvement for words that appeared in sentences the child asked for help reading vs words that the child did not ask about. Interestingly, when the child did not ask for help on a sentence containing a target word, the improvement is higher for both pronunciation and knowledge of that target word.

that were asked about. Pronunciation pretest scores: ("help": $\mu = 0.34615, \sigma = 0.42743$; "no_help": $\mu = 0.65333, \sigma = 0.40247$; $p = 0.00887$). Knowledge pretest scores: ("help": $\mu = 0.07692, \sigma = 0.27735$; "no_help": $\mu = 0.36667, \sigma = 0.45272$; $p = 0.01158$). This shows that children with lower pronunciation score pretests were more likely to ask for help when reading sentences, which is unsurprising.

Some children did not ask for help at all. It is therefore possible to group children into those who did and did not ask for any help on sentences during the interaction. A Mann-Whitney U test on the normalized improvement of these two groups shows a trend that those who did not ask for help had a higher improvement in pronunciation, but these results were not significant ("help": $\mu = 0.59095, \sigma = 0.20724$; "no help": $\mu = 0.68889, \sigma = 0.31529$; $p = 0.32329$). The normalized improvement in knowledge showed a stronger trend, but still without significant results ("help": $\mu = 0.39523, \sigma = 0.12210$; "no help": $\mu = 0.63209, \sigma = 0.26358$; $p = 0.08496$), meaning that whether or not the child asked for help at all in the interaction is not significant to the child's learning outcome. The trends for children who did and did not ask for help are shown in Figure 7-6. It is interesting to note that the results per-word in Figure 7-5 were significant, the but results per-child are not.

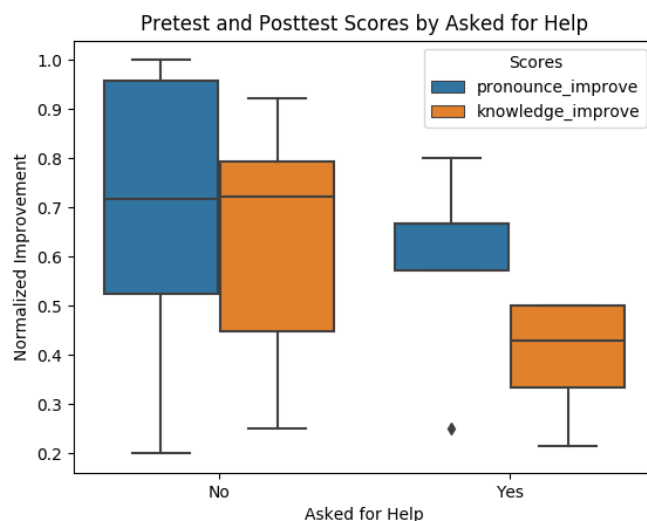


Figure 7-6: Children who did not ask for help to read sentences tended to exhibit greater improvement in pronunciation and knowledge. This is likely due to the fact that children who are more proficient readers are less likely to ask for help, and can learn from other Jibo interventions and exposure to the words.

First vs. Second Utterance of Target Words

Another analysis done with video annotations was to evaluate the effect of Jibo interventions on word pronunciations in the context of story reading. Previous analyses have used pretest and posttest scores as the metric of improvement. This analysis compares the pronunciation score of the first time a child utters a word in a sentence in the story to the score of the second time a child utters that word. The pronunciation score scale is the same 0, 1, 2 scale. The analysis is only done for words that appear at least once in the story, and the improvement is always calculated between the first and second utterances, even if there are more than two utterances. There are 9 possible combinations of first-second utterance pairs. These are categorized into three groups to categorize three different levels of improvement. This is shown in Figure 7-7.

Jibo's interventions take the form of questions that Jibo asks the child at the end of each page. As mentioned in the study protocol in the previous chapter, Jibo asks a predetermined question about the meaning of each target word at the end of the page when that target word first appears in the story. In addition, Jibo sometimes asks another question related to target words on the page. Therefore, the combination of questions associated with a target word can fall into one of these four categories:

		Second Utterance		
		0	1	2
First Utterance	0			
	1			
	2			

Figure 7-7: The nine possible combinations of first and second utterances are divided into three groups. The blue shaded group represents improvement. The red shaded group represents no improvement or regression. The green shaded group represents perfect scores both times. These three groups represent the three categories of improvement used in the analysis.

- Ask for meaning and ask for tap on word
- Ask for meaning and ask for tap on object in the picture
- Ask for meaning and ask for pronounce word
- Only ask for meaning

A fourth type of question exists, where Jibo asks the child to re-pronounce an entire sentence on the page, but this was a feature implemented later during development and it did not occur in the reading interactions during the study pertaining to the target words that appeared more than once in the story. Therefore, this type of question is omitted from analysis.

Among the 11 children, there were 39 annotations of first vs. second utterances of target words that appeared in the story. Of the 39 pairs of utterances, 25 of them exhibited perfect behavior. So, there were only 14 opportunities for the intervention to improve the in-story pronunciation of the word. The breakdown of improvement types is provided in Figure 7-8.

