Title
Speech-enabled Systems for Language Learning

Author
Tewari, Anuj

Publication Date
2013

Peer reviewed|Thesis/dissertation

Speech-enabled Systems for Language Learning

by

Anuj Tewari

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Canny, Chair
Assistant Professor Bjoern Hartmann
Associate Professor Greg Niemeyer
Associate Professor Laura Sterponi

Spring 2013

Speech-enabled Systems for Language Learning

Abstract

Speech-enabled Systems for Language Learning

by

Anuj Tewari

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor John Canny, Chair

Levels of literacy and the variance in them, continue to be a problem in the world. These problems are ubiquitous in the sense that they change form from developing to developed regions, but do not seize to exist. For example, while teacher absenteeism is a fairly large problem in the developing world, student motivation can pose challenges in the developed world. Prior research has demonstrated that games can serve as an efficient medium in bridging these literacy gaps, generating student motivation (or engagement) not just in short term but also in the long term. This dissertation is dedicated to the investigation and application of spoken language technology to language acquisition contexts in the developed world. We explore the broader research question in two major contexts.

Firstly, lack of proper English pronunciations is a major problem for immigrant population in developed countries like U.S. This poses various problems, including a barrier to entry into mainstream society. Therefore, the first part of the dissertation involves exploration of speech technologies merged with activity-based and arcade-based games to do pronunciation feedback for Hispanic children. This also involves using linguistic theory to determine computational criteria for intelligibility in speech and computational adaptations to reflect them. We also present results from a 3-month long evaluation of this system.

Secondly, a large body of research has shown that the literacy gap between children is well-established before formal schooling begins, and predicts academic performance throughout primary, middle and secondary school. Therefore, in the second part of the dissertation we explore natural interactions for preschoolers that would engage them in game-like activities that involve short follow-up conversations. We explore the design and implementation of a conversational agent called Spot, that acts as a question-answering companion for preschool children. We present a month long study with 20 preschoolers with some insight on the potential, efficiency and usage of such a system. We end with a discussion on computational complexities in building Spot, and rules that it uses to work around speech recognition and natural language understanding errors.

Dedicated to my elder brother Ambuj, for being the greatest source of inspiration.
Dedicated to Maa and Papa, for their unflinching faith in me.

Contents

List of Figures

List of Tables

Acknowledgments

A PhD is a journey of perseverance and the sheer amount of time involved makes it even more challenging. Obviously, such a journey can only be covered with the help of loving and caring individuals. This is my attempt to show gratitude to some of them. This is the least I can do.

First and foremost, I would like to thank my advisor John Canny, for having the faith in me and recruiting me under his tutelage. I came from a relatively unknown undergraduate institution at the time, that didn't enjoy the same reputation as the very well known IITs (Indian Institute of Technology). The fact that John saw potential in me and my undergraduate research, is something I will always appreciate and be thankful for. John has been a great advisor, and someone I have continuously learned from. His area of expertise is extremely broad and this always gave me a sense of safety. I always knew that whatever be the problem at hand, John would have some thoughts about the solution. John and I have sailed fairly smooth through my time in graduate school, and I attribute it to his calm and composed self. There were times when I ran into dead-ends and roadblocks, but John was always able to help me out and also redefine the course to take, if necessary. His repertoire of technical knowledge coupled with his curiosity for human behavior around technology, made him an ideal advisor for this thesis. John has helped me shape myself over the years, and has been extremely understanding of my personal problems. He has supported me through my time at Berkeley, and for that I will always be indebted.

I would also like to thank Matthew Kam, who was also John's student at Berkeley, and was my undergraduate mentor. I worked with Matt on project MILLEE for close to four years, most of that work coming as an undergrad. Matt's work and his tutelage trained me to think like a researcher and critically analyze problems at hand. Matt was (and is) the perfect combination of a strict and yet understanding mentor who leads by example. He set very high standards for me to meet, and I took it as a challenge to meet them. By involving in multiple rounds of field studies in India, paper writing, system design and development, Matt truly prepared me for graduate school. Had it not been for Matt, I never would have realized my true potential. Working with Matt, I became much more confident of myself. I remember being a secluded and quiet young man in my first year of undergraduate studies. From that I transitioned to an outgoing, talkative and social individual who was not afraid of talking to a room full of people. Matt prepared me for graduate school and the professional life to follow, and I am incredibly thankful for that.

I would also like to thank the rest of my committee for their critical yet constructive feedback over the years. Laura Sterponi has been a great committee member over the years. She tends to be a little cynical about technology (as per her own accounts), but that helped me a lot in framing certain motivational components of this thesis. Laura given her background in linguistics and education was an ideal guide for this thesis, and I have greatly enjoyed my interactions with her over the years. The first time I met Laura was as a part of my external minor in education. From her I learned a lot about theories of literacy, and the difference of opinion in the academic world on the definition of literacy itself. I was the only computer

scientist in the class, but Laura was able to create an atmosphere where I felt I could contribute with a unique perspective. She is one of the kindest professors I know or have interacted with and will continue to be my go-to for anything educational. Bjoern Hartmann and Greg Niemeyer have also been great committee members. I met Bjoern fairly late in my graduate school career, but right before my qualifying examination. Since then Bjoern has been a great support and I have always felt very comfortable bouncing ideas off him and getting his opinion on research. Bjoern carries an informal yet professional air around him, which makes him the perfect professor. I will be thankful to him for being so friendly and full of energy. Greg comes into my committee from a completely different research angle, and which makes him a very interesting member. With his knowledge of art practice, new media and technology, Greg is someone who can put the abstract and the practical together with ease. Over the years, I have been amazed by Greg's take on research problems and issues. I still remember him recommending a science fiction novel as my qualifying examination reading. Moreover, the novel was a perfect match for my research work, which explains Greg's importance in my committee.

Talking about friends, I have been lucky to have found friends in school who have stayed by my side through the rest of my journey until now. Gautam Singh and Kapil Godhwani deserve a special mention here. The three of us have been friends for around fifteen years. Gautam and Kapil are friends I have blindly relied on during my PhD. Over numerous phone-calls, chats and hangouts we have shared and dissipated any problems that came my way, professional or personal. Gautam has a knack of making people laugh irrespective of their mood, and it is a quality that has helped me tremendously over the years. Gautam is also the one who made me skydive recently. He is someone I can discuss everything with, from efficient sorting algorithms to discussions about faith, Gautam has been a true friend. I am thankful for friends I can fall back to, I will continue to exploit this feature of our friendship.

In terms of friends at Berkeley, the list is really long and I will try and capture a few important names here. Debanjan Mukherjee has been a great friend throughout my time in Berkeley. For two out of the five years, we have also been roommates. He has supported me through a lot of ups and downs which are common in the course of a PhD. Debanjan is a friend I could talk to for hours, without either of us getting distracted or bored. I will cherish our discussions on pretty much everything under the sun, for years to come. A special mention also goes out to Sharanya Prasad. I only became friend with Sharanya midway through my PhD, but she turned out to be one of the best friends I have ever had. I would also like to thank Avinash, Yasaswini, Mohit, Kranthi, Deepthi and Adarsh for making my time in Berkeley something worth cherishing.

The Berkeley Institute of Design has always been a great place to work ever since I came to Berkeley. For five years, I have seen personnel change from year to year, but some people deserve a mention for making my stay exciting. Andy Carle, Kenrick Kin, Kenghao Chang, Ana Ramirez Chang, Divya Ramachandran, Reza Naima, David Sun, Kristin Stephens, Lora Oehlberg, Wesley Willett, Drew Sabelhaus, Shiry Ginosar, Valkyrie Savage, Celeste Roschuni, and Mark Fuge, thanks for everything. Special mention for Nicholas Kong who has been a great hangout and late night work buddy. Nick's pragmatic approach to life has been

quite inspiring, and it has been a pleasure sharing a cubicle with him. Pablo Paredes has truly been a great friend and lab-mate. Technically I am supposed to be his "peer advisor", but in terms of life experience and ability to deal with stress, I have learned a lot from him.

I have also had the pleasure of two internships, one at Nokia Research and the other one being Microsoft Research. Special mention to RJ Honicky and Kimmo Kuusilinna for mentoring me at Nokia. At Microsoft, it was a great experience to collaborate with Mohit Jain, Indrani Medhi and Ed Cutrell.

A big thanks to all the research collaborators over the years. This PhD wouldn't have been possible without them. Especially, Anuj Kumar, Nitesh Goyal, Karen Baker, Priyanka Reddy, Simon Tan, Matthew Chan, Timothy Price, Ingrid Liu, Carrie Cai, Ozge Samanci and Tim Brown, thanks for working with me, it has been my greatest pleasure.

Special thanks to the Menlo Atherton High School and ECEP (Early Childhood Education Program) for helping us with deployments and being instrumental in making research cycles easy for us.

Any PhD is almost impossible to complete without the support of family and loved ones. I have been lucky to have a family that has always believed in me, even when I was low on confidence. Ambuj, my elder brother, who I have dedicated this thesis to, is the one who has always shown the path forward. Whenever I have been out of ideas related to career and life, I have gone back to him and have never been disappointed. Thanks for being there for me! Words will probably not capture Ambuj's contribution to this thesis, I can just say that had it not been for him, I wouldn't have thought of ever getting a PhD. Ambuj switches between being a friend and being a mentor, with utmost ease, and has truly been a driving force. Thanks to my little niece Paavani and my sister-in-law Shilpa, for being the awesome people that they are. A big and hearty thanks to Maa and Papa for their love and support over the years. Thanks for listening to my numerous rants about failures in research, and thanks for inspiring me to always move on. I am nothing but a reflection of what they are, and the grit, determination and sincerity required to finish a PhD, I attribute to them. My father has been a scientist through his professional life, and my mother has been very active when it comes to educating underprivileged children. I feel that it is not a coincidence that this thesis lies at the intersection of those two domains. My family members have been my greatest friends in all the ups and downs of this PhD, I have tried but mere words will not capture my gratitude.

Last but not the least I would like to thank my fiancé, Devanshi (Meesha). Meesha's ever enthusiastic nature and usual zeal for things has kept me going ever since I have known her. She has been with me through this thesis, in one way or the other. Meesha's calm and composed attitude towards life and her tendency to not bow down in face of trouble, is something that inspires me tremendously. Before I met her, I believed that the capability to understand me and my set of values and principles, doesn't exist in the world outside my family. Meesha proved me wrong, her faith in me and everything I do (or believe) makes her the perfect partner. Moreover, I hope to be the companion who gives what he receives. I am extremely excited about our path forward. I look forward to spending a lifetime of happiness and adventure with my ever enthusiastic and charming partner.

Chapter 1

Introduction

1.1 Background

Levels of literacy and the variance in them, continue to be a problem in the world. These problems are ubiquitous in the sense that they change form from developing to developed regions, but do not seize to exist. For example, while teacher absenteeism is a fairly large problem in the developing world [31], [33] student motivation can pose challenges in the developed world [78].

Prior research has demonstrated that games can serve as an efficient medium in bridging these literacy gaps [33], generating student motivation (or engagement) [34], not just in short term, but also in the long term [40]. This thesis is dedicated to the investigation and application of spoken language technology to language acquisition contexts in the developed world. However, the motivation for this work comes from years of research focused on developing regions [35].

1.2 Prior Work (MILLEE)

Before tackling the problem of language learning in the developed parts of the world, prior research was done on language acquisition related problems in the developing world, as part of project MILLEE (Mobile and Immersive Learning for Literacy in Emerging Economies). Exploratory studies revealed social and infrastructural challenges to using desktop computers to promote learning in school settings. On the other hand, there was a tremendous opportunity for out-of-school learning via educational games on cellphones [33].

In this process, a human-centered design process was followed, in which experienced local English teachers were consulted on instructional and game design aspects. The foundational games that were built went through numerous iterations, including formative evaluations with four communities of rural and urban slums learners in both North and South India. By field-testing with multiple communities, user behaviors with the technology that generalize across

Figure 1.1: Children with MILLEE (Mobile and Immersive Learning for Literacy in Emerging Economies) games

Figure 1.2: A screenshot from a MILLEE game that recreates a scene from a traditional village game.

settings were observed. Through ethnographic studies, factors such as gender and caste were also studied with respect to game-play in everyday rural environments([33], [34]).

Some of the research was also dedicated to tailoring the games to local practices. In particular, the traditional village games were adapted to socially appropriate design to connect more with the target audience. The end-product was therefore not only the games themselves, but a suite of tools and methods for adapting and extending them for local use.

A summative evaluation was also done where 27 students attended an after-school program at a village in Uttar Pradesh, India three times per week over a semester to learn English using such mobile games. Participants exhibited significant post-test gains at the end of this intervention. These learning gains were achieved by combining theory and practice. The games drew on the latest research in language acquisition. Almost 35 successful commercial language learning packages were also reviewed to identify best practices [40].

1.3 Thesis Structure

Even though much of the research mentioned so far was dedicated to solving the problem of language acquisition in the developing world, none of this work dealt with skills related to spoken language. A part of the reason for this was that the technology required to accomplish such complex learning environments is neither ubiquitous, nor cheap. Also the environments encountered in the developing world are fairly challenging (noisy in case of speech) for such systems to perform efficiently. Moreover, proficiency in spoken language is an important component of literacy. Therefore, there is good motivation to test the feasibility of spoken language systems in the developed world, and then optimize them for robustness and cost [35].

Hence, this thesis is dedicated to investigating the use of spoken language technology for language learning. In terms of research questions, the thesis is split into two major parts, as explained in the sections to follow.

1.3.1 Part I: Pronunciation Feedback Technology for Hispanic Children

Lack of proper English pronunciations is a major problem for immigrant population in developed countries like U.S. This poses various problems, including a barrier to entry into mainstream society. This part of the thesis involves exploration of speech technologies merged with activity-based and arcade-based games to do pronunciation feedback for Hispanic children within the U.S. Chapter 2 discusses a 3-month long study with immigrant population in California to investigate and analyze the effectiveness of computer aided pronunciation feedback through games. Furthermore, in chapter 3 linguistic theory is used to determine computational criteria for intelligibility in speech and computational adaptations are proposed and evaluated to reflect them. The output of this phase of research (in addition to the research questions) was also a pronunciation feedback library that includes perceptually relevant characteristics of speech in CAPT (Computer-Aided Pronunciation Training). Research projects like SMART [41] have already started to take inspiration from this research.

While the research mentioned in this part of the thesis involves investigating and building spoken language technology for high school children, research also suggests that greatest impact on literacy will come from interventions at the preschool stage [23]. Therefore, in the second part of the thesis we explore spoken language technology for preschool children.

1.3.2 Part II: Question Answering Technology for Preschoolers

A large body of research has shown that the literacy gap between children is well-established before formal schooling begins, that it is enormous, and that it predicts academic performance throughout primary, middle and secondary school. Indeed rather than closing this gap, there is much evidence that formal schooling exacerbates it: once behind in reading and vocabulary, children read with lower comprehension, learn more slowly and have lower motivation than their more language-able peers. Many national organizations like National early literacy

panel, National Centre for Family Literacy and NIH recognize the essential role of early literacy in a child's later educational and life opportunities.

In this part of the thesis, we try to explore natural interactions for preschoolers that would involve them in game-like activities that involve short follow-up conversations. Chapter 4 establishes a theoretical framework for this research deriving on research from child psychology and adding to it through our qualitative experiments with CHILDES [47]. Chapter 5 is more of a computational approach towards child question-answering. In chapter 5, we try to use clustering techniques and belief propagation algorithms to do object identification in conversational discourses. Chapter 6 builds on the previous two chapters and describes a focused study that involves building a question-answering game for preschool children, called Spot. Chapter 7 discusses the technical framework that was used to build Spot. In essence, chapter 7 presents a framework that can be used to do speech-controlled machinima [44].

Part I

Pronunciation Feedback Technology for Hispanic Children

Chapter 2

Speech and Pronunciation Improvement via Games

2.1 Introduction

"Inclusive Education"[1] is a part of Improving Education Quality, one of the themes of UNESCO. Children belonging to indigenous groups and linguistic minorities are classified as vulnerable to exclusion from the benefits of the education system. Traditionally such minorities have been believed to exist only in the developing and the less developed world. However, statistics and our experiences suggest that such minorities exist even in the developed world. International Bureau of Education (IBE), an international centre for the content of education, is an integral, yet autonomous part of UNESCO and International Academy of Education (IAE). IBE's Teaching Additional Languages booklet classifies "speaking" as an integral part of language learning for additional language learners.

As a contribution to include such minorities and address the challenges involved, we conducted a three month long study with Hispanic immigrant children with limited exposure to spoken English language at a public high school to explore the potential role of pronunciation-feedback coupled games as motivational tools, henceforth referred to as SPRING, to teach and improve pronunciation of immigrant children.

Around 6 percent of the total population in USA is of Mexican origin - authorized or not [8]. About 70 percent of these Mexican born immigrants live in closed communities in just four of the fifty states in USA: California, Texas, Illinois, and Arizona. These communities do not just live together for cultural and social benefits. Their similar economic and financial conditions also bring them closer because the Spanish-speaking immigrant population is almost twice as likely to live in poverty, much higher compared to any other immigrant group. According to MPI's release in February 2010, about three-quarters of Mexican immigrants in 2008 were limited English proficient.

This highlights the plight of Hispanic, and specifically Mexican, immigrant cultural and

[1] This work was done in collaboration with Nitesh Goyal

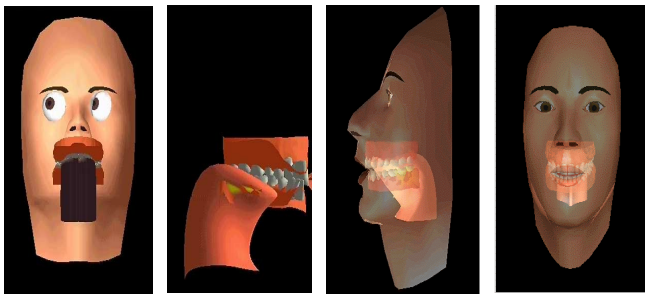

Figure 2.1: Massaro's Baldy showing a cross section of the anatomical movements while pronouncing

linguistic minority living in USA, one of the most developed countries. Evidently, this community suffers from exclusion of benefits of the infrastructure and society available in the developed world. Moreover, their lack of knowledge of primary language of communication: English hampers prospects of improvement.

Thus, this line of work focuses on the age group of 12-18 year old immigrant Hispanic children by employing games similar to the games that they already enjoy playing as an aid to the existing classroom teaching in English Language Learners (ELL) classes.

2.2 Related Work

Computer Assisted Language Learning (CALL) has existed for almost 70 years now. Several methods and systems have been proposed to help improve particular focus areas in language learning using computers. Most work in the CALL domain does not explore the ability of technology to teach English pronunciation using persuasive computer games to immigrant high school children.

Horowitz et al. [27] describes an 8-week long study that promotes literacy in USA with participants from households below the poverty line. The focus of this study was to improve literacy and teach the English alphabet using videos. While the videos were persuasive, they lacked focus on improving the English pronunciation and were targeted at very young children.

Massaro in 2003 [51] described Baldy (Figure 2.1), a virtual talking head on a screen with focus on helping users learn how to pronounce the phonemes properly as a virtual teacher. This is by far the only study we know that focuses on teaching pronunciation and provided a visually detailed feedback and training. Powers et al [64] (Figure 2.2) is also a similar system and goes a step further by acting as Embodied Conversational Agents (ECA). These systems improve upon Massaro by including other features like vocabulary learning etc. While these systems mention encouraging results, they lack information about how motivational these systems might be.

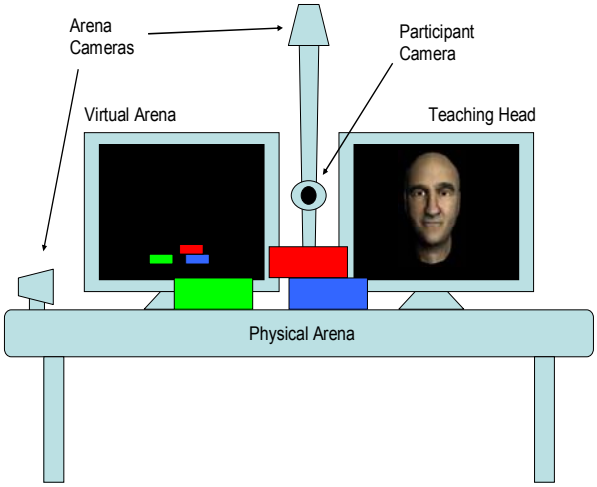

Figure 2.2: Powers et al's system installation showing the ECA connected to a camera

Multimodality has also been briefly investigated for pedagogical benefits in English Language Learning. Chen Yu et al [83] suggests that spoken language can be grounded sensory perceptions of the real world. It describes a learning interface that bridges a gap between the real world physical objects and the virtual interface. Sluis et al [75] describes a collaborative table top based simple matching to help develop the reading skills of young groups of children. Fallahkhair et al [15] describes a system with 2 inter coupled-interfaces: TV as an audio visual aid, and mobile phone as a supporting aid to help learners learn the vocabulary. These systems also continue to focus on the writing, reading, and vocabulary parts of the language.

However, recently there has been a growing interest in including computer based tools that use automated speech recognition to provide a guided reading experience for the users. Mostow et al's Project LISTEN based Reading Tutor [56], [57], [55] has been used with a variety of audiences in improving the English reading ability of children, with English as a first language and with English as a second language (ESL/ELL) in USA by Poulsen et al [63] and Canada by Reeder et al [65].

While the Reading Tutor involves use of stories, Kam et al [32] has successfully shown the use of games, especially the use of mobile games as persuasive tools for improving the English literacy of the illiterate English as Second Language (ESL/ELL) children in India. Johnson et al in 2005 [30] and 2007 [29] present a system being used by the US Army to learn Arabic in Iraq. However, we feel that due to nature of the intended use and lack of a particular pronunciation focus, this product is unsuitable for use by young children.

Anna et al's DEAL [26] uses both the ECA and task-based game design in its system. The users hence, learn how to structure the sentences properly and learn appropriate word placements. This system is focused more on the grammar than spoken language.

As explained above, the existing works successfully describe using games or speech recognition or both for literacy improvement. However, unlike our system, none of them employs usage of both games and speech recognition for pronunciation improvement amongst English as Language Learner (ELL) children.

2.3 Overview of Pilot Study

To investigate the problem mentioned above, a pilot study was carried out at a public high school located in a highly populated Hispanic immigrant location in California, USA for about three months from December, 2009 to March, 2010. The study took place within the school premises during the extended school timings and involved demographic study, the pre-test, the experiment, and the post-test with permissions from the school authorities, the teachers involved, the students and/or their parents.

Three sessions were held, on an average, per week for four weeks. Each session accommodated approximately three students, one after the other. So, each student played freely in seclusion from the other students for about ten minutes per week. There were two different games that each student was able to play. These games were alternated each week to keep up the interest level of the students. Hence, during the four week long endeavor, each student received a total of about forty minutes worth of play time from the two games of SPRING.

2.4 Study Locale and Setup

This section describes the steps we followed to find our user group. We began by contacting the teachers and school authorities at several public middle and high schools located in the vicinity. Our aim was to locate a school with a high immigrant population having a low level of spoken English fluency. Based upon the anonymous demographic and diversity data that we received from these schools, we shortlisted three schools where we submitted a request for conducting research with the students within the school premises. One of the public High Schools that accepted our request fulfilled our requirements. According to a data survey by the school district in 2007-08, an overall 50% attrition for ELL was reported. For this particular High School, the rate was 75% for ELL. We were guided to one of the English Language Learner's (ELL) classes at this school. The class consisted of 20 students, at ELL level 2. These students had been in USA for less than two years, and had over the time attended, and cleared ELL level 1. The class had a 100 percent immigrant population. In this class, 90% immigrants were from Latin America. Of the students in this class, 95% are labeled as SED (Socio-Economically Disadvantaged). That means one of two things (or both things) is true of all but one student. Either (1) the students are living at an economic level qualifying them for the federal free/reduced lunch program or (2) his/her parents did not graduate from high school, or both are true. This situation at this school compares favorably

with the previously quoted national data. So, we decided to choose this particular ELL class at this school.

2.5 Data Collection

The pilot study was managed solely by the four researchers involved in this project. However, due to the nature of the participants, a local member of the school volunteered to help translate between English and Spanish for children who could not understand our use of English language.

The class had a strength of 20 students. We divided this class, for the purpose of our study, into two groups of 10 students each. One of the groups was the CONTROL GROUP, which received the regular classroom training from the teacher and did not attend the play sessions with SPRING. The other half, EXPERIMENT GROUP, received exactly the same training in the classroom as the CONTROL GROUP. However, they also received the opportunity to attend play sessions with SPRING.

To reduce any bias due to pre-existing knowledge between the two groups, we randomly picked and assigned the students to either of the two groups. Next, we administered a simple qualifying test to all the 20 participants to gather their existing level of knowledge. The test consisted of a slide-show of 30 words, one after the other, on a computer. The test taker was required to speak the word shown on the screen and a speech recognition engine (discussed later) recorded and scored the utterance. The scores were not made visible to the students to reduce anxiety. The tests were done in private with each student to minimize any learning effects. The words selected for the test were kept constant for the entire pool of the participants and were selected from the syllabus and the recommended textbook for that class. These sessions were also audio recorded. During the course of the study, we evaluated the participants using a similar test to prevent test anxiety and for consistent comparable results. These were administered as a series of pre-tests and post-tests.

Each play session with SPRING was video taped to record the emotional state of the participants while playing. This was captured by facial and body expressions, exclamations, sighs, gasps and other auditory feedback. These recordings created the contextual data by providing us with more data about the play-ability of the different stages, elements and parts of our games.

2.6 Participants

This pilot study was one of the first kinds to be established at our partner school, especially with the immigrant population. So, our participants were very new to this new arrangement and we benefited from their enthusiasm to participate in "something new". Initially, in total we obtained consent from 20 children and/or their parents to participate in the study. They were all part of the same ELL Grade 2 class at the school and represented the total strength

Figure 2.3: A participant playing one of the games.

of the class, as well. We began our pilot study with all the 20 of them. However, during the due course of time, 2 of them left the study. Unfortunately, the reasons for attrition could not be conclusively determined due to their continuous absence from the school itself during the three month long duration. However, reasons of attrition, after consultation with teacher, seemed to be family and financial problems for the male participant, and teen-age pregnancy for the girl participant.

2.6.1 Demographics

The 18 students (after attrition of 2 from 20) exhibited the following characteristics:

- Six (6) were male and twelve (12) were females.
- All eighteen (18) in the study were in ELL level 2.
- The students were in the age range of 14 to 17.
- All eighteen (18) students were of Hispanic ethnicity

Many of them lived with family members such as uncles, aunts, and cousins; some did not live with their mothers or fathers. The fathers, uncles, and brothers held jobs working in a market, as a florist, washing cars, as a gardener, or other lower-end jobs. Few had younger/older brothers or sisters still in school.The mothers, aunts, and sisters had jobs that involved cleaning homes, babysitting, or no job at all.

Amongst the 18 children, many had ambitions of becoming a lawyer/attorney, doctors, teacher etc. 16 of the 18 students either had a cell phone or had access to a cell phone (from

a family member) and only use it for texting or talking on the phone; none play the games on the phone. When asked about what kind of games they played, students listed board games such as checkers to several PlayStation games such as soccer (FIFA), Boxing, racing games, Mario, or some computer games. There were a small number of students who didn't play games at all, too. None of them knew about Guitar Hero.

When it comes to learning English, all the students pointed out vocabulary acquisition and pronunciation/speaking as their key issues; other issues were reading and writing. All the students except one recognize the importance of learning of English, so they can attain better job prospects and communicate better. However, the teachers also mentioned that there is some resistance to learning English because these students are surrounded by a community of other Spanish-speaking peers and lowers their incentive to learn. These students also mentioned peer pressure because they did not want to sound silly when they mispronounce English words. Evidently, there are issues with intrinsic motivation.

While lack of intrinsic motivation is a discouraging factor, the extrinsic motivation is also lacking. While the children want to succeed and aim high for their life, there are not many good examples available in their community. Furthermore, for illegal immigrants, avenues for higher education and professional growth are virtually non-existent. This reduces the motivation of some of the students to try harder because they know that they will eventually get low skilled and low-waged jobs like their parents.

2.7 Design

This sections describes how we designed our study and the associated apparatus and content for a successful implementation. We begin by explaining the current curriculum taught at the school to our user group and how we derived a syllabus for our study. Next we explain the methodology behind our game designs and end with a description of the implementation and system design.

2.7.1 Curriculum Design

A student in the ELL Level 2 spends roughly 3 hours in the ELL classroom daily. This includes instruction and teaching, drills, practice sessions, silent readings, and tutor-time. We developed our curriculum worth teaching 7 percent of the entire vocabulary, for the entire academic year, in about 10 minutes session once a week after discussing with the ELL teacher for the class. This represents quite a negligible self-learning time. The students at ELL Level 2 at the chosen school attend classroom teaching by an experienced teacher, aided by audio-visual media to improve the attention and understanding. They follow the curriculum designed according to the textbook "Milestones California Edition". The book is divided into six units, each describing a different facet of life like "Dreams", and "Survival" etc. This curriculum is heavily based on reading, vocabulary, and grammar lessons in content, and exercises. However, it offers limited opportunities to speak English formally. Each ELL

level requires a certain minimum level of knowledge of English vocabulary. These words are discussed in the class but spoken and pronunciation correction drills of these words do not happen at the class or an individual level. The only opportunity that these children have at listening these words are when used by the teacher in the class during the discussions.

The "Milestones" book includes a list of about 300 words from the 6 units that are expected to be known by the students at the end of the academic year. We divided these 6 units into 3 parts: Group A: Units 1 and 2 which had been taught by the teacher in the class before we began the study; Group B: Units 3 and 4 which were being taught during the study; Group C: Units 5 and 6 which had not been taught during the duration of the study. We randomly chose 10 words out of each Group (A, B, and C), giving 30 words, a 10% sample set out of the pool of 300 words and created a syllabus of our study based on them. The aim to divide the words into the groups was to investigate if the games caused significant deviation between the learning gains of preexisting knowledge (Group A), or unknown knowledge (Group C), or aided what is being taught (Group B). Some of the words in the sample set included "Menacing", "Attic" and "Soggy" etc.

The study was designed to test the pronunciation ability of this sample set of words by the users, teach the users how to pronounce those words using a game, and then finally testing to detect the effects, if any.

2.7.2 System Design

The entire game logic for SPRING was written in Flash ActionScript. SPRING was eventually deployed on an Ubuntu Linux installation. Details of the individual pieces are as follow:

Speech recognizer

For the purposes of the speech recognition, we used the CMU Sphinx-III speech recognition engine. However, instead of using it in decoding mode we used it in forced-alignment mode. In force-alignment, rather than being given a set of possible words to search for, the search engine in the recognizer is given an exact transcription of what is being spoken in the speech data. A reason for using the force-alignment mode was that we were able to obtain scores at the level of individual phonetic units. We used this information to point out which part of a particular word was uttered incorrectly.

Speaker adaptation

Since we wanted the games to give feedback after comparing to standard American accented pronunciations, we trained the recognizer on large corpora of data (15GB, raw format) from American accented speakers. However, we did account for the change in texture from a male to female voice. We recorded audio utterances from 2 American males and 2 American females and used MLLR (Maximum Likelihood Linear Regression) transforms to adapt the

recognizer to male and female voices as and when required. Use of MLLR transforms is the most commonly used method for speaker adaptation in automatic speech recognition systems.

Feedback routine

The recognizer could generate acoustic scores, but they had to be compared against standard American accented pronunciations, before giving feedback to the participants on how they did on a particular word utterance. Therefore, we coded a library that returned back a Likert scale (1-3) rating for each phonetic unit in the word under consideration. This rating could then be used to give feedback to the participant.

Graduated interval recall

The game logic for Voz.Guitar was implemented in a way that the syllabus queue was chosen according to a well-established algorithm called Graduated Interval Recall. [62] The algorithm helps in determining the order of the questions, given a syllabus. It is modeled in a way that performance on a particular question determines the number of times it will be posed in the near future, thereby causing long-term retention of syllabus items. The game concept of Voz.Guitar had an aspect of repetition, as opposed to Zorro, which allowed the player to explore an exciting but static and predefined game world. Therefore we just used the algorithm for Voz.Guitar and not for Zorro. However, we countered the lack of repetition in Zorro, by making the participants play the game again. Moreover, we ensured that the time for which the participants are getting instructed (also playing the game) stays constant across the two games.

2.7.3 Game Design

The aim of the study was to design and create games, enriched with pedagogy that might motivate the players to play them, despite the challenges posed by the learning material in the game. We based the design of our games on the following resources:

- Demographic interviews of the children clearly indicated a penchant for certain types of games.
- Popular and best selling commercial software available in the market.

This gave us an advantage of creating games that were likely to peak interest of the children while they incorporated the best practices of game design and elements from existing games. Using this knowledge, we decided to create two games: Zorro (based on Mario), and Voz.Guitar (based on Guitar Hero) for chiefly the following reasons:

- Activity based vs. Arcade based: The demographic interviews pointed out the predilection for two different genres: card games and action games. However, in either genre, the children preferred fast paced, non-time restricted gaming sessions.

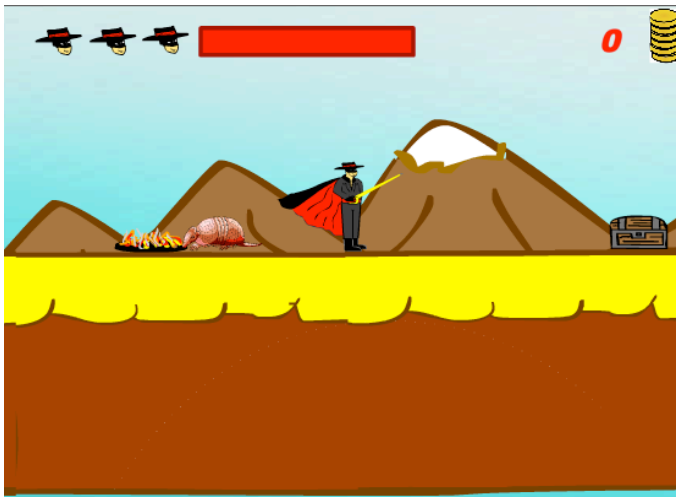

Figure 2.4: Screenshot from Zorro game.

- Novel vs. Comfortable: We based our design on two popular and proven games: Mario and Guitar Hero. The demographic interviews indicated the previous playing experience of most of the participants with Mario, while none knew about Guitar Hero. So, we decided to give them a mix of a comforting known game and a novel, and hopefully exciting, game.

- Adaptive vs. Non Adaptive: We chose Mario based game because it is non-adaptive and gives a consistent experience of play, with onus on the player to act fast. On the other hand, Guitar Hero based game was adaptive and had an element of surprise.

Both games followed the principle of teaching, drill, immediate feedback, scores, and repetition. Both games feature the word, associated playable American accented female voice, and spelled-out-pronunciation to aid the users. The spelled-out-pronunciations were obtained from the online dictionaries [20] and then modified accordingly by a trained linguist with five years of experience.

Zorro, as shown in Figure 2.4, is a character based arcade game, which involves moving Zorro, the main character, of the game from left to the right of the scene using arrow keys until he reaches the castle.

As shown in Figure 2.5, on the way, he encounters five closed chests, dangerous animals, tricky terrain, and obstacles, which must be overcome. The obstacles can only be overcome by opening up the chests. Each chest contains a word, associated pronunciation, and the associated audio pronunciation coupled animated spelled-out-pronunciation. The word is pronounced three times every time it is played. Next, the user gets an opportunity to record their pronunciation of the word by the click of a button.

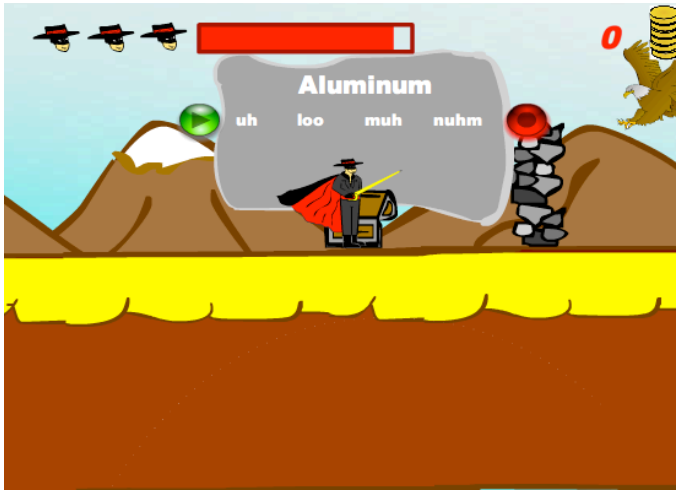

Figure 2.5: Zorro opening a box and revealing a syllabus item.

Figure 2.6: Feedback on the quality of pronunciation.