2x2 contingency tables were used to analyze the effect of Jibo interventions on the improvement between first and second utterance of target words, using Fisher's exact test. Each table compares a pair of interventions, each of which involves a default question about

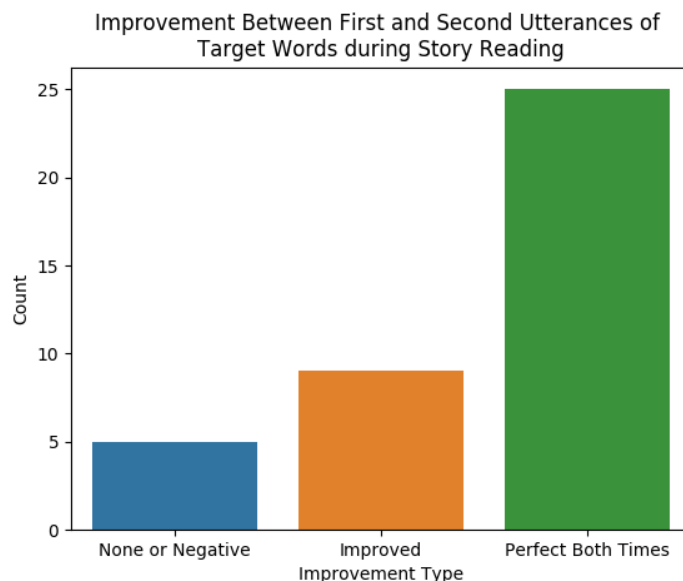


Figure 7-8: More than half of the first-second utterance pairs were both correct pronunciations. Of the remaining pairs, there were about twice as many improvements as no-improvements.

the meaning of the word and also an additional question. For example, one table compares the case where Jibo asked about the meaning of the word and asked the child to pronounce a word to the case where Jibo asked about the meaning of the word and asked the child to tap on the word in the sentence text. A comparison was also done between the aggregate results of when Jibo delivered only the default question vs. when Jibo delivered the default question plus *any* additional questions. This entire analysis was performed twice, once where the two outcome categories were "no improvement" and "improvement" and the other where the two outcome categories were "no improvement" and "improvement + perfect both times." This was done to see if excluding the times where both the first and second utterance were correct had an effect on the results. While there is a clear trend that asking the child to pronounce words is more effective in improving their pronunciation during the next encounter of the word, there is not enough data for the results to be statistically significant. The contingency tables and reported p values are given in Figure 7-9.

7.2 System and Interaction Design

The second major analysis evaluates the system's functionality and usability. For this analysis, the reading interaction videos were annotated for a different set of features than for the

	No Improve	Improve		No Improve	Improve
Meaning and Pronounce Word	1	7		1	7
Meaning and Tap Word	2	1		1	0
	p = .1515			p = .2222	
	No Improve	Improve		No Improve	Improve
Meaning and Tap Word	2	1		1	1
Meaning and Tap Object	1	0		4	8
	p > .5			p > .5	

(a) No Improve vs. Improve

	No Improve	Improve*		No Improve	Improve*
Meaning and Pronounce Word	1	19		1	19
Meaning and Tap Word	2	4		1	3
	p = .1231			p = .3116	
	No Improve	Improve*		No Improve	Improve*
Meaning and Tap Word	2	4		1	8
Meaning and Tap Object	1	3		4	26
	p > .5			p > .5	

(b) No Improve vs. Improve*

Figure 7-9: The contingency tables for the analysis of the effect of Jibo interventions on improvement between first and second utterances of target words. The outcome category Improve only includes instances where the utterance score increased, while Improve* also includes instances where the utterance score was perfect both times the child pronounced the word. There is a clear trend suggesting that Jibo asking the child to pronounce words was the most effective, but not enough data exists for the results to be significant.

above learning objectives analyses. The annotations noted positive, frustrated, confused or undesirable behavior from the child, and noted technical failures (network down). Included with each annotation is the state of the finite state machine the controller was in at the time of the annotated event. The annotations can be grouped into three broad categories: interaction design error, child's mistakes, and technical failures. A breakdown of all annotations used for this analysis can be found in Appendix B. One important note is that network failures are considered to be outside of the scope of the analysis. In other words, the goal is to analyze the design of the interaction, not the execution, and so errors that occurred due to parameters outside of the control of the study are noted but not discussed.

The annotations were analyzed for the most common failures and confusions. The number of occurrences of each annotated event, as well as a breakdown of the states during which the event occurred, are given in the figure below.

The most common errors were children tapping multiple times on something they only needed to tap once on, children needing to be prompted to press the button to see the next sentence, children repeating themselves when Jibo was slow to respond, and Jibo ASR misunderstanding what the child said. Most of these errors occurred during the same FSM state "WAITING_FOR_CHILD_AUDIO." This is when the system is waiting for the child to read text. Possible issues include latency in new sentences appearing, and children forgetting that they need to press a button, since that's the only action they are required to take without prompting from Jibo during the interaction.

It is also interesting to consider how different types of failures prompted ideas for changes to the design of the interaction. For example, there were a number of times where the child used a phrase that Jibo was not programmed to recognize, even though the phrase was perfectly reasonable. For example, the first child to participate in the study said "I do not know" instead of "I don't know" and the system did not respond correctly, and this was fixed immediately by updating the natural language processing rules that Jibo uses for this interaction. Another example of an update to the interaction design as a result of observing failure modes during user testing involves repeating Jibo's questions when the child did not hear them. Some children were excited about the activity and tried to talk to the researcher or Jibo, but then missed the question that Jibo asked because they weren't paying attention. To address this, after all the studies had ended I added logic to the controller finite state machine that commands Jibo to repeat the question if the child asks

	waiting_for_next_page	waiting_for_child_audio	waiting_for_jibo_help_with_sentence	waiting_for_end_page_jibo_question	waiting_for_end_page_child_response	waiting_for_end_page_jibo_response	waiting_for_end_story_child_response	TOTAL
jibo_interrupt_child					2			2
repeat_frustrated		2			17		1	20
proud						6		6
confused_which_sentence	2							2
confused_why_wrong						3		3
response_confusing						3		3
unrecognized_ask_jibo		2						2
child_did_not_hear_question					5			5
child_tried_wrong_action					8			8
multi_tap		19			10			29
need_prompt_button_press		15						15
need_prompt_ask_jibo		10						10
need_prompt_idk					1			1
child_interrupt_jibo			6	2		2		10
speak_before_blue					5			5
tap_accident		5						5
system_freeze	1	6		1		1		9
asr_misunderstand		4			15			19
asr_missed_help		1						1
TOTAL	3	64	6	3	53	15	1	