As shown in Figure 2.6, a feedback screen shows the correct and wrong parts of the pronunciation, and the associated score follows this. She also hears her own pronunciation and the intended pronunciation. After crossing the five obstacles by practicing the five words and avoiding the deadly animals, the user wins the game. In case, she finished short of 10 minutes, she is obliged to play the game again.

Voz.Guitar, as shown in Figure 2.7, is an activity-based game that displays the word, associated spelled-out pronunciation, and plays the associated pronunciation.

Next, it allows the users to hit the falling alphabets of the words at the right time. Next, the user is obliged to pronounce the word, as shown in Figure 2.8.

The feedback screen displays the hits and misses in the spelled-out-pronunciation and corresponding errors. The user hears her pronunciation followed by the intended pronuncia-

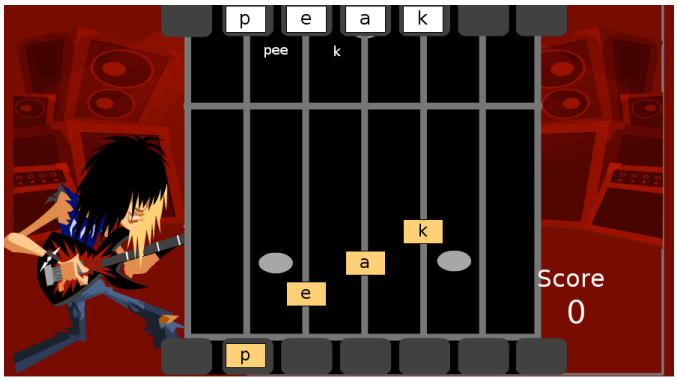

Figure 2.7: Screenshot of Voz.Guitar

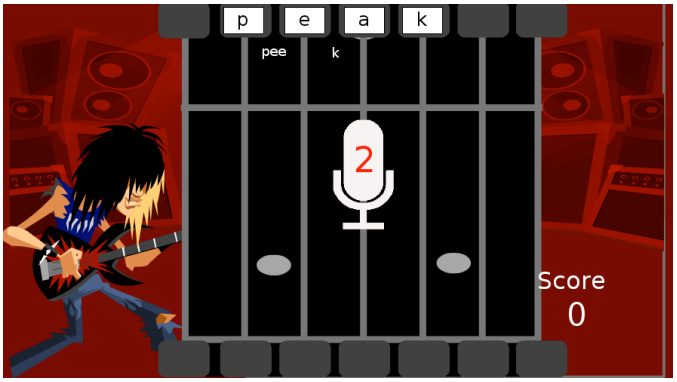

Figure 2.8: The mic icon was the cue for users to pronounce a word.

tion. The game is adaptive and hence, tends to automatically repeat the words, which have not received a satisfactory pronunciation response from the players, as shown in Figure 2.9. Each positive utterance increases the score of the users. The session continues until the time limit of the session reaches.

2.8 Study Sessions

As previously mentioned, the study was designed across three groups of words, for two sample sets of population: Control Group, and Experimental Group.

The sessions lasted for around two hours per day, three times per week, and four weeks in a row. There were two types of sessions: Pre/Post Test sessions, and Learning sessions. A 2-hour Learning session was typically structured as follows: preparing the game database with the pre-determined group of words (A, or C), arrival at the school premises, arrangement

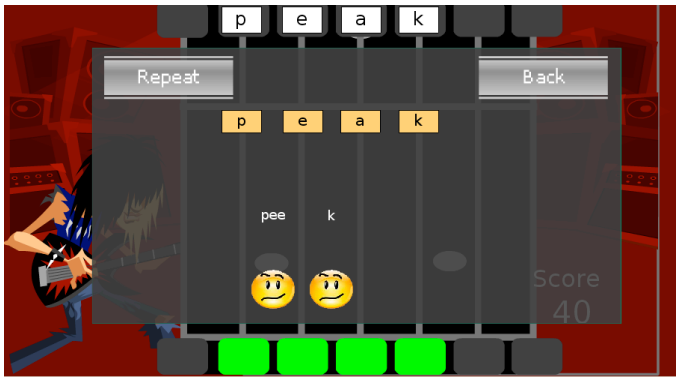

Figure 2.9: Feedback in Voz.Guitar

Control Group		Treatment Group	
CG1	1.24%	TG1	0.26%
CG2	-1.72%	TG2	0.67%
CG3	1.15%	TG3	1.02%
CG4	0.71%	TG4	-0.13%
CG5	1.02%	TG5	3.14%
CG6	-7.40%	TG6	1.32%
CG7	-1.68%	TG7	-0.12%
CG8	0.71%	TG8	1.43%
CG9	-0.12%	TG9	5.14%

Table 2.1: Acoustic score gain percentages for control and treatment group

and setup at a quiet location in one of the pre-arranged labs, greeting with the teacher, a list exchange of the students needed for that day, escorting a student to the lab, explanation of how the game is played, the goals, and a demonstration, the gaming/learning session for 10 minutes, a post game session qualitative interview, escorted return of the student, and bringing back the next student. A Pre/Post Test session involved the same as above except the student faced the test instead of the gaming session.

2.9 Quantitative Observations and Findings

2.9.1 Metrics

Before we go on to explain our quantitative findings from the experiment, we need to define the metrics that we used to gauge the change in pronunciation. We used the following two metrics:

- Acoustic score gain percentages (ASGP): These were numerical scores generated by the CMU Sphinx-III speech recognizer. We did a batch decoding of all the audio samples (pre-test and post-test) that we had from the participants and generated acoustic scores to quantify the quality of pronunciation. The acoustic scores take all aspects of spoken language into account (like intonation, fluency, clarity etc). Moreover, the acoustic scores were generated for each phonetic unit in a word, and hence judge the actual quality of each phonetic unit. These individual phonetic unit level scores can be added together to generate word level scores. The ASGP for each participant was calculated as follows:

$$\frac{(Total\ Posttest\ Score\ -\ Total\ Pretest\ Score) * 100}{Total\ Pretest\ Score} \tag{2.1}$$

- Word gain (WG): The word gain was nothing but the difference in the number of words that the recognizer could decode during the pre-test and the post-test. In simple words, this metric is a high-level representation of the number of words a participant learned to pronounce (with acceptable pronunciation) during the course of the experiment.

It should be noted that we had initially divided the 20 words in our curriculum (that was taught), into two parts. As explained earlier the first part came from pool of words they had already encountered in class and the second part came from pool of words that were completely unfamiliar to them. When we analyzed our post-test and pre-test data, we realized that the correlation between the category (familiar or unfamiliar) of the word and average gain on the same over the duration of the experiment was negligible. Quantitatively speaking, the correlations between the average scores (both ASGP and WG) across all participants and the category (familiar/unfamiliar) was $<=0.27$ for all the 20 words. Moreover, this was true for both, control and the treatment group. Hence we decided to group our results together, and analyze the gains across all the 20 words.

2.9.2 Post-test gains

In each experiment, we used a standard statistical t-test to compare the gains of the treatment and the control group. This test yields a p-value indicating how significant the difference is between the means of the two groups. A two-tailed t-test on the pre-test scores of the treatment and the control group yielded a p-value of 0.25, which shows that there was not a statistically significant difference between the means of the two groups before the start of the experiment.

Acoustic Score Gain Percentages (ASGP)

After the post-test, the mean acoustic score gain percentage for the control group was -0.68 (σ=2.77, n=9) and that for the treatment group was 1.41 (σ=1.72, n=9). The ASGP are small numbers because they are percentages of total pre-test scores across 20 words (more

Participant ID	Number of words attempted in pretest	Number of words attempted in posttest	Word Gain
CG1	13	14	1.00
CG2	15	15	0.00
CG3	16	15	-1.00
CG4	12	12	0.00
CG5	16	16	0.00
CG6	17	16	-1.00
CG7	19	19	0.00
CG8	12	12	0.00
CG9	17	18	1.00

Table 2.2: Word Gain for control group

than 110 phonetic units). However, a two-tailed t-test between the ASGP for the control and the treatment group yielded a statistically significant p-value of 0.08. Table 2.1 lists the ASGP for participants in the control and treatment group. A negative percentage denotes that the participant's total acoustic score for the post-test was lower than her acoustic score for the pre-test, and therefore the increase was actually negative.

Word Gains (WG)

After the post-test, the word gain scores had a mean of 0 (σ=0.71, n=9) for the control group and a mean of 1.11 (σ=1.54, n=9) for the treatment group. This gain was in addition to the improvement in the quality of the pronunciations that is represented by the ASGP. Tables 2.2 and 2.3 list out the words attempted in pre-test, post-test and the resulting WG for the control and the treatment groups.

There wasn't a significant difference in the number of words the control and the treatment group could pronounce to some extent at the start of the experiment. The t-test on the number of words attempted at the start of the experiment yielded a value of 0.42.

However, the t-test on the number of words attempted at the end of the experiment (by the control and the treatment group) yielded a value of 0.06, which shows a statistically significant difference. It also points to a possible confidence boost during the study in terms of pronouncing less familiar and more complex words. Moreover, a two-tailed t-test on the WG values for the control and the treatment group yielded a p-value of 0.07. This shows that there was a statistically significant difference between the WG of the control and the treatment group.

Participant ID	Number of words attempted in pretest	Number of words attempted in posttest	Word Gain
TG1	20	18	-2.00
TG2	19	20	1.00
TG3	16	17	1.00
TG4	15	16	1.00
TG5	12	15	3.00
TG6	13	14	1.00
TG7	17	20	3.00
TG8	19	19	0.00
TG9	15	17	2.00

Table 2.3: Word Gain for treatment group

2.9.3 Gender Related Findings

Our control and treatment group had the same distribution in terms of gender. Therefore, we also did some analysis to quantitatively measure the influence of gender on game play and learning. The correlation between gender and ASGP for the control group (0.65) suggests that boys performed worse than the girls overall, over the period of the experiment. However, the correlation between gender and ASGP for the treatment group (0.32) suggests that gender did not influence the improvement in pronunciation quality that was exhibited by the participants, after playing the games. This is in contrast to the findings from similar ESL (English as a Second Language) acquisition studies in the more underprivileged parts of the world [32].

2.9.4 Effects of pre-test on post-test gains

The correlation between pre-test scores and ASGP for the treatment group was 0.11 and the correlation between pre-test scores and WG was 0.25. This shows that the participants in the study showed similar learning gains across both metrics irrespective of their performance on the pre-test. Therefore, there was no notion of bimodality as suggested by similar ESL acquisition studies in developing parts of the world [32]. This might be happening due to various different factors like better ESL levels, prior exposure to technology, and access to education.

2.9.5 Learning gains during game play

We also collected data logs of how the participants performed during a session. Across a total of 10 (one of them dropped out of the school before the post-test) participants and a total of 40 game sessions the treatment group exhibited an average ASGP of 12%. Calculating the differences in acoustic scores of the first and last instance of a particular word in a single

game session and averaging it across all participants in the treatment group resulted in these percentages.

2.10 Qualitative Observations and Findings

In addition to the quantitative sources of data, we also had videos that served as an important part of the analysis. We recorded approximately 600 minutes (10 hours of video). After transcription and qualitative coding of the data we came up with the following major qualitative findings:

2.10.1 Player profiles

Through the duration of our study, we observed several key distinctions in our pool of subjects. The first major separation appeared in gender difference. The females appeared to be indifferent to the game play and were more focused on the speech/voice features of the game. Females also needed more assistance with the games compared to males, whether it were additional verbal cues or helping them with the obstacles in the game. When a translator was used for one female, the two put together were more engaged with playing both Zorro and Voz.Guitar; the two laughed, gestured and were more focused on game play in addition to the speech/voice features.

The males were more focused on game play than the speech/voice features; for example, when they opened the chest with Zorro and the feedback screen appeared, the males were still playing with the Zorro character (trying to move it around). When the males did interact with the speech/voice features, they said the words with more confidence than the females and had less stuttering and hesitation.

For both male and females, they exhibited a certain learning curve when playing, and that was true for both games. Almost none of the players used the "repeat" feature in the Zorro game, as opposed to Voz.Guitar that forced them by having them go through each word again. In addition, although both genders found the games entertaining as a whole, they did occasionally display gestures of frustration including rolling their eyes and hand waving to brush off mistakes. We felt that these gestures were partly attributed to general game playing and demonstrate the student's attention and involvement in the game, which is a positive factor.

We further broke down our subject pool and found four specific player profile classifications in the subject population. They are represented by the following four names: Pablo, Juna, Estera, Sandra. And we suggest some design decisions to be kept in mind for future designs to create an inclusive game in Table 2.4.

Name	Sex	Likes to play games	Body language while playing	Game involvement	Pedagogic involvement	Game design suggestions
Pablo	Male	Yes	Active, Focused, Nimble	High: Focused on Scores	Low, Bypasses the learning	Game requires higher percentage of accuracy to bypass the pedagogy elements
Juna	Female	Yes	Indifferent	Little: Takes a call on phone during the play	Low	Other Genre Games like Shopping Spree, Pop culture, Dressing up
Estera	Female	Yes	Not too excited	High: If game is socially interactive	Average	Online Social Networked Games with discussions, and chatting
Sandra	Female	No	Confused	Frustrated: Loses Lives constantly	Average	Games with easier levels, and abundant practice for learning game controls

Table 2.4: Player profiles: game design suggestions

2.10.2 Pronunciation Measures

We tested and identified pronunciation measures using a speech recognizer. Since all the processing was happening off the field and on a dedicated machine, we got accurate scores. However, to bring more credibility and to add more human aspect to our research, we would like to seek help from trained linguists. Our overarching goal is to better the pronunciations to a level that is socially acceptable. Using the recognizer for evaluation is the first step, but using human inputs from various different sources would be beneficial. This is exactly what is done and discussed in the next chapter.

2.10.3 Other Findings

In the post-game play session interview, 7 out of 9 participants reported that they felt they were learning pronunciations during the game, the rest said they did not know if they learned.

We also asked who they would want to help them with pronunciations. 6 out of 9 participants said they would want help from both their teacher and SPRING. The rest of the 3 participants said they would want to learn from the game only. Since this is self-reported data, we don't attach a lot of value to it, but it definitely points out that SPRING was a pleasant change for a majority of the students. When asked which game they enjoyed more, 6 out of 9 said they liked Zorro better than Voz.Guitar, the rest of the 3 said the opposite. This was intuitive because Voz.Guitar had a lot of repetition and Zorro was exciting. We would want to mix these two factors in the next phases of the study. It would have been hard to mix game play and pedagogical concepts right from the start, but now we can use the current phase of study and the design decisions we took to inform the next phase of the study and design.

2.11 Challenges Faced

2.11.1 Motivation

The community we worked with was a very complex one. There was little or no motivation for them to acquire English as a Second Language. Through our games, we were trying to break this barrier to entry. Our aim was to develop games that are inherently more engaging and have pedagogical concepts merged into game concepts. Throughout the duration of the study we constantly tried to keep up the interest levels of the students we were working with. This was generally done through interface changes. We made sure that we modify any interface element that causes a loss of perception, or is frustrating to the participants. This required iterative design and rapid prototyping.

2.11.2 Technical challenges with Speech

Use of speech had a lot of attached technical challenges to it. As discussed earlier, male and female voices were hard to adapt to, but it was accomplished by using MLLR transforms for speaker adaptation. Speech recognition systems are generally very sensitive to background noise and environment changes; therefore we had to be careful about keeping the environment constant and stable across various sessions. We also used a high quality noise reduction microphone to capture audio during game play and during tests, to minimize effects of background noise.

2.12 Future Directions

2.12.1 Conversational agents and adaptive games

As stated earlier, the community we worked with did not necessarily have an intrinsic motivation to learn English. Therefore similar future ventures should involve efforts towards motivational games. Emotional analysis of speech can be used to detect the motivation levels

or emotional states of the children. This information can then be used to start emotional conversations with the students. Games with conversational agents seem to be a good fit in such cases.

Our qualitative results point to some common player profiles that we observed during the game sessions. Future research could cater to all the profiles through our future games. There is a possibility of developing adaptive games, which try to gauge the profile of the player based on her interactions with the game and match that accordingly.

2.12.2 Context-based Games

Moreover, we found that a "one size fits all" approach doesn't work in term of gender. Therefore, there is a need for games that have multiple story lines, characters, goals, reward structures and endings. In such cases interactive fiction seems like a good fit, where story, characters and plots could change based on the personality of the player. The kind of decision he/she takes in a game session would then determine the overall direction of the game. This would result in games that are still "one size fits all", but are considerate of gender, context and culture.

Our qualitative findings suggest that there was demand for multiplayer or collaborative games. Future research should try to explore implications of speech and games in the domain of shared learning. In such games, the players can collaborate and help each other with pronunciations.

2.12.3 Mobile Devices

Future research could also look into the domain of mobile devices. With the increase in the processing power of the phones, it is possible to run speech recognizers on cellphones. We have already ported the CMU Sphinx-III speech recognition engine to mobile devices (Nokia N810), and the performance is comparable to the computer version (average time taken in decoding one word on Sphinx III is 0.92 seconds, and average time taken in decoding a word on the ported mobile version is 2.2 seconds).

2.13 Conclusion

This chapter presented work that is aimed at the use of educational games for pronunciation feedback for Hispanic children at a high-school in California. The aim was to develop general English language competency in students, and this certainly includes being able to speak English intelligibly. In this chapter we just focused on quality of pronunciation based on raw acoustic scores. Of course, there are many aspects to learning a foreign language. Acquisition of vocabulary, knowledge of grammatical rules and structures, and learning about cultural norms and traditions are all important parts of the process. However, there is also pronunciation of the target language and oral proficiency. Students often begin learning a

foreign language with their peers (most of whom are also lacking in experience with the FL), and regardless of mispronunciation, they are usually understandable to the teacher and class. However, at some point, a student may find himself or herself in a face-to- face interaction with a native speaker. If the student has not learned proper pronunciation and is not orally proficient, he or she could be incomprehensible to the speaker. We all want to be understood when we speak, and proper pronunciation allows for that. Intelligibility is critical to face-to-face conversational exchanges. There are strong relationships between oral proficiency and literacy, and we feel we should simply focus on the value of oral proficiency in the country's native language.

As a part of this work, two games that derived inspiration from popular games like Mario and Guitar Hero were implemented and deployed. The games were interfaced with a speech recognizer that could measure the quality of pronunciations and give feedback on the same. Therefore, the next chapter contributes to development of evaluation methods that are based on perceptually important criteria for intelligibility, not just on recognizer scores. Eventually we build these annotations into our feedback mechanisms with the goal of producing automated feedback to students that approximates the intelligibility of their pronunciation, not just recognizer generated acoustic scores (acoustic scores are log-likelihood probabilities). We also present some preliminary evaluation of these optimizations, and also results from the same, in the next chapter.

Chapter 3

Optimizing pronunciation feedback for perceptual characteristics

3.1 Introduction

As we may recall from the last chapter our participants were specifically Hispanic children from low-SES backgrounds studying in a English Language Learners program. The age range of the participants was 12-18 years. All the participants were in ELL level 2, therefore their pre-existing knowledge of English was not strong. Our study lasted for three months, and we spent the first few weeks trying to understand the demographics and needs of the participants. We divided the class of 20 into a control group (10 students) and an experiment group (10 students). The experiment group went through both the games, where each game had a syllabus of 10 words. Each participant spent 40 minutes per game over a month. This means that each participant went through 80 minutes of game play for 20 words in total.

The participants also appeared for a pre-test and a post-test on the syllabus for the two games. These pre-test and post-test utterances were individually evaluated by two trained linguists with more than 10 years of experience. However, this process was blind, as the evaluators did not have any information on which test the recordings belonged to or who the participant was. The details of the experiment and its computational evaluation were discussed in the last chapter. The linguistic evaluation and the resulting computational adaptations are discussed in this chapter.

3.2 Background

From a linguistic standpoint, the each speech sample contains encoded in it phonetic data that is measurable as sound waves. Using acoustic software, e.g. Praat, specific characteristics of an individual participant's pronunciation of a test word can be analyzed for characteristic wave pattern features. The human voice produces a variety of acoustic disturbances, most of which cluster in structures called formants visible in spectrographs, i.e. waveform print outs

made accessible by the software. Each individual's pronunciation of a particular word differs slightly from another's pronunciation, as well as from the same individual's pronunciation of the same word at a later time. The spectrogram is capable of measuring and displaying such differences as a matter of time and frequency. The primary structures needed for measurement are the first three formants, typically labeled F_0, F_1 and F_2. These waveform groupings appear on the spectrogram as darker constructions that huddle about a frequency range that changes depending on the sound being produced. Given the set structures that pronunciations tend to have, we came up with a rubric to evaluate intelligibility of pronunciations, as discussed in the next section.

In short, each of the aforementioned formants corresponds roughly to a physiological phenomenon occurring in the speech apparatus of the speaker. Thus, F0 appears as a baseline on the spectrogram and corresponds to voicing, i.e. whether (and to what degree) the vocal folds in the so-called voice box are active ("buzzing" or "humming") or passive (stationary). The quality of the sound produced is recorded in Formants 1-2 and still higher formant structures. F1 corresponds roughly to a cavity in or near the front of the mouth, which itself can be altered in shape and volume by moving the tongue up-and-down and/or in-and-out. Similarly, extension and rounding of the lips in contrast with spreading (i.e. "smiling") and withdrawal of the lips can effect the size of the volume being measured in F1, effectively altering the output on the spectrograph. F2 corresponds roughly to a cavity in or near the back of the mouth, i.e. top of the throat, which similarly can be altered in shape and volume by movement of the base and/or root of the tongue, as well as muscles of the throat. In short, the larger the volume in question, the lower the frequency produced; the smaller the volume, the higher the frequency produced. Vocalic sounds are produced by setting the vocal folds in motion and allowing the air flowing through them from the lungs to exit the mouth and/or nose without obstruction.

The position and shape of the tongue in combination with the position of the lips produces marked variations in the two upper formants that show as a line dancing upward or downward over time. Consonants, on the other hand, occur as an obstruction of the airflow. This obstruction can occur at any point along the line(s) of exit, and is caused thus by the tongue, soft palate, hard palate, teeth and/or lips moving into contact with one another. Full obstruction of the air passage at the point of the vocal folds can also occur. The obstruction at any position can constitute a full closure of the air flow or merely a partial closure to some degree or another. These obstructions of the air flow also produce perturbations on the spectrograph, generally appearing as blank columns in the case of full closure. Sudden movement of the vocalic line prior to the blank space indicates the point of articulation, the depth of shade representing the degree of obstruction, and the presence of dark striations along the base, i.e. F0 the voicing quality. Acoustic perturbations at frequencies higher than F2 are also possible, but mostly represent a "noisy" component of the sound in question, i.e. generally some form of frication (e.g. [s]), laterality (e.g. [l]), or nasality (e.g. [n]) that occurs in addition to the measurable variance in the aforementioned formants.

3.3 Methodology

Our grading rubric was based on the Likert scale used commonly in sociological studies to elicit a study participant's intuitions about the topic at hand. Consequently, the project rubric elicited a grade from the consultants as to the comprehensibility of the EFL-speakers' recorded pronunciation of the test words. Accordingly, a score from 1 to 5 was to be assigned to the recordings of the EFL-learner's pronunciation of each test word in the Pre-test and Post-test. The division of the scores 1 through 5 was made with the aid of a rubric based on three qualitative dimensions:

3.3.1 The degree to which the pronunciation varies from the phonemic standard and thus produces "interference" in interpretation of meaning.

The first dimension was the "Predictive Assumption". Since misproduction of a certain sound has the potential to change meaning or, even more drastically, to confound meaning altogether, this dimension accounted for disruption as "interference". Accordingly, a score of 5 was applied to a pronunciation that was "comprehensible with no interference". A score of 4 was applied when the test word pronunciation was "comprehensible with minor interference". A score of 3, when the test word was comprehensible but exhibits a "distinct foreign pronunciation". A score of 2, when the pronunciation was "incomprehensible unless greatly contextualized", i.e. only understandable to the native-listener when occurring in association with a present object or a printed form of the word spoken. A score of 1 was assigned to the pronunciation of the test word when it was "incomprehensible" to the native-listener under all circumstances.

3.3.2 Acousto-phonetic description of the utterance

The second dimension of the grading rubric offers a further "Description of the utterance". A score of 5 is assigned to non-native pronunciation that is "indistinguishable from a native-listener's pronunciation of the same word". A score of 4 was assigned to non-native pronunciation that was "easily understood by native speaker with some slight indication that the producer is a second-language speaker". A score of 3 was assigned to non-native pronunciation that is "understood by most native-listeners yet also clearly marked as foreign". A score of 2 was assigned to non-native pronunciation that was "recognized as a possible word by a significant part of native listeners, albeit even then likely confused for a word with a meaning not intended by the foreign speaker". A score of 1 was assigned to a non-native pronunciation that was "unrecognizable; most native-listeners would consider it gibberish".

	Avg. Pre-test	Avg. Post-test	Gain
EG1	2.1	2.8	0.7
EG2	2.4	3.6	1.2
EG3	2.0	3.3	1.3
EG4	1.85	2.5	0.65
EG5	1.9	2.4	0.5
EG6	2.6	2.9	0.3
EG7	2.7	3.0	0.3
EG8	2.4	2.6	0.2
EG9	2.7	2.6	-0.1
Avg	2.29	2.84	0.55

Table 3.1: Learning Gains (Exp. group)

3.3.3 Position on the spectrum "native - dialectal - non-native - non-speaker" relative to the native-listener's own speech

The last and third dimension of the grading rubric simplifies the native-listener reactions to basic qualifiers that describe the non-native pronunciation. It is thus titled "i.e., pronunciation is judged as". A score of 5 was "native". A score of 4 was "marked". A score of 3 was "foreign". A score of 2 was "foreign and mistaken for another word". A score of 1 was "foreign and meaningless".

A score of 5 thus reflects native pronunciation, a score of 1 reflects pronunciation of a non-native speaker with no knowledge of the target language, and the scores in between reflect pronunciations of a non-native speaker at various experiential stages relative to the extremes.

It should be noted that the effort is not to perfect pronunciation or teach particular accents. Rather we are trying to teach non-native English speakers how to pronounce new English words in a way that is intelligible to listeners. The theoretical scores in our games ranged from unintelligible to native pronunciation, but it is noticeable from our tables that average scores were in the range 2-3 on a 5-point scale. Roughly speaking we were trying to move speakers from 2 (intelligible but confusable with other words) to 3 (intelligible but heavily accented).

3.4 Evaluation Results

Based on the rubric mentioned above, the linguists (in consultation with each other) gave each utterance a numerical score from 1-5 on a Likert scale. The scores were then averaged across all the words that were taught through the two games. The measure that we used to process results was the average gain in Likert scale readings across the entire syllabus. We merged the scores from the two games, because there was no significant difference between

	Avg. Pre-test	Avg. Post-test	Gain
CG1	1.8	2.8	1.0
CG2	2.0	1.9	-0.1
CG3	2.0	2.0	0.0
CG4	2.6	2.7	0.1
CG5	3	2.2	-0.8
CG6	2.2	2.1	-0.1
CG7	2.1	2.2	0.1
CG8	2.4	2.35	-0.05
CG9	2.1	2.8	0.7
Avg	2.27	2.36	0.09

Table 3.2: Learning Gains (Control group)

the individual learning gains from each individual game. A Wilcoxon Mann-Whitney test between the learning gains from Zorro vs the learning gains from Guitar Hero resulted in a p-value of 0.56. Moreover, both the games followed very similar interface design process during creation and the aim was to study the effect and feasibility of such an intervention, as opposed to comparing the effects of different design choices against each other. Tables 3.1 and 3.2 present the average pre-test and post-test scores. It also presents the average learning gains for all the participants in both the control and the experiment group.

It turns out that the average learning gain on the Likert scale across all control group participants was 0.09, whereas for experiment group participants it was 0.55. It should also be noted that a Wilcoxon Mann-Whitney test between the average learning gains for the control and experiment group yielded a p-value of 0.02. This denotes that our results are significant. Moreover, a Mann-Whitney test between the average scores of participants during the pre-test for the control and the experiment group yielded a p-value of 0.96, which shows that there was no significant difference in the knowledge of participants when the study started, and that the experiment and control groups were comparable to each other in terms of their pronunciation baselines.

It's clear from this evaluation that the games we designed were able to generate short-term learning gains, not just from a computational standpoint [78], but also in terms of perceptually important criteria for intelligibility. There were also several merits to building these intelligibility criteria into the existing feedback algorithms that were employed. Some of these were:

- A thorough linguistic evaluation of intelligibility is costly and time consuming. An automatized evaluation can be much faster and cheaper.
- If this feedback on intelligibility could be made real-time (by building it into interfaces/games), it would increase the learning gains that participants can derive out of such systems.

Therefore, we tried to take first few steps towards including these perceptually important criteria for intelligibility into our pronunciation evaluation system. The rest of the chapter talks about this process and some preliminary results from the same.

3.5 Computational adaptations

3.5.1 Challenges

Whether judged by a computer program designed to measure changes in the acoustic wave form or by a human (either with our without linguistic training), the speech produced by any human can be deciphered for fingerprint-like, tell-tale clues that can in turn be associated with qualitative information about the individual speaker. That is, people from different backgrounds pronounce things differently. The association between a particular pronunciation of a certain word entails the listener being able to recognize the more-or-less homogeneous pronunciation of his own speech and those among whom s/he commonly operates. As soon as another person speaks a few words, a native-listener can typically determine whether the speaker "belongs" to the listener's native speech community. That is, a few unexpected variations in the acoustic waveform is all it takes for the native-listener to know that the speaker "isn't from around here".

Training a computer program have the same kind of sensitivity to acoustic data is a matter that goes beyond mathematical and/or programming knowledge. That is, it is often not a simple question of how much variation in a particular phonetic segment is to be allowed before a computer would be able to raise a 'red flag' to denote non-native pronunciation. All languages consist of a finite number of phonetic sounds that, if changed, have the capability of conveying a new meaning on the one hand, and causing an utterance to become completely meaningless on the other. That is, variation in the sound wave may or may not be meaningful, depending 1) on the language involved, and 2) on the "sound" involved. This "sound" has much less to do with acoustic reality than with the native-listener's mental perception of the sound s/he encounters and produces. Depending on the language involved and the finite sounds, i.e. phonemes, that have the potential to cause a change in meaning in the listener(s) present to interpret the speech, the variation involved in the acoustic product, i.e. "phone" or "sound", might vary drastically or very little.

This qualitative differentiation in the pronunciation of a phoneme is complicated even more by the fact that "correct" pronunciation, i.e. unmarked pronunciation (that which does not raise suspicions), differs from one native speech community to another. Thus, while both considered "native" speakers/listeners of English, the native of southwestern England and that of the southwestern United States will have characteristic pronunciations that each considers "correct", and therewith will consider different types and degrees of variations of individual and particular phonemes to cause or not cause meaningful changes in meaning or recognizability of a word. This issue compounds even more so when a non-native English speaker acquires the ability to speak English, but does so with the phonetic habits of his own native language,

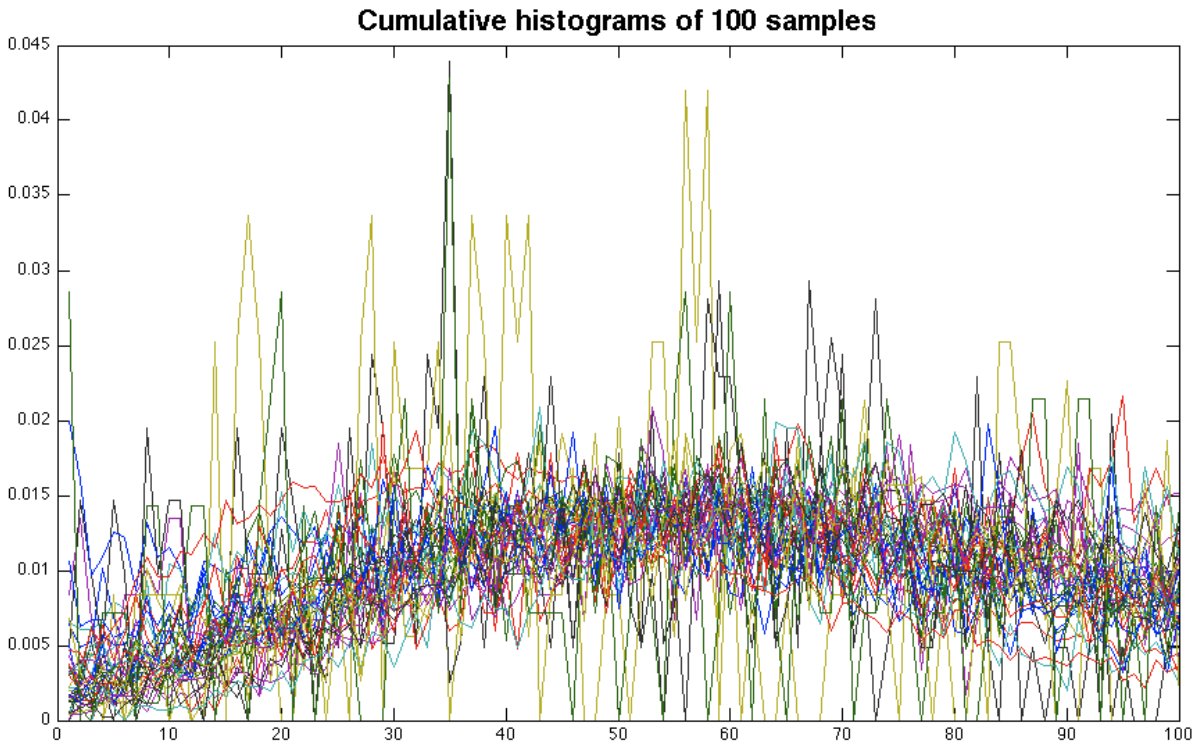

Figure 3.1: Cumulative histograms for 100 samples of the 43 phones

e.g. Spanish or Hindi. In either case, the EFL (English as a foreign language) speaker will produce phones that vary to some degree from 1) that of the native speaker of English (regardless of whether it is England or the US), and 2) that of another EFL-speaker with a different native language. Thus, we see that what is commonly referred to by the layman as "accent" is actually the congregation of a number of variables that affect what a native-listener would easily recognize as being out of place. These include, but are not limited to: geography, native language of the "foreigner", gender, age, time that the "foreigner" has spent among the native-listener's cohort, etc. Only a sophisticated computational process could account for all that a native-listener can.

Our ultimate goal is to incorporate the outcome of these expert consultations in better tuning our pronunciation evaluation system in its ability to provide meaningful feedback to the user (an EFL-learner) in order to induce changes in that user's physiological production of sounds until that which is produced matches acoustically that which the computer has been trained to judge as a "native" pronunciation, in terms of intelligibility. The next few subsections talk about the first few steps towards building such linguistic expertise into real-time pronunciation evaluation/feedback systems.

3.5.2 Requirements

In order to further refine the feedback mechanism that the games offered and also to automatize such evaluations in the future, we analyzed the rubric defined during the linguistic analysis in search for annotations and techniques that we could build into the pronunciation evaluation components of our system. It should be noted that the real-time pronunciation evaluation in our games was restricted to acoustic scores produced by the speech recognizer. The scores that the recognizer produces are just log-likelihood probabilities and demonstrate the overall quality of the speech encountered by the recognizer. Intelligibility, on the other hand is a slightly different problem and is determined by the three qualitative dimensions (also mentioned above): a) the degree to which the pronunciation varies from the phonemic standard and thus produces "interference" in interpretation of meaning, b) acousto-phonetic description of the utterance, and c) position on the spectrum "native - dialectal - non-native - non-speaker" relative to the native-listener's own speech.

From a computational standpoint, this means that to evaluate the intelligibility of a particular speech segment is necessary to do a phoneme level analysis. Moreover, the analysis should compare the acoustic characteristics of the speech units (phonemes) being evaluated, with a native-speaker's enunciation of the same.

3.5.3 Implementation

Given the requirements above, the feedback mechanism needed to do the following:

1. Break down an incoming speech segment into its component phonemes and generate phoneme level acoustic scores.

2. For each phoneme level score in a word, statistically compare it to a pool of acoustic scores of the same phoneme, uttered by a native speaker.

For example, if the incoming word is "peak" (phonetic representation: "P IY K"), it would be broken up into three acoustic scores. The intelligibility of "P" will be determined by comparing its acoustic score to a pool of acoustic scores. This pool will be obtained by repeated usage of the phoneme "P" by native speakers over a large number of cases. A similar process will be adopted for all the phonemes in a word, and the result averaged out to obtain overall intelligibility.