Figure 7-10: All system and interaction design annotations from the videos of children in the study. rows represent different types of annotations, and the columns represent the state during which those annotations occurred. The states are states of the controller's finite state machine.

with a particular phrase, such as "Can you repeat the question?" or "Can you say it again?". Issues involving the child not knowing to press buttons or to ask for help when stuck led to the refinement the demonstration process throughout the studies, and informed which parts of the system to let the child try during the demonstration to make sure they could do it during their own reading interaction. In particular, from observing run throughs of the study, I have determined that it is most important to have children try asking Jibo for help during the demonstration, press the button to see the next sentence, and practice answering an open ended response question, and future iterations of the study will definitely feature those actions in the demonstration stage of the study.

There were some observed problems that have not yet led to changes in the interaction design, and that are good places to start for implementing improvements. For example, sentences were not conspicuous enough when they appeared, especially when children were looking at Jibo instead of the tablet, leading to children sometimes not knowing which sentence to read when they looked back at the tablet. One possible fix would be to have new text visually animate more obviously and continue to animate to indicate that it is the active sentence. Another two issues that can be fixed are the child interrupting Jibo because the child thinks Jibo is done talking, and Jibo's responses to questions being confusing. The solution for both of these problems is simply to modify the TTS text sent to Jibo to remove unnecessarily long pauses and edit responses to be more straightforward.

To further categorize the types of issues the system encountered, an analysis was done to explore if there was a correlation between the age of participants and the frequency of various child errors. For errors when the child needed prompts to proceed through the interaction, there was a significant difference between children in different age groups. A Mann-Whitney U test shows that younger children had a harder time understanding the system, and needed more prompts than older children to proceed through the interaction by properly tapping buttons ("old": $\mu = 0.80000, \sigma = 0.83666$; "young": $\mu = 3.50000, \sigma = 2.50998$; $p = 0.01582$). These results are shown in Figure 7-11. It was also interesting to consider the effect age had on Jibo's ability to correctly parse the child's speech. It turns out that this was not a significant factor, although a Mann-Whitney U test shows that Jibo's ASR system did make more mistakes for younger children than for older children ("old": $\mu = 2.50000, \sigma = 0.44721$; "young": $\mu = 1.20000, \sigma = 1.76068$; $p = 0.12241$). The distribution of ASR misunderstandings by age is shown in Figure 7-12.

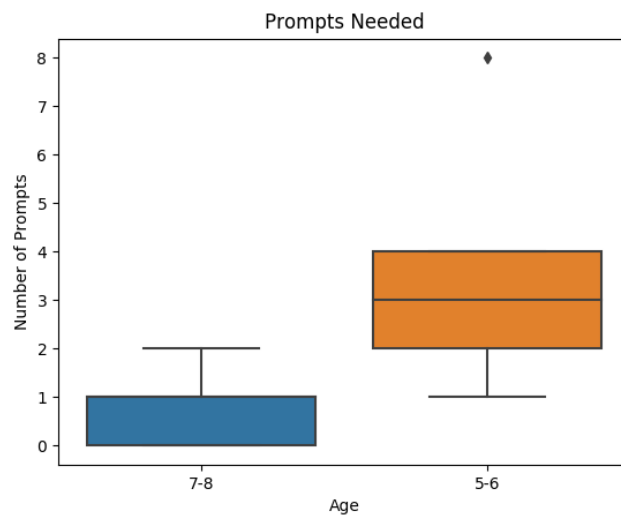


Figure 7-11: Younger children exhibit more mistakes while using the system, and need more prompting from the researcher. These prompts are for when the child forgets to press the button to see the next sentence, or forgets to ask for help when stuck.

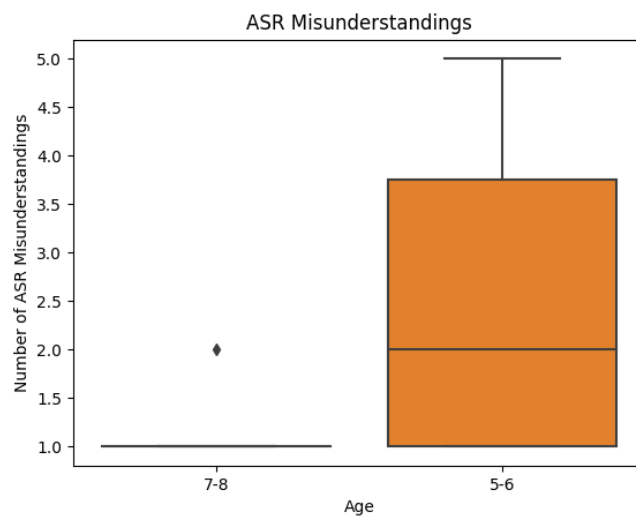


Figure 7-12: Jibo's ASR system has a harder time understanding younger children than older children.

7.3 Engagement

7.3.1 Authoring Interface Qualitative Discussion

This section qualitatively discusses the effectiveness of the authoring interface in engaging parents and enabling them to create content for their children to read.

Overall, the authoring interface was successful at showing parents how simple story creation can be. Every parent who created a story using the interface was able to see their child read through that story with Jibo. In other words, there were no cases throughout the course of the study where a story failed to be transferred from the authoring interface to the tablet.