The database that we used for comparison was the ICSI meeting corpus [28]. The corpus contained audio recorded simultaneously from head-worn and table-top microphones, word-level transcripts of meetings, and various metadata on participants, meetings, and hardware. The corpus contained approximately 72 hours of recording from 53 unique speakers. The demographics of the speakers suited our requirements, as it was dominated by native English speakers, but also had speakers from Spanish backgrounds. More intricate details of the corpus are covered in the ICSI meeting corpus paper [28]. The entire corpus was forced-aligned to the transcripts using the CMU Sphinx speech recognition system. Overall, we

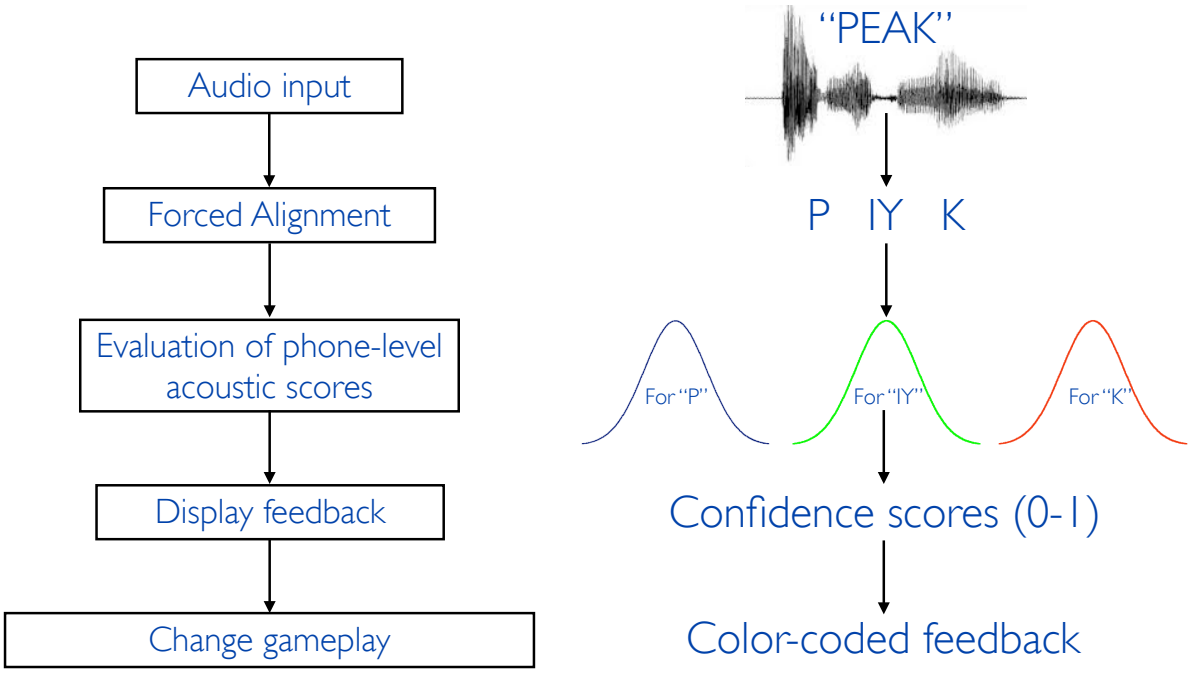

Figure 3.2: Explanation of the feedback mechanism

generated 320,938 acoustic scores across 43 phonemes (used by CMU Sphinx), an average of 7,463 observations for any particular phoneme.

To make sure that this data follows some high level patterns, we did a bunch of statistical measures and test. In interest of time and space we discuss the most important ones. Since the scores generated were log-likelihood probabilities, we converted them to original values, by raising them to an exponent. We generated cumulative histograms for 100 samples of each of the 43 phonemes and plotted them together to make sure that these values follow a high-level probability distribution. As is clear from Figure 3.1, apart from a few outliers, most of the scores followed a reasonably similar and dense probability distribution as suggested by the estimate (histogram).

As the next step, for each phoneme we used a beta distribution to fit the data and calculated the characterizing parameters, α and β. The reason for choosing a beta distribution was that it offers reasonable freedom in terms of the types of data it can fit. Once this was done we had 43 pairs of α and β values, that could be used to recreate the beta distribution that corresponds to any particular phoneme. Therefore in order to determine the quality of an input speech segment, we could now just estimate the position of the acoustic scores for component phonemes on the respective beta distributions. Like in the examples mentioned above, we could now estimate as to where P, IY and K would lie on their respective population distributions, and what percentage of the data is superior and inferior as estimated by the beta

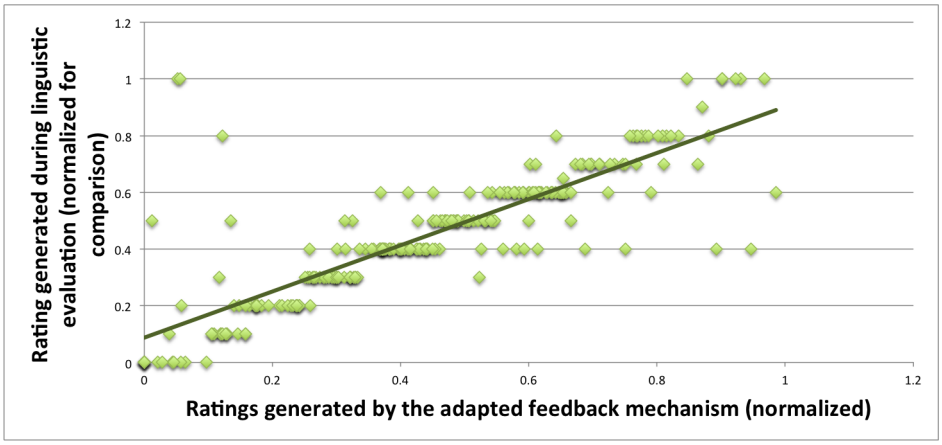

Figure 3.3: Scatter plot of ratings generated through linguistic evaluation vs ratings generated by the adapted feedback mechanism

curve. The overall quality score would of course be an average across the three phonemes. We call this the **intelligibility score**. It should be noted that this enabled us to give feedback based on the intelligibility of a speech segment on individual phoneme level, as well as word level. A diagrammatic representation of the functional system is presented in Figure 3.2

3.5.4 Evaluation

For an initial evaluation of the adaptations, we chose to use the hand-graded data (from the linguist) that we already had. The aim was to test if our adaptations could produce ratings similar to the ratings obtained during linguistic evaluation. It should be noted that since the rubrics used during linguistic evaluation were established *a priori*, the data being evaluated did not bias it in any way. Moreover, the data used to adapt the feedback mechanism came from the ICSI database, which was totally independent of the data collected from our participants. Therefore, it was reasonable to assume that the biases in evaluation setup were minimal.

The total number of words evaluated during the linguistic evaluation, across pre-test and post-test were 260. It should be noted that we did not distinguish between pre-test and post-test data, as we were just trying to test the ability of the adapted feedback mechanism to imitate a linguistic evaluation. Therefore, we ran the adapted feedback mechanism to produce intelligibility scores across all of these 260 words. Since the granularity of the linguistic evaluation was at the word-level, we excluded the phone level intelligibility score from this comparison. Once we used the adapted feedback mechanism to evaluate the data obtained from the participants, we normalized it for comparison. We also normalized the Likert scale ratings obtained during linguistic evaluation. The correlation between these two sets of ratings was 0.80 across 260 data points, a strong correlation. For a visual feel of the

two data sets, Figure 4.1 clearly depicts a strong correlation between the two sets of ratings through a scatter plot. It should be noted that the data points form tight clusters on one axis because one set of ratings were on Likert scale to begin with.

3.6 Realtime Intelligibility Feedback

All the code discussed above was written in Java, and has plugs for connecting to the CMU Sphinx speech recognition system. It can directly take the word-level and phone-level acoustic scores generated by the speech aligner and generate scores (between 1 and 100) for each phoneme, that represent the degree of intelligibility. This means that any speech-enabled game/system that uses CMU Sphinx or any other recognition systems for generating acoustic scores, can use our Java classes to generate intelligibility scores. This work has inspired other projects like project SMART [41].

3.7 Future Directions

3.7.1 Evaluation of Realtime Intelligibility Feedback

In this chapter we have tried to investigate computational optimizations into the real-time feedback given by the Computer Aided Pronunciation Training (CAPT) games/systems. In simpler words, instead of just reporting feedback based on recognizer scores, the games could produce feedback on intelligibility of the input speech segments. An interesting future direction could be including this real-time feedback into the games, and a longitudinal evaluation of the system with a larger audience. It should also be noted that in our initial evaluation of the optimized feedback mechanism, we just used the word-level scores, because the linguistic ratings were also at that level. Phone-level intelligibility scores can be valuable in pointing out exactly what is "non-native" about a particular piece of pronunciation. Even though the computational adaptations that we have made produce phone-level scores, we haven't evaluated them real-time in a game. Future research could possibly explore this as well.

3.7.2 Prosody Feedback

A limitation of the current work presented this far, is that the focus has been on word-level feedback. Literacy theory research suggests that prosody can play an important role in the interpretation and sense-making of spoken language [76]. Future work that builds on the research presented here, could potentially study sentence-level prosody in combination with word-level intelligibility to help children improve their spoken language skills.

3.7.3 Mobile Devices

Another future work would be to look into the domain of mobile devices. With the increase in the processing power of the phones, it is possible to run speech recognizers on cellphones. Getting the real-time intelligibility feedback to work on mobile devices would also lead to greater coverage in terms of user populations, and this could result in interesting patterns in data.

3.8 Conclusion

In this chapter we presented the design and linguistic evaluation of two speech-enabled games at a high school in California, meant to target Hispanic children. There is a growing need for such interventions due to a variety of factors. We systematically established a criteria for evaluating perceptually relevant characteristics in speech that can be used to judge intelligibility. Using these criteria we evaluated the pre-test and post-test data from our participants and proved that our games are able to generate short-term learning gains, even from an intelligibility standpoint. We also used the criteria developed during linguistic evaluation, to motivate and inspire some computational optimizations to the existing feedback mechanism used by our games. To evaluate these optimizations, we used the pre-test and post-test data and tested the ability of our optimizations in mimicking the ratings generated during the linguistic evaluation of intelligibility. We saw a strong correlation in the ratings generated by the two methods.

Even though this line of work has tremendous potential, and the results that we obtained were very encouraging, there is a large body of research that indicates that the greatest impact on literacy will come from interventions at an early stage of development. Therefore, as the work from project SPRING continued to get adopted and evolved [41], we also started exploring technologies for early literacy. The rest of the thesis will be dedicated to discussions on the same.

Part II

Question Answering Technology for Preschoolers

Chapter 4

Theory and Motivation: Child question-answering

4.1 Introduction

A[1] large body of research has shown that the "literacy gap" between children is well-established before formal schooling begins, that it is enormous, and that it predicts academic performance throughout primary, middle and secondary school. Indeed rather than closing this gap, there is much evidence that formal schooling exacerbates it: once behind in reading and vocabulary, children read with lower comprehension, learn more slowly and have lower motivation than their more language-able peers. Many national organizations recognize the essential role of early literacy in a child's later educational and life opportunities [69],[12],[24]. Hart and Risley [23] report a factor of two difference in the working vocabularies of high vs. low-SES three-year-olds. The average low-SES child has heard 30 million fewer words than a high-SES child by this age. However, they also observed that SES alone is not a predictor of cognitive development at the pre-school stage. "The richness of nouns, modifiers, and past-tense verbs in their parents' utterances, their parents' high propensity to ask yes/no questions, especially auxiliary-fronted yes/no questions; and their parents' low propensity to initiate and use imperatives and prohibitions were more strongly predictive of the children's performance on the Stanford-Binet IQ test battery than was the family SES." Hart and Risley note that to close this gap is an enormous challenge and will require lengthy and regular language experiences for the child.

As noted in the above studies, the greatest impact on child literacy will come from intervention at pre-school ages. The pervasive achievement gap begins at birth. Positive adult-child interactions provide the necessary conditions for language acquisition, expressive and receptive language skills, interpersonal coping skills and eventual literacy that includes reading and writing. Speaking and listening skills, usually learned, refined and reinforced through interactions with the environment, lag behind in children living in environments that

[1] This work was done in collaboration with Ingrid Liu and Carrie Cai.

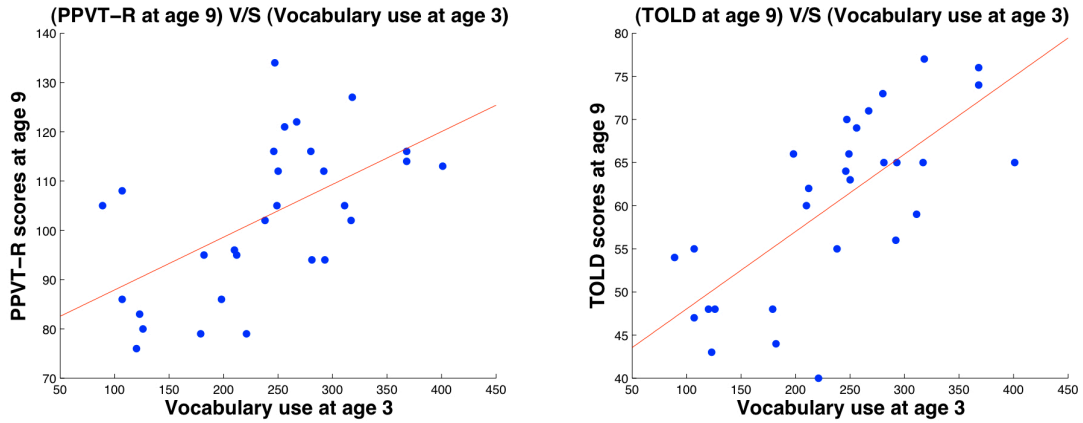

Figure 4.1: Correlation of vocabulary use at age 3 to vocabulary growth till age 9

are under-resourced in terms of language features.

While it is becoming increasingly clear that conversations and language interactions serve as an important tool in the child's cognitive process, a growing body of research is also suggesting that pre-school children are voracious inquisitors. One recent study found that preschoolers ask approximately 80 questions/hour [10] which constitutes more than one-fourth of their utterances. These questions are an essential part of language development: they provide primary experience with question construction, statement construction, explanation construction, complex tenses etc. The child question-asker is primed for an answer. Unlike other forms of interaction (reading, games) no external influence is needed to garner the child's interest or build motivation. The questions reflect the child's current state of knowledge and should take them just beyond it, i.e. child-initiated questions are naturally in the child's Zone of Proximal Development (ZPD). At the preschool level, most questions are fact-based, e.g. "do fish fly?" although around age 3 there is a sharp increase in explanation-oriented questioning. Fact-based questions are readily answered by short statement responses. Children may ask follow-up questions, but in general the chain of conversation is short. Explanation-oriented questions seek richer answers with causal links or chains [17]. They will often be met with additional requests for more information, or a universal "why?" Children seem to have strong preferences for the form of the answer (a causal explanation vs. fact-based answer), although less so for content. The "why?" question in particular is often indicative of a general desire for more information, and can be meet with any number of responses with relevant material (i.e. explanations in terms of human or non-human agency, history, rules or principles, or stories). Question-asking, not surprisingly, goes beyond literacy and is an integral part of children's cognitive development [10].

It is safe to assume that parents are the primary teachers for preschool children, but many interventions directed at parents reproduce the gap. Educational interventions for children involving parents appear to be dependent on the parent's educational level, and so literacy differences persist across generations. For instance, dialogic reading (defined

later) interventions involving high-SES parents were far more effective than with low-SES parents (effect size of 0.58 vs. 0.13) [54]. Children evidently need some form of linguistic engagement for many hours a week, with a language-able partner who can engage with them in age-appropriate language-learning activities. Since research in early child development suggests that for pre-school children question-answering serves as a frequent and heavily-utilized medium of synchronizing mental models with adult-like understanding of the world, this linguistic engagement can come in form of interactive question-answering systems. Since children spend a significant amount of time playing alone, or out of home, there might be instances when they don't find an adult around to answer their questions. There might also be times, when the adult doesn't have sufficient information at hand to answer a child's question. This explains the need for expert interactive systems that can work as engaging question-answering agents. However, before any type of technology push, we want to establish a theoretical framework in which such interventions can be based. This framework would come from previous work that has analyzed question-answering and also some analysis of existing databases of children's speech. It should be noted that the findings in this chapter do not just hold for question-answering systems, but for any conversational/pedagogical system that intends to target preschoolers through conversational exchanges.

4.2 Related works and Motivation

4.2.1 Language Learning

Child development research has shown that children rapidly acquire knowledge of new words starting at 18 months of age. According to Jean Piaget's theory of development, it is during the preoperational period (ages 2-7) during which children become able to represent ideas through language and mental imagery [61]. Vocabulary size more than doubles between 18-21 months and again between 21-24 months of age, and a typical child understands at least 10,000 words by first grade. These patterns suggest a high propensity for children to acquire vocabulary at a very young age, and that preschool age is likely an appropriate time to engage children in language learning ([73]).

Moreover, scaffolded linguistic interactions with adults significantly advance children's learning. For example, toddlers whose mothers follow their attention by labeling objects of joint attention tend to have larger vocabularies later on ([73]). While children appear to be naturally susceptible to acquiring linguistic competence in their early years, their sentences are initially error-prone and only become more well-formed over time, partly through exposure to others. Not only does adult grammar provide semantic clues that aid children in deciphering the meaning of words, social cues also help children develop competency through the corrective feedback that adults give when children use words incorrectly.

According to psychologist Lev Vygotsky, such interactions are not merely external forces that provoke internal change in an individual, but rather integral to the very mechanism of cognitive development [79]. Because childhood word learning both increases rapidly at an

Question type	Examples	Percentage (CHILDES)
Information seeking questions		71%
Fact	Where's the ball?	56%
Explanatory	Why is the sky blue?	15%
Non-information seeking questions		29%
Attention	Hey mom?	6%
Clarification	What did you say?	9%
Action	Will you close the door?	3%
Permission	Can I have an apple?	5%
Play	To doll: Are you hungry?	1%

Figure 4.2: Structure of children's questions

early age and demands support from adult modeling, it is valuable to examine ways in which adult-child interactions at the preschool age can be modeled through software interfaces.

4.2.2 Developing knowledge of the world

It is common knowledge that young children ask a considerable number of questions, but to correlate children's inherent motivation to develop theories about the world with their question asking, the amount, content, and responses to adult's answers have been analyzed. In a longitudinal study of transcripts involving four children, ages 2.5-4, overall 71% of the questions were information-seeking questions, and overall 56% were fact-seeking and 15% were explanation seeking questions [7]. Non information-seeking questions ranged from seeking attention, clarification, action, permission, play, towards a child or animal, or were unknown [7]. To view the statistics of questions-asking across a greater socioeconomic spectrum of families, a diary study of 68 children's questions are recorded for one week, between the ages 1;0 - 5;0[2] . Before the experiment, the parents are furthermore trained to recognize specific gestures indicating a child's desire for an explanation. From as young as 1;0 : 1;5, the 69 information seeking questions per child on average during the week were recorded by the parents, indicating that children begin building their theories of the world through questions before they can even speak coherently. For the ages 3;0 : 5;0, which is the age range focused on in this chapter, the 30 children observed had 54 - 70 questions recorded per child during the week by their parents, where 80 - 90% of the questions were information-seeking questions (Figure 2). The number of questions recorded by the first study appear much decreased from the second because it is difficult for the parents to fill a form for every instance of the children's questions, whereas it is much simpler and accurate to record conversations and transcribe them later.