Interviews

In informal interviews, parents shared that they liked how they could craft a story based on the interests of their child by using any images and text they wanted. However, a big pain point for the parents was the scene object drawing mechanism for tracing bounding boxes. Many parents complained that boxes they wanted to delete would still be visible, and that the location of the labels was confusing.

Some parents thought of interesting applications for the authoring interface and reading interaction. Two moms expressed that they were wary of technology but that they were interested in exploring educational apps to engage their children in activities more useful than games. One mom felt that her children were more comfortable answering questions from a robot rather than from a teacher, which she saw as an opportunity to encourage her children to share their thoughts. A few applications that were brought up when parents were asked what they would use the authoring interface + storybook experience for if it were free and available include:

- Create storybooks that deal with ninjas and dinosaurs to appeal to a young son
- Examination prep for reading comprehension questions
- Create interactive storybooks out of the printed PDF picture books that children bring home from school

These applications are already feasible with the current system, which is heartening.

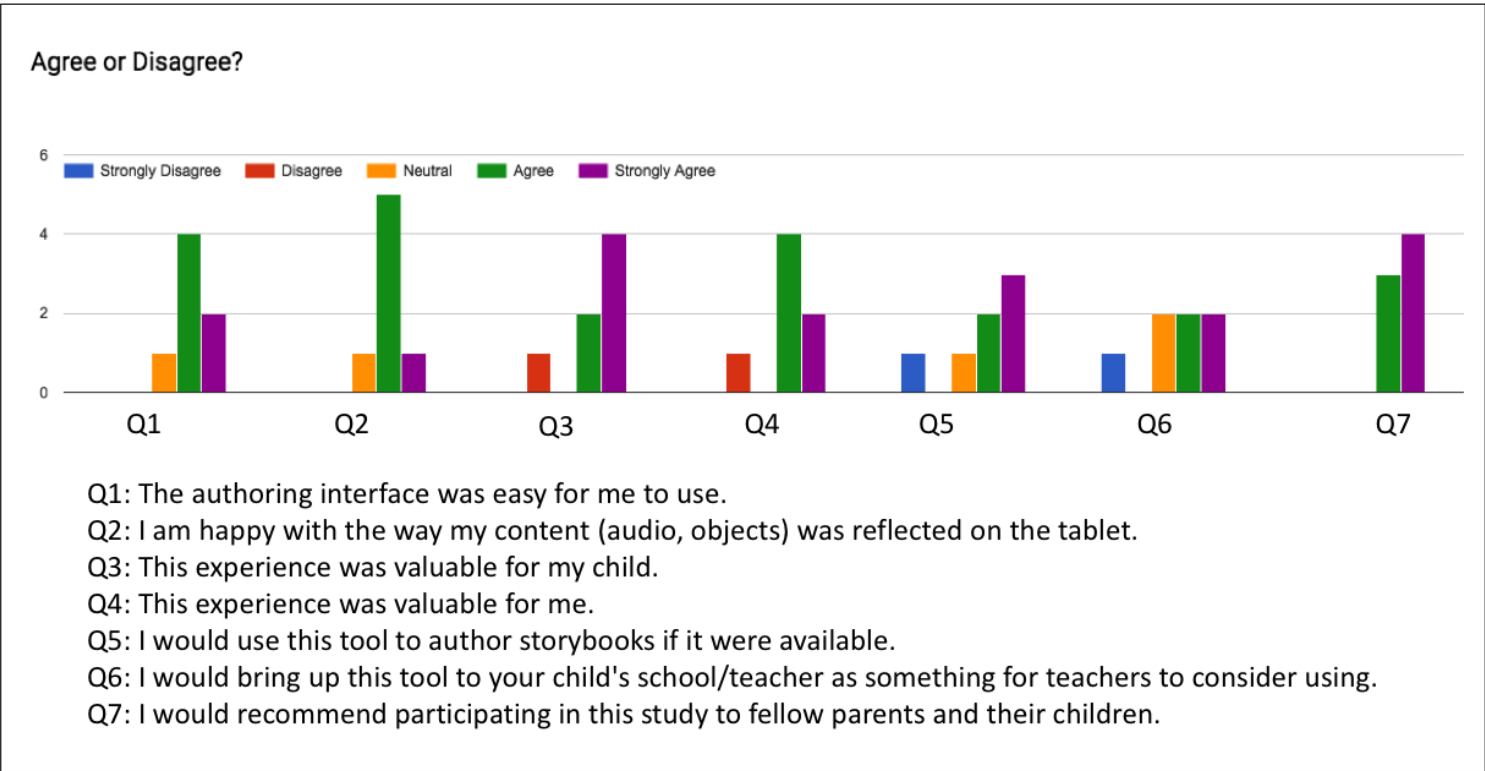


Figure 7-13: The results of the agree/disagree questions given on the parent survey.

Survey Responses

In addition to interviews, parents also provided responses to a survey after the study concluded. The survey had a mixture of structured agree/disagree questions and unstructured open ended questions. The results of the agree/disagree questions are shown in the figure below.

In the unstructured questions, parents were asked to provide reasoning for their responses to the agree/disagree questions, and asked what they liked and didn't like about the experience. I then categorized responses to the unstructured questions into common sentiments that were expressed. These sentiments are depicted in the chart below. In general, parents found the recording feature easy to use, thought the experience was useful for their children, and struggled with the scene object labeling feature. Some parents also offered their views on technology and media for children in general. The results of the open ended questions are presented in Figure 7.2. Parents gave lengthy answers, and for brevity only opinions expressed multiple times are reported in this table.

Comment	Frequency
Drawing boxes for labeling objects didn't work well	2
Easy to record audio, liked being able to do so, personalization	3
Story creation process is quick and simple	5
Reading system would be useful in schools with fewer resources, but not in my child's school because there are enough teachers	3
Desire for auto labeling repeated objects on different pages to avoid tedious relabeling	2
We as a household tend to limit technology usage, but enjoyed the experience nonetheless	
Interface felt familiar and intuitive (similar to other apps)	2
Engaging for kids	3

Table 7.2: Parents' general comments about the experience, collected from an online survey after the conclusion of the study.

Improvements Made in Response to Parent Feedback In response to the complaints from some parents, after the study concluded, the scene object labeling mechanism of the authoring interface was redesigned, so that the implementation is cleaner and the feature is more stable. The specific bugs that parents pointed out, with regards to drawn boxes not showing labels and sometimes not allowing deletion, are now fixed.

7.3.2 Child Engagement and Enjoyment Qualitative Discussion

Interviews

At the conclusion of the study, children were interviewed briefly about their impressions of Jibo and the reading experience as a whole. Some negative feedback received from children involved frustration when the system would freeze ("Jibo kept getting stuck"), or confusion when there was delay in Jibo responding ("Jibo didn't hear me the first time"). These issues are difficult to resolve since they result from network problems, limitations of the tablet compute power, and the latency of API calls to cloud services. Potential solutions include offloading the recording and SpeechACE functionality to Jibo instead of the tablet, but this is difficult since it is not currently possible to access Jibo's microphone stream. The interviews also brought out positive commentary. One child said he particularly liked that Jibo didn't give help unless he asked for it, because his parents sometimes jump in to help without giving him a chance to try. Another child remarked that she liked Jibo because Jibo was knowledgeable and asked funny questions. One girl loved listening to the story

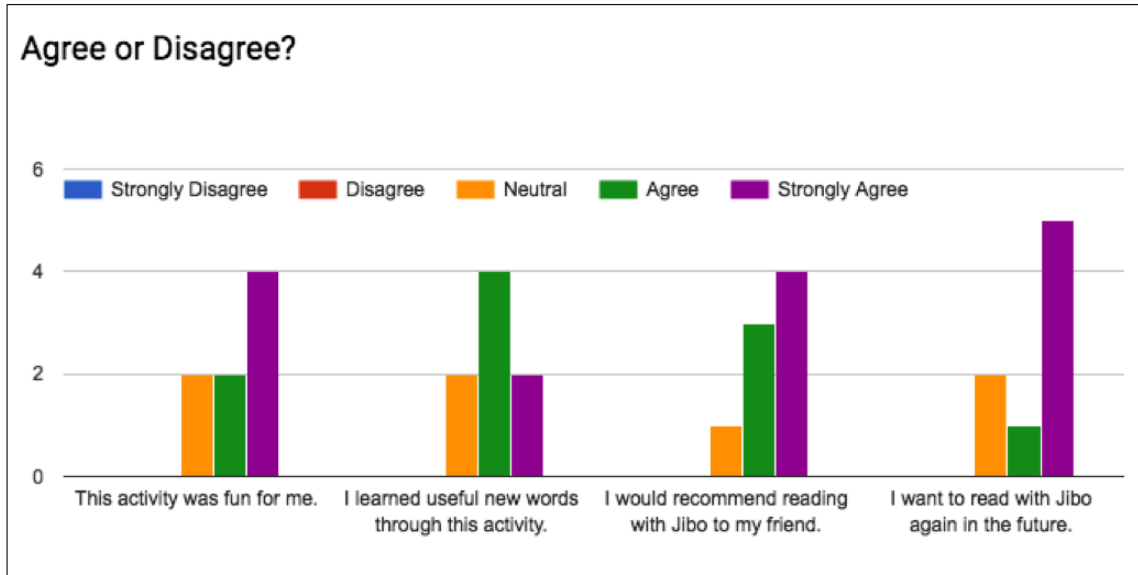


Figure 7-14: The results of the agree/disagree questions given on the child survey.

with her mom’s recorded voice, and went over to give her mom a hug after finishing the story. These interview snippets, while anecdotal, demonstrate that children engaged with the system and formulated opinions about it.

Survey Responses

Children also filled out a survey after they went home after the study. The survey was a combination of structured agree/disagree questions, and unstructured open ended questions. The responses to the agree/disagree questions are presented in Figure 7-14. Only children who did the study in lab (as opposed to in schools) filled out the survey. None of the children expressed that they would not want to read with Jibo again in the future, and about two thirds of the children indicated strong interest in reading with Jibo again.

There were two questions on the survey pertaining to the child’s preference for either explore mode or evaluate mode. The majority of children felt that evaluate mode was more educational (i.e. that they learned more) than explore mode, and half of the children liked it better when they were asked to read the story sentence by sentence rather than at their own pace. A possible reason for this besides the learning aspect is that Jibo is more active and engaging in evaluate mode. The results of the survey questions regarding story mode are reported in Figure 7-15.