[2]Ages are represented with year;month.day, where day is optional. For example, 1;5.10 is a child that is 1 year, 5 months, and 10 days old

4.2.3 Developing concept of causality

Based on questions with young children, such as asking the children for sentence completions, Piaget concluded that young children had very primitive notions of causality under 5 or 6 years old [61]. However, recent works are re-examining Piaget's claims. Shultz performed an experiment, where children of ages 3, 5, 7, and 9, were shown three pairs of two objects, where one object was the cause of an effect, and asked to identify the object which created the effect. Children of all ages were able to correctly link the causes and effects using attributes of the source or result.

Hood and Bloom find that children make causal statements and responses to causal questions by adults from at least age 24 months, and by 30 months, they can ask causal questions productively. Furthermore, these causal questions are oftentimes more sophisticated than one word questions such as "why" and "how" that are meaningful in the context of specific domains such as natural phenomena, biological phenomena, physical mechanisms, motivation/behavior, and cultural conventions. In a study by Callanan and Oakes [7], parents of children ages 3, 4, and 5 were asked to record forms for children's questions, with special focus on causal questions for two weeks. At age 3, 20% of "why" questions were simply "why?" at age 4, 0% were "why?" and at age 5, 4% were "why?".

Causal questions are actively asked by children to acquire explanations of causal processes in the world, and indeed, when children ask a causal question but are given a non-explanatory response, the child will express dissatisfaction. In an analysis of causal questions involving "how" and "why" from various datasets in the CHILDES database, children are shown to re-ask the original question or provide their own explanation following a non-explanatory response with high significance. In contrast, children are likely to agree or say oh, ask a follow-up question, disagree, or provide no response with high significance following an explanatory response. To demonstrate this consistently across a controlled environment, Frazier et al. [17] performed a laboratory experiment where investigators engaged children in conversation about a set of unusual toys and alternated between providing explanatory versus non-explanatory answers to the children's questions. The children's responses to adult answers were then analyzed and coded. Again, children significantly agreed or asked follow-up questions following explanatory answers and re-asked their question or provided their own explanation following non-explanatory responses. The difference across explanatory responses and non-explanatory responses with respect to children giving no response was not significant.

In their questions of causality, children develop their knowledge of the world. Conversely, by developing children's knowledge of objects in the world, children also build the mental structures necessary to reason causally. Shultz's experiments provide evidence that children can judge causality by using their knowledge of object attributes, or by generative transmission, rather than on attributes such as spatial or temporal contiguity [72]. Carefully selected apparatuses and effects, such as a lamp and a spot of light on the wall, removed the possibility of spatial contiguity as a causal reason. To check whether children made causal assumptions based on temporal contiguity, the children were asked to explain the reason for the effect,

and the analyzed results showed that of the 84% correct attributions, 81% were explained by the children based on the object's or transmission's properties.

4.2.4 Categorization of children's question

Many recent research papers have focused on categories of children's questions through manual coding. These studies use a subset of the datasets described in the appendix, perform diary studies, or perform laboratory experiments observing the question/answering dynamic for young children.

Questions can be coded along several dimensions: information-seeking versus non information-seeking, response desired, content, response type, and information given in the response [10]. The response desired can be a fact or explanation if the question is information-seeking, or it can be attention, clarification, etc. if the question is non information seeking [10]. Content can range from the label, appearance, property, etc. of the questions' subjects [10]. Within Chouinard's monograph of children's questions, longitudinal studies from CHILDES have been manually coded and the statistics calculated for each of these dimensions. Chouinard's monograph also provides a diary study of children ages 1;0 - 5;01, where questions recorded by 68 families were coded by content and question type, and all questions were grouped within age buckets. In the causal domain, studies have also coded children's questions along several dimensions. Analyzed multiple aspects of children's causal questions. , the responses that are given by adults, and the children's responses to the adults' responses. Causal question can be grouped into the categories "how?", "why?", "what would happen if?", "what is this for?", "where did this come from/where is this going?", "do you know why/how/what?", other questions, and non questions. Along these causal question types, Callanan and Oakes derived the statistics of the ratio of causal question types, the situations in which they emerge, and their content through a diary study of 30 preschool children [7]. Frazier et al. derived the statistics of parent's responses to children's causal questions and children's responses to different responses by their parents by examining longitudinal studies from CHILDES and through laboratory experiments with 42 preschool children [17].

4.2.5 The structure of parent's responses

In terms of responses, children seem to have strong preferences for the form of the answer (a causal explanation vs. mechanism based answer). As per the literature, a causal explanation is preferred for explanation seeking questions like "why", and mechanism based answers are preferred for fact-based question like "how".

The "why?" question in particular is often indicative of a general desire for more information, and can be meet with any number of responses with relevant material (i.e. explanations in terms of human or non-human agency, history, rules or principles, or stories).

Children overwhelmingly receive informative answers to their questions. They are significantly more likely to get an answer to their questions than not, and they get the information they ask for as much as 86% of the time [7],[10]. This tells us that parents interpret these

Response type	Examples
Mechanism	There's a membrane in it that you blow.
Prior Cause	Jimmy is crying because Susie pushed him.
Consequence	They work on the road, to make it wider.
Combined Cause-Consequence	That's the way his outfit is, it keeps him warm.
Noncausal	Going to a meeting
Questions	Why do you think?
Yes/No	Yes/No

Figure 4.3: Structure of parent's response

questions as serious requests for information, and give the target information to the children. So, from very early on, this mechanism is effectively eliciting information from others, gathering information about the world. On top of this, parents give additional information that supplements the information children request; as much as 24% of their responses contain such information [7],[10]. This is particularly true when children are at the youngest ages; this information may be complementing the younger children's more limited ability to identify the specific information needed to fully understand the situation at hand. For example, when a child asks "Whats mom doing?", if the father were to say "Shes shivering" this would answer the child's question, but not get to the heart of the matter; by saying "Shes shivering, shes cold" the father tells the child that the activity really is not what is key here, it is his mothers internal state that is important. If the child does not know what a poodle is and asks "Whats that?", if the parent answers "its a poodle" this would answer the question; but the answer "Its a poodle. Poodles are a kind of dog" gives the child a bit of extra help in setting up the proper animal hierarchies. So, children's questions open the door for targeted input from the parents that helps guide their child's thinking in the right direction at precisely the moment they need and want it. [7],[10]

4.2.6 The content and form of children's questions

The codes "label", "property" , "appearance", "function", "part", "generalization" and "hierarchy" look more closely at what information children seek out during the categorization process, while learning what an object is and which new objects should be treated as part of a category [7], [17].

The codes "Theory of Mind (ToM)", "activity" and "possession" are used specifically to investigate learning about people.

Importantly, however, all of these also apply to animals, and some can apply to objects, and an important part of the learning process is discovering which things apply to people only, or to people and animals but not objects and so on [17].

Content type	Examples
Label	What's a jack-o-lantern?
Appearance	What color is an apple?
Property	What is Mickey made of?
Function	What does this do?
Part	What is the lion doing?
Activity	Is the car broken?
State	How many stars are there?
Count	Do you have a cat at home?
Possession	Whose coffee is that?
Location	Where is the ball?
Hierarchy	What kind of car is that?
Generalization	Why do cats like milk?
Theory of Mind	How does the pilot know to fly?

Figure 4.4: Content of children's questions

4.3 Children's questions in various activities

There are several components, requiring research from various fields, necessary to construct any technology that promotes question asking by pre-school aged children. The conversation dynamics between children and adults have their own structures and processes, with complex rules of turn-taking. In this domain, we are primarily interested in how to best encourage a child to ask meaningful, instructional questions while keeping them engaged and happy. To anticipate and correctly answer the questions that children may ask, it is necessary to properly identify and group the questions with the type of response needed. Determining the types and levels of engagement children have during specific activities in their daily lives will guide us in designing technology that promotes their question asking.

4.3.1 Materials

The focus of our question categorization is to investigate how engagement differs with interaction. For our analysis, we had to choose between labeling dialogs of spontaneous child play or dialogs of children with controlled play activities in a laboratory setting. For spontaneous child play, the dialogs would have to be coded for the activity type, and there would be variation within groups of interaction types, such as the type of toys a child had access to in a game of pretend. Furthermore, the transcripts of child and parent interaction lack any details regarding the surrounding objects, simultaneous events, and other extraneous circumstances, making them difficult to code for interaction type and difficult to annotate for disturbances. For controlled play activity, there are always pitfalls related to the naturalness of interaction in an unfamiliar setting with new objects. The observer's paradox is an additional concern, which affects both child and parent, since cameras and investigators easily distract the chil-

dren, while parents are concerned with their appearance as guardians [42]. As with any data analysis, there must be a large enough set of data, which is varied among different children.

After preliminary of analysis of both types of datasets within the CHILDES database, the spontaneous child play was determined to be very difficult to annotate in a consistent manner, and a laboratory study of adult with child interactions was chosen. Appendix details the dataset chosen: Gleason, and we use the age range 3-4. Since there are laboratory studies of the child with the Mother and Father separately and the studies are spaced out, we only analyze the transcripts of children who are between the ages 3-4 for both visits. This left 6 children, ages 3;1.04, 3;7.01, 3;2.21, 3;2.12, 3;2.03, and 3;7 during the Father's visit and ages 3;0.20, 3;6.07, 3;2.02, 3;3.16, 3;2.21, and 3;7.25 during the Mother's visit.

4.3.2 Procedure

Since we are interested in building an interactive interface for addressing children's questions, we code the questions in the Gleason study across various dimensions of question types. The first dimension chosen was **questions of causality**. The causal categories were chosen from the Callanan and Oakes [7] study as a comprehensive overview of children's causal questions, and no other causal question types were found during coding. The second category was the **response type expected**. If a child's question is in a causal category, then the question requires an explanation. If a child's question does not fit in a causal category, but is still information seeking, then the response needed is a fact. This includes clarification of a previous statement, confirmation that a belief or answer is correct, or any other question that seeks information. If the child asks a question for attention, to direct the conversation to a different topic, to direct attention of the adult to an object, to request something, or to signify interest or impatience, then the question is non-information-seeking. Of the information seeking questions (fact-seeking and explanation-seeking), the question can be directed completely towards the activity at hand and provide the child with no new information of the world. These questions are labeled "within scope". Questions that are "outside scope" can still be about the current toys the child is interacting with; however, it should add to the child's knowledge base of object names, properties, or mechanisms in the world. Lastly, the adult prompts many of the questions that children ask. To engage a child, adults will often ask the child a question. When the child repeats the question, the question is not the result of the child's inherent interest, but of the adult's mode of interaction, and is thus coded.

4.3.3 Discussion

The Gleason dataset was relatively simple to divide into the three activities, since the parents were encouraged to split time between the activities evenly across their half hour in the lab. There was a section at the end where the investigators holding the study gave the children a gift-this section was not included in the category analyses. There were also instances where the children noticed a camera in the room and conversed about the camera-this section was also removed in category analyses. Lastly, there was overlap between activities. When fully

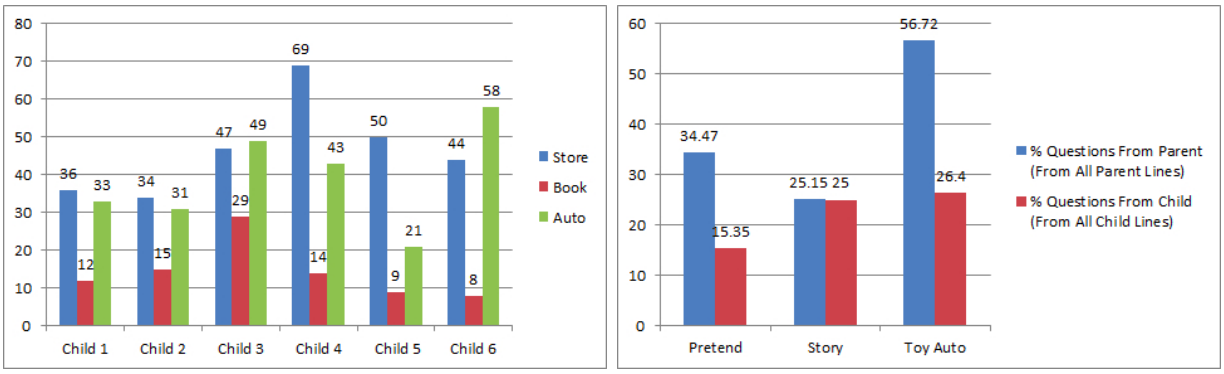

Figure 4.5: **Left:** Ratio of conversational turns by parents and children percentages in various activities. **Right:** Percentage of questions by parents and children in various activities.

engaged in an activity, the child would stay focused on the task at hand; however, between activities, it would take several turns of persuasion by the parent to continue with the next task. A new task is considered started when the parent or child suggests the activity, and there is no further debate after that line of starting the new activity.

Types of roles

The type of role children take based on activity can be inferred by the ratio of child statements to questions and the number of child turns to adult, as they vary greatly between activities. Figure 2 presents these ratios. For example, in the game of pretend, children took up relatively more turns in the conversational exchange, but asked fewer questions relative to their increased speech. From the transcripts, it is clear that children are more interested in the role-playing aspect of pretend, than asking questions about familiar objects. Thus, they direct the conversation towards the make-believe world they wish to enact, while asking questions only when they are uncertain what a toy prop is. In contrast, children on average took a more passive role in interpreting the picture book. In general, the parents made up the story for the child, and most children took the role of listener, with varying degrees of participation in story-making. When constructing the toy auto, parents tended to take a more verbally active, tutorial role, answering questions, giving suggestions-in both statements and questions-and giving and asking for additional information about the different components of the car. This is reflected in the relatively high ratio of questions to statements by parents. These numbers hold across all children, and Figure 3 presents the percentage of child-initiated questions asked per child per activity, out of all questions asked in the activity by the child and adult. From this figure, it is more apparent that in the story activity, parents ask many more questions than children.

Overall, children were more active when playing store or playing with the auto. When reading the story with their parents, however, the amount of child-initiated questions varied greatly. It should be noted that this cannot be seen as a result of just the child-there are

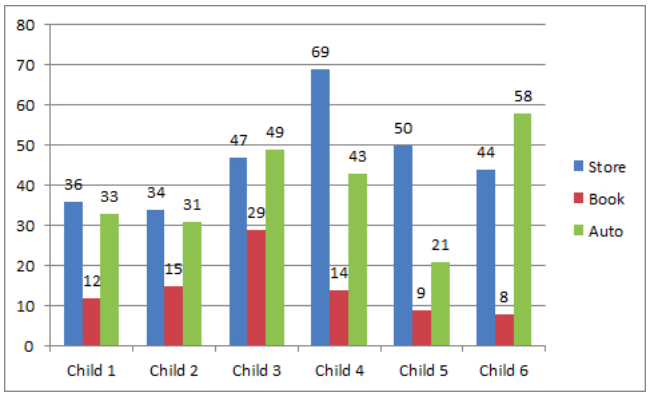

Figure 4.6: Percentage of child initiated questions per child per activity, out of all questions asked by child and adult in the activity.

	How?	Why?	What is this for?
Store	1	4	5
Book	0	15	0
Auto	20	13	6

Table 4.1: Causal questions across activities

	How?	Why?	What is this for?
Store	0	4	5
Book	0	9	0
Auto	18	13	4

Table 4.2: Number of causal questions across activities initiated by children

two sides to the conversation, and the variance in child-initiated questions may also be the specific dynamics between the parent and child. The percentage of child-initiated questions is presented in Figure 3 and 4.

Causal questions

Based on numbers of causal questions, we can postulate what logical reasoning is required by a child for different activities. The numbers of causal questions are recorded in Table 4.1. In these laboratory settings, very few causal questions were asked. "Why" is most prevalent in the fictional story reading activity. The story was based on a cat being chased around by various characters in funny scenarios, and the child's why questions were directed towards making sense of the parents' explanation of scenes in the story. There were many more opportunities for "why" questions because the events occurring on the page could be interpreted in many

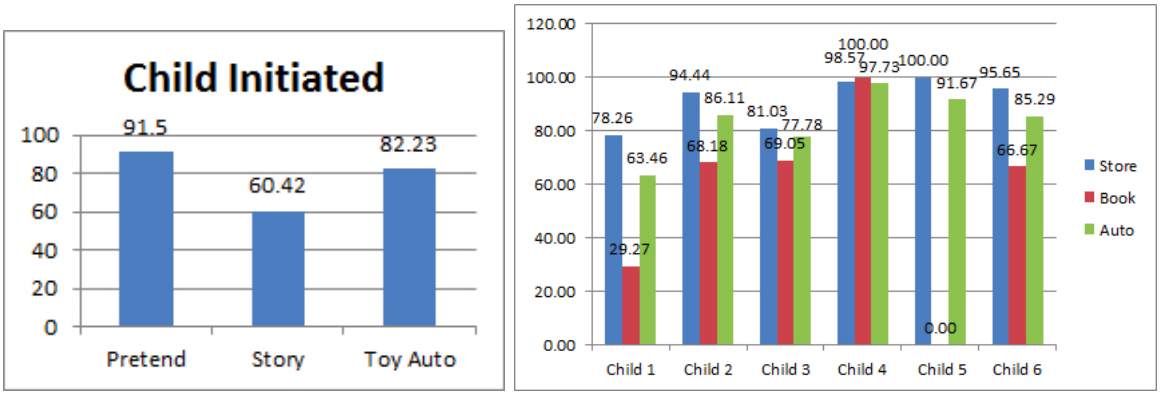

Figure 4.7: **Left:** Percentage of questions initiated by children across activities out of all questions asked by children in an activity. **Right:** Percentage of questions initiated by each child across activities out of all questions asked by a child in an activity.

ways-for example, a child asked why a picture of a girl on the page was crying. "How" and "what is this for" is more prevalent when the child is piecing together the toy auto, since every request for suggestions on building the car is a "how" question, while "what is this for" questions result when the child is trying to determine what a novel piece of the car does. The increased questioning during the auto construction is the result of an activity where a child feels the need for a lot of direction, which is reflected in their increased "how" questions. Since there is very little data to infer the number of causal questions children naturally ask, we also provide a Table 4.2 which shows the number of causal questions asked per activity initiated by the child, to rule out the possibility that the children were not simply repeating causal questions asked by an adult. The Callanan and Oakes study [7] showed a veritable amount and range of causal questions asked by children, so it is possible that either none of the activities analyzed prompted children to ask causal questions or the environment prevented the type of exchange conducive to children asking causal questions.