The last questions of the child survey asked for positive and negative general comments

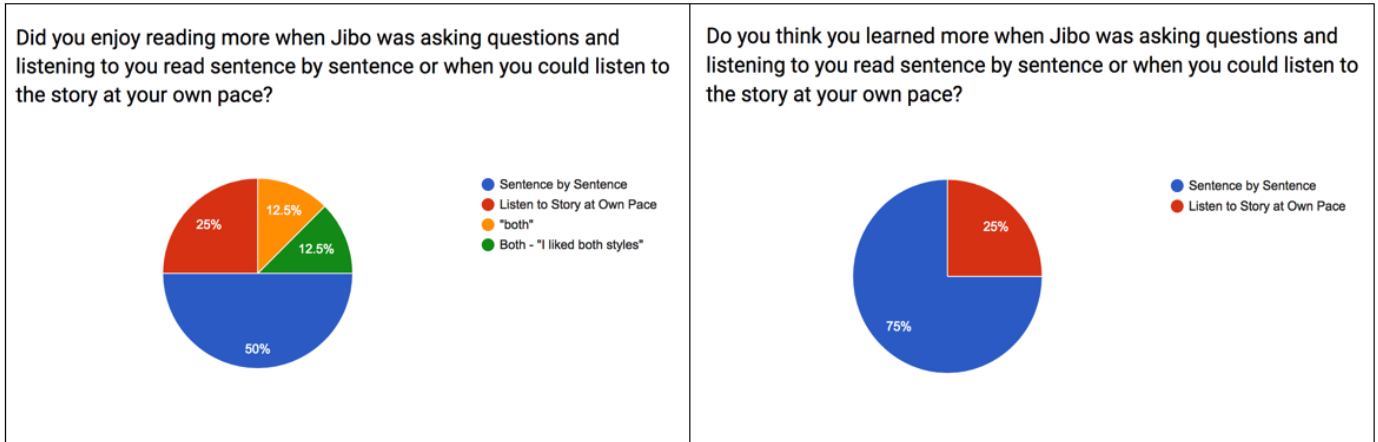


Figure 7-15: The results of questions on the child survey pertaining to evaluate and explore mode.

Comment	Frequency
Interacting with Jibo was fun	6
Liked the story, questions, and Jibo teaching	3
Jibo funny and engaging	2
Being asked to read a lot was tiring	1
Jibo had a hard time listening	1
Jibo looked strange with only one eye	1
The story was too short	1

Table 7.3: Child responses to likes and dislikes about the reading experience.

about the experience. In Figure 7.3, child responses are reported by categorizing similar responses and indicating the frequency of each response. The most common response was that the experience was "fun." Some complaints were that the reading task was tiring, and that Jibo was "bad at listening". To address the problem of a long continuous reading task being too tiring, a good feature to add would be to gracefully pause the interaction in the middle and resume from where it left off, to allow children to take a break. Currently, the app can be started from a previous saved state but only when the components are restarted entirely, which breaks the flow of the interaction. It is a much more difficult problem to try to improve Jibo's listening capabilities, since they are limited by the quality of the ASR services that it uses. However, future work in the area of developing better speech recognition models trained on children's speech could eventually improve ASR capabilities and make speaking to Jibo during the interaction a smoother experience for children.

Conclusion The analyses presented in this chapter suggest that the reading interaction does improve children learning, answering the first research question posed in Chapter 1. All children experienced an increase in score in both pronunciation and knowledge of the target words. The degree to which the children improved was dependent on a number of factors, most notably their age. Parents found the concept of being able to author storybooks for their children to be compelling. Parents found the interface's design to be intuitive, but were frustrated by technical issues, particularly with object labeling. That feature has since been fixed, and is now functioning without the disruptive behavior parents were displeased with. Children were engaged by Jibo, with almost all of them describing the interaction as "fun" and expressing interest in reading with Jibo again in the future. This answers the second research question posed in Chapter 1, by demonstrating that the authoring interface can engage parents in the content creation process and be an important step toward making interactive storybooks and reading experiences with robots widely available. The next chapter ties together the ideas presented in the thesis and offers more ideas for future directions of the project.

Chapter 8

Conclusion and Future Work

In this thesis, I presented the design of a system that engages children in interactive reading experiences with a Jibo robot and allows parents to author content for the reading experience. Children were able to learn new words through this experience, and the interaction engaged both parents and children., There were some technical limitations that prevented the interaction from being as effective as it could have been. This chapter suggests possible next directions for the project, and discusses more limitations and potential solutions to those limitations.

Some limitations beyond those discussed in the results above are that the storybooks on the tablet do not currently support as rich of an interaction via tablet touches. One feature to add is sprites and animations, which will bring the storybook to life. Displaying sprites on the tablet app is already possible, and the remaining challenge is to obtain the sprites themselves via tracing. One option is to crowdsource the tracing tasks on MTurk, and another option is to ask users of the authoring interface to trace the objects they wish to label. A disadvantage of using MTurk is inaccurate results, and a disadvantage of asking authoring interface users to trace is that it burdens them with a tedious task. A better solution would be to allow users to draw a bounding box around the object, then do the tracing automatically by employing techniques from computer vision to follow edges. Incorporating sprites into the app would also enable animations to further engage the child.

A second way to further engage children would be to add gamification features to the reading experience, such as a way for children to gain points by reading more books or answering more questions. This would hopefully not encourage competition, but rather

create easily reachable goals and incentives for children to interact more with the system. Such a modification to the reading experience would require creating the concept of accounts for different children, and devising a point scheme for different achievements.

Another limitation of this project is that each child only interacted with the system once, making it impossible to evaluate long term effects of the experience. Fortunately, this summer, there are plans to deploy interactive storybooks in schools in Georgia as part of an effort to pilot test an entire suite of learning and educational apps involving robots. Ideally, the controllers for these apps will run in the cloud, and collect data from multiple interactions and multiple children, to better improve the underlying models and create a better experience for the children.

More work can be done to promote and enhance the authoring interface and get parents and educators more involved in the project. It would be great to run workshops to teach parents how to make stories and test them out in the app, and to augment the authoring interface with the ability to have an account, edit past stories, and create variations of the same story. These features are within reach, and can be implemented in the future.

Improvements to the pronunciation evaluation and question asking protocol could improve the story reading experience. These improvements could include higher confidence identification of right or wrong answers even for open ended questions, which could be achieved through providing the ASR recognizer with particular phrases to listen for. Another improvement would be to use ASR models specifically trained on child speech data, since commodity ASR is most attuned to adult speech, and struggles with children. Our lab is actively collaborating with other groups to develop a better child speech model, which could then be used in the interactive storybook experience.

I hope the work in this thesis has provided a foundation on which the lab and the field can continue to grow. I look forward to seeing the future direction of this project, and believe that this research is another step forward in developing educational and engaging learning experiences between children and technology.

Appendix A

Predetermined Questions for Jibo to Deliver in Evaluation Mode

The tables below provide the predetermined questions Jibo asked to provide learning moments for the children in each of the three stories. For each question, the relevant target word is listed, or the question is marked as open ended.

Number	Question	Target Word
1	What exactly is a throat?	throat
2	What did Toad eat that he shouldn't have?	soap
3	Do you know what a toaster is?	toaster
4	What did the doctor do to help Toad?	(open ended)
5	Do you know what foam is?	foam
6	How do you think the doctor feels right now?	(open ended)
7	Do you know what a coat is?	coat
8	Do you know what groaned means?	groaned
9	What is a rowboat?	rowboat

Table A.1: The predetermined questions for The Hungry Toad.

Number	Question	Target Word
1	What is a jet?	jet
2	Why won't Jim's jet fly?	fog
3	What is a cab?	cab
4	What does it mean to jog?	jog

Table A.2: The predetermined questions for Clifford and the Jet.

Number	Question	Target Word
1	Can you explain the word donned to me?	donned
2	What is a spatula?	spatula
3	What kind of cake did Henry want?	(open ended)
4	What does plain mean?	plain
5	Do you know what a bunting is?	bunting
6	How do you think Henry feels right now?	(open ended)
7	What kind of pattern does Henry's favorite T-shirt have?	zigzag
8	Can you tell me what unruly means?	zigzag
9	What did Henry's mom want him to don?	donned
10	Do you think Henry liked getting a birthday kiss from Aunt Sue?	(open ended)
11	What do you think is in the big yellow box?	(open ended)
12	Oh I love games, what's your favorite party game?	(open ended)
13	What does Henry look like with the tail on his back?	donkey
14	What does miserable mean?	miserable
15	The book says the room became very unruly, what made the room unruly?	unruly
16	Why is Henry miserable about his birthday?	miserable
17	Do you know what helium is? Can you explain it to me?	helium
18	Can you tell me what a kite is?	kite
19	Wow that's a great gift. Do you know what a raft is?	raft
20	What is a siren?	siren

Table A.3: The predetermined questions for Henry's Happy Birthday.

Appendix B

Video Annotations for System Design Evaluation

system_freeze	System froze for reasons unrelated to the interaction or child, usually a network failure	technical failure
asr_misunderstand	Jibo's ASR system parsed the child's speech incorrectly	technical failure
asr_missed_help	Child asked for help but Jibo did not produce any transcription or notice	technical failure

Table B.1: The target words for each of the three evaluation stories for the user study. Each word appears once in the story, unless otherwise indicated.

Annotation	Description	Category
jibo_interrupt_child	Jibo interrupts the child while child is speaking	interaction design
repeat_frustrated	Child repeated speech in a frustrated manner because it seemed like Jibo didn't hear	interaction design
proud	Child displayed happiness and pride when Jibo said they answered a question correctly	interaction design
confused_which_sentence	Child was confused about which sentence to read because they didn't notice a new sentence appearing	interaction design
confused_why_wrong	Child didn't understand why their pronunciation of a word was wrong, but Jibo did not provide a good explanation	interaction design
response_confusing	Child heard Jibo's response to the child's answer but found the response puzzling, for example when Jibo's response begins with "My teacher told me" and the child was surprised Jibo could have a teacher	interaction design
unrecognized_ask_jibo	Child asked for help using a valid phrase that Jibo didn't recognize to the speech matching rule	interaction design

Table B.2: Video annotations in the category of interaction design.

child_did_not_hear_question	Child was not paying attention and didn't hear the question	child mistake
child_tried_wrong_action	Child misunderstood question and performed an action that didn't answer the question	child mistake
multi_tap	Child tapped a button, word or object many times quickly, instead of just once	child mistake
need_prompt_button_press	Child needed to be told to press button to see next sentence	child mistake
need_prompt_ask_jibo	Child needed to be told to ask Jibo for help when stuck	child mistake
child_interrupt_jibo	Child spoke when Jibo was still speaking	child mistake
speak_before_blue	Child began speaking quickly before Jibo was listening (Jibo's eye on the display screen turns blue when Jibo is listening)	child mistake
tap_accident	Child tapped on a button by mistake and skipped a sentence	child mistake

Table B.3: Video annotations in the category of child mistakes.

References

- [1] O. Balet. Inscape an authoring platform for interactive storytelling. In *International Conference on Virtual Storytelling*, pages 176–177. Springer, 2007.
- [2] C. Breazeal, P. L. Harris, D. DeSteno, K. Westlund, M. Jacqueline, L. Dickens, and S. Jeong. Young children treat robots as informants. *Topics in cognitive science*, 8(2):481–491, 2016.
- [3] A. G. Bus, M. H. Van Ijzendoorn, and A. D. Pellegrini. Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of educational research*, 65(1):1–21, 1995.
- [4] J. Chall. *Stages of reading development*. McGraw-Hill, 1983.
- [5] R. Champagnat, G. Delmas, and M. Augeraud. A storytelling model for educational games: Hero’s interactive journey. *International Journal of Technology Enhanced Learning*, 2(1-2):4–20, 2010.
- [6] A. Chang and C. Breazeal. Tinkrbook: shared reading interfaces for storytelling. In *Proceedings of the 10th International Conference on Interaction Design and Children*, pages 145–148. ACM, 2011.
- [7] M. Chatterji. Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the early childhood longitudinal study (ecls) kindergarten to first grade sample. *Journal of Educational Psychology*, 98(3):489, 2006.
- [8] C.-M. Chen. Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, 51(2):787–814, 2008.
- [9] N. Cooc and J. S. Kim. Peer influence on children’s reading skills: A social network analysis of elementary school classrooms. *Journal of Educational Psychology*, 109(5):727, 2017.
- [10] W. A. Foster and M. Miller. Development of the literacy achievement gap: A longitudinal study of kindergarten through third grade. *Language, Speech, and Hearing Services in Schools*, 38(3):173–181, 2007.
- [11] M. Fridin. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & education*, 70:53–64, 2014.
- [12] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children’s second language skills. In *AAAI*, pages 3951–3957, 2016.