Scope of questions

The most difficult dimension to code across was "outside scope" versus "inside scope". At what point is a question considered relevant to adding to the child's knowledge of the world? Even a question that asks for assistance could add to a child's knowledge of the world, since it informs the child of what a specific adult can and is willing to do for a child. In Chouinard's monograph of question in the CHILDES dataset, a more conservative definition of a knowledge structure is adopted. A knowledge structure can be "facts about a given category/concept/domain" or "explanatory information that organizes those facts within the category/concept/domain" [10]. For our study, a question is considered "outside scope" if a response would either help categorize an object, describe properties of an object, or explain how an event, such as how to connect an engine to a car, would take place. In this case, unlike Chouinard's definition

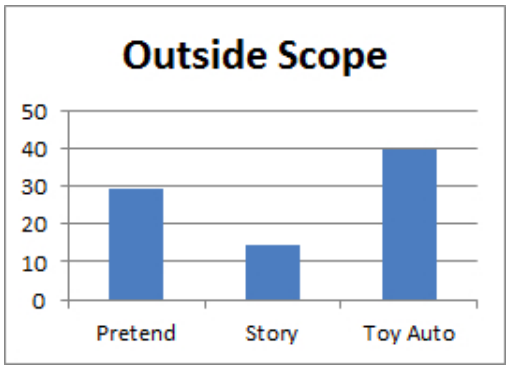

Figure 4.8: Percentage of questions that are outside scope across activities.

of information-seeking questions, "outside scope" can include clarifications of what an adult said. On the other hand "outside scope" is much more conservative. Fact-seeking questions such as "where is it?" are considered within scope because they only pertain to the activity at hand. Specific to this study, a particularly difficult piece of this was determining whether children asking about scenes in the picture book should be considered "outside scope". It is difficult from text only to determine whether children are asking what is happening because they want to know what i.e. the event transpiring on the page is called or because they simply want the adult to continue fabricating the story. In the end, questions about events in the picture book were deemed to not be "outside scope" because the pictures in the book are of well-known objects and scenes that if not in a storybook, would not be ambiguous in meaning. Another difficult piece was determining whether questions of how to piece together the toy auto was "outside scope". In the end, this was considered "outside scope", since it teaches the child reasoning skills of how mechanical objects fit together.

Since most of the questions children asked while reading the picture book involved what explicitly was transpiring on the pages, children could ask few "outside scope" questions during this activity. On the other hand, since the toy auto activity was focused primarily on constructing the auto, the child could ask many questions on the mechanisms, pieces, and properties of the car. There were quite a few "outside scope" questions in the game of pretend as well, since even though children were involved in role-playing, the children were unfamiliar with many props and asked for their labels. As a note, there were few "outside scope" questions during reading most likely resulting from the fictional nature of the story and the listener-role adopted by the child. If the book were non-fiction, the results might be very different. The percentages of child questions that are outside scope are presented in Figure 5.

By determining the type of interaction an activity promotes, we can better decide what types of activities might best engage a child in asking information-seeking questions that will build their developing knowledge structures. If there are unfamiliar objects, the child will ask object identification questions. If there is ambiguity in the activity, such as in a

	#	%
Why Questions Following Negative Statement	175	14.31(from "why" questions)
Total Why Questions	1223	
Questions Following Negative Statement	799	6.41 (from all questions)
Total Questions	12474	

Table 4.3: Number and percentage of why questions and all questions following a negatively phrased statement

storybook reading, then the child will ask ”why” questions. Activities that involve guidance, such as constructing something, will prompt the child to ask ”how” question. Despite the few activities analyzed in this corpus, the statistics of line types as a function of activity will provide insight into developing an interface suitable for question answering exchanges with children.

The use of "what?"

Beyond the categories labeled, a few other observations were made that are relevant to the design of a conversational agent. The question ”what?” in response to a parent's question was often ambiguous, in that the child could be asking for clarification of the question or could be asking the parent to provide an answer to the question. Upon closer analysis, it became more apparent that after the "what?" question by the child and a repeat of the original question by the adult, the child generally responded in two ways-either the child answered the question or repeated "what?" again. The limitation of the transcripts are the lack of marking for intonations in a child's voice. A "what?" requesting for clarification would have a rising intonation, whereas, a "what?" requesting for an answer would sound more like a statement. If a machine were to answer "what?" question, it would need to take into account the acoustics, and not simply the text representation of the child's speech. There may also be other questions similar to "what?" that require this acoustic analysis.

Another observation was the increased number of "why?" questions following a negatively phrased statement by an adult. Since we are interested in promoting meaningful questions from children, this observation motivates the study in section 3.3.5, which examines the ratio of "why" questions following negative phrased statements versus other question types.

Effect of Negatively Phrased Statements on Children's Inquisitiveness

Since children ask questions to solidify their understanding of the world, it should be expected that questions contradicting children's current beliefs should prompt them to ask more "why" questions. In the Callanan and Oakes study, this was coded for by checking for "why" questions which contain negative words and phrases, such as "not" and "can't", and they found that the proportion of "why" questions with negative words and phrases to be low overall. As we are interested in ways to promote meaningful question asking in children, we decided to approach

this from a different angle. Instead of looking at "why" with negative words or phrases, we look at the number of "why" questions as a result of an adult making a statement in a negative way. The percent of "why" statements following adults that use any of the words "not", "no", "neither", or "never", including contractions, was compared with the percent of other questions following the negative words. For this study, we use all free-interaction studies for the age range 3-4 described in the Appendix. The results are available in Table 4.3.

Without further analysis, we can only make hypotheses for the greater percentage of "why" questions following negative statements. Children could, as mentioned above, be asking "why" questions because their expectations of the world were violated. Other possibilities include conversation formalities, greater comfort with the language structure of "why", increasing the likelihood of being granted permission to do something that was originally forbidden, etc. In conversation formalities, the child may ask "why", as a way to express interest, which is an important for maintaining conversational discourse. There are many studies on children repeating statements by parents, so there is also support for children using "why", because repeating a commonly used phrase can be related to repeating a previously said line. Asking "why" to persuade adults for granting permission is a probable hypothesis as children often ask "why" when denied permission; however, more often than not parents will still deny the child's request after being asked "why" which reduces the plausibility of children asking "why" to be given permission. Regardless, the significantly higher percentage of "why" questions following negative statements suggests that using negative statements in a conversation can be a useful technique for prompting children to ask "why".

4.4 Conclusion

These context-dependent levels of engagement motivate further exploration of what activities would maximally elicit children's question-asking during interaction with a machine agent. We propose a qualitative, wizard-of-oz study that could be deployed in the near future serving the following two purposes: 1) to compare how children interact with an intelligent interface across different contexts, and 2) to test the accuracy with which our database answers questions commonly asked by children. To simulate interactivity with the conversational agent, the experimenter would manually "wizard-of-oz" or input the child's questions into the system, and the system would respond either with an answer to the question or with a response that would perpetuate the conversation. Implementing such a wizard-of-oz study would help inform which contexts are more likely to engage children during interaction with a machine interface, and moreover evaluate the effectiveness of our question-answering system.

Re-examining our categorization results from the Gleason database [52], we also observe that storytelling activities tend to elicit a greater percentage of "why" questions, while constructive activities such as the auto-building scenario yield a higher frequency of "how" questions. Therefore, we could also envision a study that explores a combination of these two scenarios through a storytelling activity that pushes children to more actively create stories within the very nature of the task. Child participants would be presented with either a regular

storybook or with an assortment of pages from the same storybook in no particular order. In both situations, the adult would help guide the child in creating or telling a story. We predict that the latter scenario would elicit more "how" questions in ways similar to the Gleason toy auto scenario, while still maintaining "why"-type questions that arose in the Gleason story-telling activities: in the absence of a prescribed page order inherent to regular storybooks, it is likely that children would not only require more direction from adults, but also be more physically and creatively engaged with the task as they manipulate and re-order the pages to construct the story.

However, in order to test the accuracy and appropriateness of our technology, we considered ways to elicit a large number of questions from children within a short amount of time. One possibility was to model after the "twenty-questions" game, in which children are prompted to ask a number of yes-or-no questions in an effort to guess which of two objects had been hidden from them. Although such a task would stimulate a high frequency of questions, it would test only a narrow subset of question types related specifically to object identification (i.e. "Is it blue?", "Is it bigger than a house?"), leaving out many other meaningful forms of question-asking that commonly arise in children's speech (i.e. "Why", "How", and other casual questions). Moreover, since the ultimate goal of the conversational interface is to expand and deepen the child's understanding of the world, an agent that confirms facts that are assumed to be in a child's prerequisite knowledge base would not be sufficient or even appropriate for testing overall learning gains. However, children use questions as an important tool in knowledge formation, and practice with that can have long-lasting effects in a child's cognitive development [10].

Chapter 6 presents and design, development and evaluation of such a system. However, before implementing an automated system that engages preschoolers in activities, we also conducted some computational experiments with the CHILDES database [47]. This was done to determine how open-ended such a system could be, and what kind of computational directions could be pursued in the future. Results from these experiments are presented in the next chapter.

Chapter 5

Computational experiments with CHILDES

5.1 Derivation of Question Patterns

Question patterns[1] and categorizations can be used to determine how a question should be answered. For children, whose mean length of questions is 5.05 words and whose median length is 4 in the free-interaction transcripts described in Appendix, question patterns are generally restricted to a few basic parts. To determine the structure of various questions types asked by children, we attempted various methods and technique of clustering.

5.1.1 Techniques for Clustering

TFIDF of Children's Questions and Parents' Responses

The standard metric for determining a term's relevance in a child's question relative to its occurrence in all child utterances is measured using term frequency inverse document frequency (TFIDF).

$$TFIDF_{w_i} = \frac{word\ w_i\ count\ in\ questions\ by\ child}{total\ word\ count\ in\ questions\ by\ child} * log(\frac{\#\ lines\ by\ child}{\#\ lines\ with\ word\ w_i}) \tag{5.1}$$

The TFIDF for each word in an adult's response can be computed similarly.

Stop Words

A stop word is a word that occurs with equal likelihood across relevant and irrelevant documents retrieved for a query [46]. We calculate stop words as the most frequently occurring words that a child uses that does not have a high TFIDF value. 63327 total lines of conversation and 12474 child questions, retrieved from free-interaction transcripts in the CHILDES database, were used to generates stop words, and a sample of these stop words are listed

[1] This work was done in collaboration with Ingrid Liu and Carrie Cai.

Stop Word	No. of occurrences from 12474 questions
yeah	4487
no	2999
going	1661
not	1229
look	1195
because	1118
oh	1081
yes	985
at	956
say	939
little	933
then	909
all	899
make	844
uh	843
two	841
big	837
eat	836
them	829
take	821

Table 5.1: Stop words for children's questions, ages 3-4.

in 5.1 and 5.2. Upon inspection of the stop words, however, many were deemed to still be relevant to categorizing questions. For example, "not" is used frequently in "why not", but because "why" questions are more infrequent, "not" does not have a high enough TFIDF value to be altered from the stop words. "Because" is another word that occurs frequently in statements and infrequently in questions, but would be important in that it marks the causal nature of the question. Since a child does not have a wide vocabulary, relatively more words used by a child are essential for conveying the child's meaning, in comparison to an adult who can string complex sentences with words such as "and", "moreover", etc. This automatic generation of stop words must be altered by a human before it is usable, and further studies are necessary to determine what words in question categorization tasks would be appropriate stop words.

Other pre-processing steps

- **Stemming:** Words were stemmed to remove plurality and tense.
- **Part of Speech Replacement:** Nouns, adjectives, adverbs, determiners, and demonstratives were replaced with part of speech placeholders. Without this step, clustering fails,

Stop Word	No. of occurrences from 12474 responses
say	1251
now	1222
all	1149
she	1135
at	1010
good	986
out	975
them	957
ya	929
tell	926
well	915
take	843
when	795
would	754
huh	750
who	735
him	699
eat	664
then	659
down	648

Table 5.2: Stop words for adult's responses to children's questions, where the children are ages 3-4.

because the utterances by the child and responses by the adult are too short for a bag of words approach to work. Verbs were not replaced, because many question types are identifiable only by the verb, i.e. "how come?".

- **Contraction Expansion:** Contractions, as well as common slang such as "gonna" were expanded into their proper forms.

5.2 Methods of clustering

Three clustering questions were attempted: clustering based on the syntax of the sentence and based on a bag of words approach.

5.2.1 Clustering by Syntax

To cluster by syntax, the lexicalized probabilistic parser developed by the Stanford Natural Language Processing Group was used to generate parse trees for every sentence. Since trees

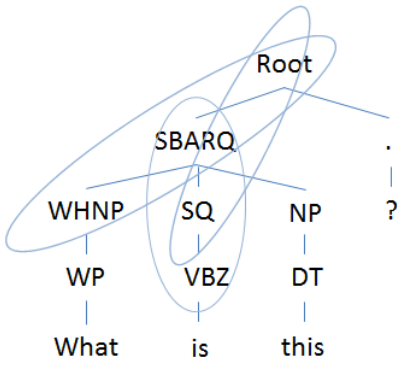

Figure 5.1: Sample features from a parse tree used when clustering by syntax.

are generated, the structure was captured in features by using paths of nodes, up to length 3, and these features were fed into the standard k-means classifier. As an example, a few of the features for the sentence "what is this?" is circled in 5.1. This attempt to cluster by syntax was not possible because of the ill-formed structure of children's questions. Many questions are fragments, two different questions that are incomplete, or out of order. Words and tenses are also used incorrectly. Parse trees are generally complex structures where the addition of an out of place word will completely alter the rest of the tree.

5.2.2 Clustering by Bag of Words

To cluster by bag of words, the terms were from the questions and responses were weighted by their TFIDF value and these features were fed into the standard k-means classifier. It was inferred, that since children lack proper structures, a bag of words approach ignoring their lack of grammar would improve the clusters found. Furthermore, the responses would have key words such as "because" that would be relevant features. Analysis showed, however, that since children generally use short questions structures, they are comfortable enough with the structures such that most parts within the structure are in the proper order. When the questions are ill-formed, it is the result of missing words. With this view of child question asking, a much more direct approach to categorizing questions was tried next: creating a hierarchy of questions.

A prevalent line of research examining child language acquisition is based on the theory that children segment phrases into meaningful units [67]. Studies have shown that whereas adults may use referential context to parse phrases, a child is less adept at using referential context and must instead use other methods based on their knowledge of individual verbs and their semantic/syntactic rules [77]. In this view, a child uses constraint-based lexicalized methods of parsing [77]. If children's processing of sentences is similar to children's language generation, then questions may be better identified by determining what meaningful units children commonly develop at a pre-school age, and categorizing questions based on these

units.

In this simple construction of a question hierarchy, questions are again viewed for the bag of words perspective, since it is a useful representation that captures the structure of sentences such as "what is xxx for?", where xxx can be anything.

To check whether this would be a good classification method, we use a baseline frequency classifier that utilizes a hierarchy of questions compiled from a set of extracted questions. The best characterization of a complex question q_i is chosen to be the question q_{j_m} of length m, whose words are a subset of the words in the complex sentence and has the highest value calculated as

$$\text{value}_{q_{ij_m}} = \frac{\#\ m \text{ length questions}}{\text{total questions}} * \frac{\#\ \text{questions that contain } q_{j_m}}{\sum_{q_{k_m}} \#\ \text{questions that contain } q_{k_m}} \tag{5.2}$$

To view the appropriateness of this categorization approach, WH questions, including why, how, when, where, and what, from the free-interaction datasets in CHIDES were used, described in the Appendix. WH questions are used, since they are almost always information-seeking questions, which is our focus. A few randomly selected questions and their chosen characterizations are presented in 5.3.

5.2.3 Discussion

Some of the characterizing questions are very similar to the original question. Since the questions are drawn from the same set, a child may repeat a question with slight modification, and this repeat question will be the most similar constituent part in the original question. If this approach were to be used, a few questions would have to be manually tagged as characterizing question types, and the classifier could choose from among those tagged questions for the characterizing question. Also, questions which are a combination of two questions confuse the classifier, which indicates that the bag of words approach should still weight words that are adjacent to each other more highly. The question "why do you call Marky have money?" can be decomposed into two questions: "why do you call Marky?" and "why do you have money?", but since the classifier views it as one question, it chooses "why do you have coffee?" as the characterizing question.

5.3 Object Identification in Conversational Discourse

For data, we used the following database within CHILDES: Brown [5] described in Appendix. Questions make up roughly 20% of the lines spoken by the child in this dataset. Object identification questions are of particular interest, because they encapsulate the main difficulty of question answering for children-identification of the surrounding context-while remaining a well-formed question. As a result, we decided that a good first step towards the hybrid problem of question answering framing, co-reference resolution, and information extraction, would be to attempt the object identification question.

Question	Characterizing Question
what this boy sleeping in something ?	what this noise ?
why are we gonna bring the suitcase ?	why is the turkey ?
why you do that to the truck ?	why you do that ?
some what I xxx in there ?	what in there ?
why do you call Marky have money ?	why do you have coffee ?
no why would you get money ?	why no paper ?
why the hand is out like this ?	why is the story like this ?
Mommy how do they put the foot in there ?	how do they put the foot in there ?
except what do you put in this ?	yeah why do he have to go roller skate ?
what you do ?	why he have roller skate ?

Table 5.3: Characterizations of questions by using a hierarchy of questions.

Line Type	Number of Lines
spoken lines	22921
questions	4698

Table 5.4: Distribution of lines spoken by a child ages 4-5.

There are several steps necessary for answering these object identification questions. Information extraction is needed to learn the possible labels of the hidden object label and to estimate the hidden objects and relations (those that are not verbally mentioned in the conversation). Context modeling is needed to estimate the flow of the conversation, or how its different parts relate. A method of evaluating the success of the system is also needed.

5.3.1 Background

The task at hand is a co-reference resolution problem, despite the unknown object's position in a question, because of the constraints on the question's form. There have been several previous works on co-reference resolution, question/answering, and semantic interactions within conversations. Question/answering has been discussed previously, so we only discuss co-reference resolution and information extraction below.

Coreference Resolution

The most successful approach to co-reference resolution to date utilizes Named Entity Resolution for generating coarse-grained entity types and a modular approach, by Haghighi and Klein. To clarify, a mention is an unresolved noun phrase instance in text, an entity is a specific individual or object in the world, and a type is a set of entity classes. The semantic module generates entity types, which group together the different types that a mention can be, using unsupervised learning. The discourse model chooses what types each mention

should be. Finally, the mention generation module assigns entities to each mention [22]. Previously, the top performing system utilized decision trees classifying on lexical, grammatical, semantic, and positional features, and a clustering algorithm to partition the noun phrases in a text [59].