- [13] E. Gregory. Sisters and brothers as language and literacy teachers: Synergy between siblings playing and working together. *Journal of Early Childhood Literacy*, 1(3):301–322, 2001.
- [14] T. Hashimoto, H. Kobayashi, A. Polishuk, and I. Verner. Elementary science lesson delivered by robot. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 133–134. IEEE Press, 2013.
- [15] M. Heilman, K. Collins-Thompson, J. Callan, M. Eskenazi, A. Juffs, and L. Wilson. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1):73–98, 2010.
- [16] D. J. Hernandez. Double jeopardy: How third-grade reading skills and poverty influence high school graduation. *Annie E. Casey Foundation*, 2011.
- [17] A. Hiniker, S. Y. Schoenebeck, and J. A. Kientz. Not at the dinner table: Parents’ and children’s perspectives on family technology rules. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1376–1389. ACM, 2016.
- [18] Y.-M. Huang, T.-H. Liang, Y.-N. Su, and N.-S. Chen. Empowering personalized learning with an interactive e-book learning system for elementary school students. *Educational Technology Research and Development*, 60(4):703–722, 2012.
- [19] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1):61–84, 2004.
- [20] C. Kelleher and R. Pausch. Lessons learned from designing a programming system to support middle school girls creating animated stories. In *Visual Languages and Human-Centric Computing, 2006. VL/HCC 2006. IEEE Symposium on*, pages 165–172. IEEE, 2006.
- [21] E. S. Kelley and K. Kinney. Word learning and story comprehension from digital storybooks: Does interaction make a difference? *Journal of Educational Computing Research*, 55(3):410–428, 2017.
- [22] G. Keren and M. Fridin. Kindergarten social assistive robot (kindsar) for children’s geometric thinking and metacognitive development in preschool education: A pilot study. *Computers in Human Behavior*, 35:400–412, 2014.
- [23] M. Kiriakova, K. S. Okamoto, M. Zubarev, and G. Gross. Aiming at a moving target: Pilot testing ebook readers in an urban academic library. *Computers in Libraries*, 30(2):20–24, 2010.
- [24] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool children’s language development. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 643–648. IEEE, 2014.
- [25] J. Kory Westlund, J. J. Lee, L. Plummer, F. Faridi, J. Gray, M. Berlin, H. Quintus-Bosz, R. Hartmann, M. Hess, S. Dyer, et al. Tega: a social robot. In *The Eleventh ACM/IEEE*

- International Conference on Human Robot Interaction*, pages 561–561. IEEE Press, 2016.
- [26] D. Leyzberg, S. Spaulding, and B. Scassellati. Personalizing robot tutors to individuals’ learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 423–430. ACM, 2014.
 - [27] S. B. Neuman. Books make a difference: A study of access to literacy. *Reading Research Quarterly*, 34(3):286–311, 1999.
 - [28] S. Y. Okita, V. Ng-Thow-Hing, and R. Sarvadevabhatla. Learning together: Asimo developing an interactive learning partnership with children. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 1125–1130. IEEE, 2009.
 - [29] F. L. Stoller, N. J. Anderson, W. Grabe, and R. Komiyama. Instructional enhancements to improve students’ reading abilities. In *English Teaching Forum*, volume 51, page 2. ERIC, 2013.
 - [30] C. R. Tamborini, C. Kim, and A. Sakamoto. Education and lifetime earnings in the united states. *Demography*, 52(4):1383–1407, 2015.
 - [31] K. Wall, S. Higgins, and H. Smith. ‘the visual helps me understand the complicated things’: pupil views of teaching and learning with interactive whiteboards. *British journal of educational technology*, 36(5):851–867, 2005.
 - [32] M. Warschauer, D. Grant, G. Del Real, and M. Rousseau. Promoting academic literacy with technology: Successful laptop programs in k-12 schools. *System*, 32(4):525–537, 2004.
 - [33] J. West, K. Denton, and E. Germino-Hausken. America’s kindergartners: Findings from the early childhood longitudinal study, kindergarten class of 1998-99, fall 1998. *U.S. Department of Education, ED Publications*, 2000.
 - [34] K. Westlund, M. Jacqueline, S. Jeong, H. W. Park, S. Ronfard, A. Adhikari, P. L. Harris, D. DeSteno, and C. L. Breazeal. Flat vs. expressive storytelling: Young children’s learning and retention of a social robot’s narrative. *Frontiers in Human Neuroscience*, 11:295, 2017.
 - [35] G. J. Whitehurst and C. J. Lonigan. Child development and emergent literacy. *Child development*, 69(3):848–872, 1998.
 - [36] M. Zipke. Preschoolers explore interactive storybook apps: The effect on word recognition and story comprehension. *Education and Information Technologies*, 22(4):1695–1712, 2017.