Information Extraction

Question Answering is generally done by information retrieval (IR) and query-based summarization. There are several ontologies available for IR, including WordNet, Suggested Upper Merged Ontology (SUMO), Verbnet, Framenet, and Open Mind Common Sense (OMCS), or even Wikipedia. In addition, IR can include web searches such as Google queries. With a wealth of knowledge, the main difficulties in IR are document ranking and answer extraction.

OMCS, developed at the Media Lab at MIT is particularly useful tool for interpreting human reasoning, in that its relations are defined and rated by humans. By August 2002, the commonsense knowledge acquisition system had collected over 450,000 commonsense assertions from over 9000 people [74]. Each relation includes a score, which is the number of positive votes given by individuals, and this serves as a soft "document ranking".

5.3.2 Problem Formulation

Data Set

We use the free-interaction transcripts to extract contexts surrounding children's object identification questions. The preceding 10 lines, or less if there are fewer than 10 lines before the question in a transcript, serve as the "context" around which the question is asked. Only questions that are of the form "* what is this, that, these those" are accepted.

Problem Statement

Given the context before a child's object identification question, the system needs to model the conversation such that a reasonable prediction of the object can be made. The system should use commonsense knowledge to retrieve the list of objects the child could be referring to and to return an object consistent with the context given so far. Note that this is much harder than the task the overarching project should be able to solve, which would be for a system to give the child a focus to their conversation (i.e. a list of objects), and to answer questions given the conversation's focus and commonsense knowledge. However, since there is no data available for the task the overarching project should solve, the task is generalized to resolving context ambiguity within CHILDES transcripts.

5.3.3 Model

From empirical examination, the dialogues between child and adult generally proceeds serially and is furthermore always centered around physical object(s). The context can thus be

modeled as two set of objects: (1) a series of physical objects, which are connected temporally within the conversation and (2) hidden context, which include the child's knowledge of the world's other objects and their relations, since speech act modeling is outside the scope of the project at this point in its development. A graphical model is used to represent the probability that any of the objects described above is the object to be identified. The prior probability of an object is unknown, but we can model the potential between the object and the evidence available in the context as a function of the location of the object's mention in the conversation, the object itself, and whether the object was brought up in the conversation or extracted from OMCS. The edge potentials between any two objects is a function of the relations between the two objects.

$$\Psi(x_i, y^*) = f_1(\text{loc}(x_i), \text{source}(x_i)) \tag{5.3}$$

$$\Psi(x_i, x_j) = \max_k[f_2(\text{relation}_k(x_i, x_j)] \tag{5.4}$$

Each object x_i is binomial and takes on the values 0,1 for depending on if it is the object to be identified. The MAP solution to this problem is

$$\text{argmax}_x P(x, y) = \text{argmax}_x[\prod_c \Psi_c(x_c)] * [\prod \Psi_{ii}(x_i, y_i)] \tag{5.5}$$

where $x_i = \delta(x^* = x_i)$. This is simplified to the following equation when the cliques are pairwise [40]:

$$\text{argmax}_x P(x, y) = \text{argmax}_x[\prod_{i,j} \Psi_{ij}(x_i, x_j)] * [\prod \Psi_{ii}(x_i, y_i)] \tag{5.6}$$

for $x_i = \delta(x^* = x_i)$.

Graph Formulation

The set of possible objects can be found by walking through levels of the OMCS relations. After walking one hidden level deep, the Bayesian network is framed as shown in 5.2. As a note, at the i=0 level, which comprises the physical objects in the conversation, mentions of an object can be repeated. I.e. if dog is mentioned twice in different parts of the conversation, it appears in two nodes of the i = 0 level. This occurs as a result of the objects possibly not co-referring, and the potentials between object and evidence being different if the object appears in different locations. At the (i+1)th level, all objects related to objects in the (i)th level are added and connected to their (i)th parent. These objects are unique, and multiple (i)th level objects can refer to the same (i+1)th level object. All edges are added where the edge's weight is given in the next subsection when the edge does not equal 0. As a note to the reader, even though the objects in the 0th layer have already appeared in the conversation, they may still be the object the child is asking for.

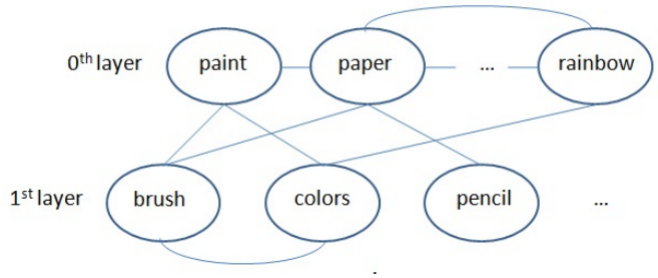

Figure 5.2: Sample graphical model of a context.

Relation Weighting

Here, we calculate the weight of a relation between two objects x_i and x_j and the relation between an object x_i and the evidence y. The object to object relation weight is a function of the score available in OMCS, analogous to a tf-idf score while the object to evidence relation weight (also chosen as the potential) is a function of the object's position, and its level.

$$\psi(x_i, x_j) = \frac{\text{score (relation}(x_i, x_j)}{\sum_{k=0}^{N-1} \text{score(relation}(x_i, x_j)) * log(N)} \tag{5.7}$$

$$\Psi(x_i, y) = f(\frac{\text{pos}(a(x_i))}{W}, \text{level}(x_i)) = \begin{cases} \delta & \text{level } i = 0 \\ f(\frac{\text{pos}(a(x_i))}{W}) & \text{o/w} \end{cases} \tag{5.8}$$

where W is the number of physical objects in the context, a(x_i) is the i=0 level ancestor of the highest position, δ is some small number, and pos(a(x_i)) is a position of the physical object x_i in the context and is equal to pos(x_i) when x_i is at the i=0 level.

Approximate Inference Algorithm

The procedure described for the model's network creation yields many loops. Furthermore, there is an added constraint that the approximation algorithm should be computed quickly, since it will be calculated in real time. We select Loopy Belief Propagation (LBP) as our approximate inference algorithm as a generally very accurate MAP approximation except when non-convergent. In multiple graphical structures, Murphy, Weiss, and Jordon demonstrate that LBP far outperforms likelihood weighting when it converges [58]. When LBP does not converge, and the hidden variable values oscillate, the performance is very poor, and averaging the values yiels poor performance [58]. Murphy et al. found that small priors or small weights will cause oscillations. In our studies, we found that a symmetrical potential matrix will cause oscillations, and our hypothesis is that the symmetry and small loops in our graph structure caused the oscillations. (The potentials of (x_i, x_j) is a 2x2 matrix, where 2 is the number of values a node xi can take, specifying the compatibility between nodes x_i and node x_j). To

assignment	consistent	inconsistent	indeterminable
percentage	0.27	0.43	0.30

Table 5.5: Percentage of object identifications that are consistent, inconsistent, or indeterminable given the context.

assignment	reasonable	unreasonable
percentage	0.24	0.76

Table 5.6: Percentage of object identifications that are reasonable or unreasonable, given that a fair guess can be made given the context.

address this problem, the potential between two objects x_i and x_j were taken as

$$\Psi(x_i, x_j) = \begin{cases} \max(\max_{x_i,x_j} \psi(x_i, x_j), \max_{x_i,x_j} \psi(x_j, x_i)/2) & x_i = 0,\ x_j = 1 \\ \max(\max_{x_i,x_j} \psi(x_j, x_i), \max_{x_i,x_j} \psi(x_i, x_j)/2) & x_i = 1,\ x_j = 0 \\ \delta_1 & x_i = 1,\ x_j = 1 \\ \delta_2 & x_i = 0,\ x_j = 0 \end{cases} \tag{5.9}$$

since the set relation(x_i,x_j) does not equal the set relation(x_j,x_i) and where $\delta_1 \ll \delta_2$. We also justify this asymmetry in the edge potentials due to the fact that relations are also asymmetric. The parameter "2" is a heuristic that works well in practice.

For a non-tree graph, if LBP converges, it will converge to a "neighborhood maximum" of the posterior probability, or in other words, the MAP will be greater than other assignments in that neighborhood, where the neighborhood is relatively large [40].

5.3.4 Preliminary Results

To evaluate the system, we had five colleague evaluate 20 random assignments by the system at two levels:

- Consistent with the context, inconsistent with the context, or impossible to tell due to a lack of context. This is given in 5.5.
- Given that a guess can be made with context, whether the guess is a reasonable answer to the object identification problem or not reasonable. This is presented in 5.6.

5.3.5 Discussion

The LBP algorithm converges very quickly for the task, but the results are disheartening. There is a lot of work to be done in more robust potential assignments, which are specific

```
*CHI:   I never heard of a baby string (.) have you ?
*CHI:   she getting milk .
*CHI:   I like to sing Happy_Birthday .
*CHI:   yeah (.) something funny .
*CHI:   are the juice ready ?
*CHI:   I want Ursula to listen to the music .
*CHI:   you weren't coming because Paul had the chicken pox .
*CHI:   I can't .
*CHI:   I don't want to .
*CHI:   I'm eating dis [: this] clown up .
*CHI:   what is dis [: this] ?
```

Figure 5.3: Error cause: Randomness in the conversation. The objects brought up range from baby string, to milk, to Happy Birthday, etc.

to the dynamics between child and adult conversations. A discussion of common errors are presented next.

Discussion of Common Errors

- The randomness of a child's ramblings will make the object they want identified too ambiguous. See 5.3.
- The many objects are related, but their most pronounced shared feature does not continue the flow of conversation. This is caused by local maximization of posterior probabilities rather than finding a global optimal. See 5.4.
- The weights don't have a threshold. Given a relation with a high weight, the object may be chosen despite not matching the rest of the context. See 5.5.
- Word sense disambiguation is not performed. For example, in some instances, a relation between a noun x_i and a verb x_j adds a node for x_j, because x_j can also serve as a noun. See 5.5 again.
- Overall, a lack of information in the dialogues prevents correct classification. See 5.6.

Applications to the Design of an Intelligent Conversational Partner

This problem was an initial step towards determining the complexities of interpreting the conversation of a child, which may be conceptually random and ungrammatical. The problem chosen was more difficult than what our planned system will do, in that classifier must also guess the context from which the conversation originated. With a focused activity, our system will ideally only have a small set of concepts from which the child could be asking about. Conversely, it only addresses a small subset of the questions an intelligent conversational agent must be able to address to answer all question types a child may ask

```
*CHI:   what ?
*MOT:   is that an eagle ?
*CHI:   eagle .
*MOT:   is that an eagle ?
*CHI:   no .
*MOT:   is it a crow ?
*CHI:   no .
*MOT:   is that an owl ?
*CHI:   yes .
*MOT:   it's an owl .
*CHI:   what is dis [: this] ?
```

Figure 5.4: Error cause: Shared similarity between objects are not consistent with the conversation. This context was labeled "eye".

```
*MOT:   what's a baby dog ?
*CHI:   a kitten .
*CHI:   a kitten .
*MOT:   a kitten ?
*CHI:   a cat .
*MOT:   a cat ?
*CHI:   uhhuh .
*MOT:   a dog (.) Rinny could be the mother of a cat ?
*CHI:   yeah (.) and de [: the] cat's gonna be the mother of a big baby .
*CHI:   oh (.) wasn't that fun ?
*CHI:   and what is dis [: this] ?
```

Figure 5.5: Error cause: No thresholding of weights and WSD. "Cry" has a very high weight and is accidentally classified as a "noun" even though the relation between "baby" and "cry" is noun-verb.

```
*FAT:   do you want me to look at it ?
*FAT:   let me see if I can figure it out .
*FAT:   v is for vase .
*FAT:   a special vase .
*CHI:   v is for vase .
*FAT:   w is .
*CHI:   w is for fish .
*FAT:   for whale .
*CHI:   for whale and x is for x_ray ?
*FAT:   yeah .
*CHI:   what is this ?
```

Figure 5.6: Error cause: Lack of context. This is a sample of a father and child reading an A-Z book together, but that information is not easily gleaned for a machine.

and only considers a small portion of a conversation. In the design of a conversational agent, discourse segmentation would be needed.

5.4 Future Directions

- A test set more relevant to the desired task of object identification given a context would give better indication of how well the inference scheme works. This must be compiled from new experimental studies. Based on the number of test samples that people were able to classify, the test set lacked enough context to answer the object identification question well, and it can be presumed that object identification from transcripts of adult/child dialogues is too difficult. Furthermore, there is no gold answer to the test samples available.

- Experimentation with deeper relation nets. This study examined a graph constructed only to two level-one of physical objects and one of hidden object directly related to the physical objects.

- Segmentation of the conversation based on a system that determines coherence between lines in the conversation.

- Reexamination of whether local approximations are adequate through more experimentation, particularly because the local approximations are faulty in representation. In practice, the system can find good guesses of the object from the context under certain conditions; however, consider a message:

$$m_{ij}(x_j = 1) = \max_i \Psi_{ij}(x_i, x_j = 1) * m_{ii}(x_i) \prod_{x_k \in N(x_i) \setminus x_j} m_{ki}(x_i) = \tag{5.10}$$

$$\Psi_{ij}(x_i = 0, x_j = 1) * m_{ii}(x_i = 0) \prod_{x_k \in N(x_i) \setminus x_j} m_{ki}(x_i = 0) \tag{5.11}$$

  for the most part, because only one node x can be the object to be identified. Next, consider the product:

$$m_{ki}(x_i = 0) = \max_k \Psi_{ki}(x_k, x_i = 0) * m_{kk}(x_k) \prod_{x_l \in N(x_k) \setminus x_i} m_{lk}(x_k) \tag{5.12}$$

$$= \Psi_{ki}(x_k = 1, x_i = 0) * m_{kk}(x_k = 1) \prod_{x_l \in N(x_k) \setminus x_i} m_{lk}(x_k = 1) \tag{5.13}$$

  which implies that in the message from x_i to x_j, there is another node x_k which is the object to be identified, even though we are solving for a message where x_j is the object to be identified.

- Attempting an EM approach to this problem. For pronoun resolution, Cherry and Bergsma decompiled the data points into triplets: (p,k,C), where p = pronoun, k = context, and C = list of candidate nouns. The E step estimates the probability Pr(C|k,p) using a set of rules while the M step estimates the probability Pr(p|l), using Pr(C|k,p) to compute the fractional counts to each of the candidates, where l is a the lexical component of a candidate c [9].

In essence, it is clear from the points made above and in the previous chapter that building a question-answering system for preschoolers requires the activity to be focused and yet engaging. An open-ended solution to this problem is hard to conceive at this stage, given the lack of data and age-appropriate content on the web. Therefore, the next chapter discusses the design and development of a question-answering system called Spot, that engages children in an activity very similar to twenty questions. The choice of this activity comes from previous research in child psychology [10].

Chapter 6

Spot: A Question-Answering Game for Preschoolers

6.1 Envisaged Solution

As a potential realization of our intended system we envision a projection system with a virtual character that acts as the question-answering agent. The aim is to create a virtual play space that can keep the conversation grounded in a context. Grounding is important because context specific speech recognition is more feasible than one in scenarios that lack context [19]. The virtual character (usually an animal), rendered over a wide area using a projection system, will engage children in language games that use question-answering as the primary dialogue structure. For example, the character will show the child several objects, then hide one and ask the child to guess what it is by asking questions about it. The game engages children in language use, and also in concrete questions about things in the world and their properties. The envisaged solution is shown in Figure 6.1.

6.2 Current Work

In this chapter we describe a two-phase study, one phase using a human language partner, and the second using a system which approximates Figure 6.1. Rather than relying on speech recognition and dialog interpretation, we used a Wizard-Of-Oz system. The goal of the studies was to explore the feasibility of the envisaged solution: whether students would ask "on-topic" questions, whether the questions matched some templates, and whether they would be engaged by the game. Phase 1 involved 20 children studying at the same preschool, playing a 20-questions game with a familiar researcher. It contributed to answering the following research questions:

- Are children's questions predictable and deterministic, when grounded in an activity like 20 questions?

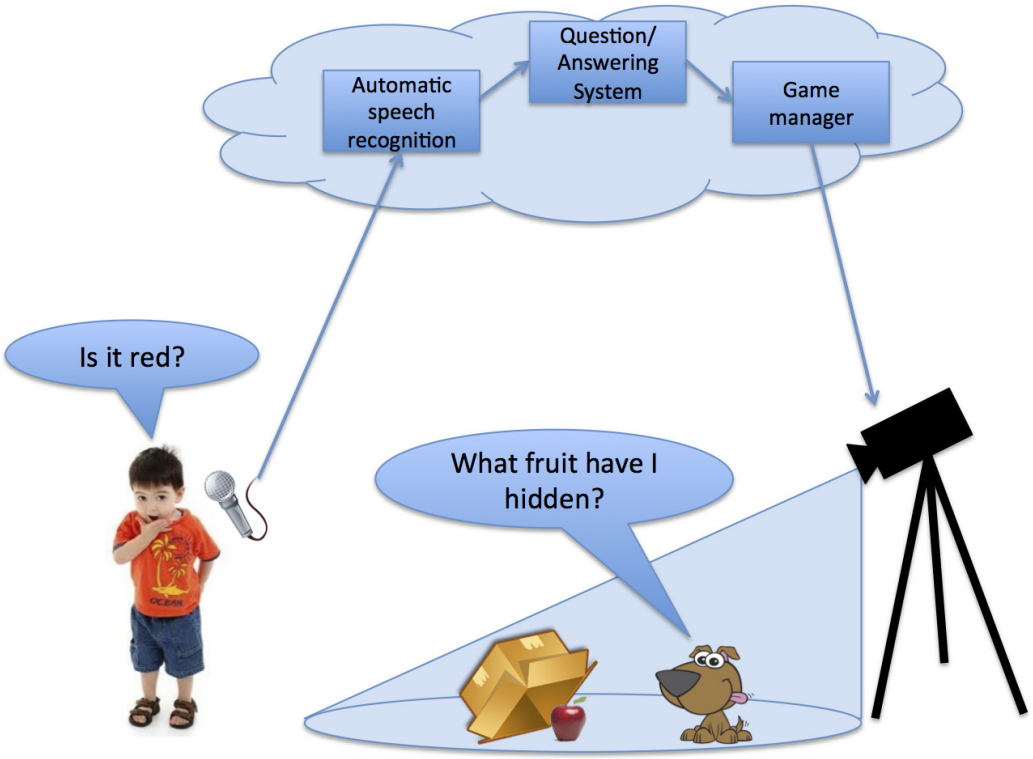

Figure 6.1: The envisaged solution

- Is the repair required in such a dialogue limited and feasible?
- Is it possible to effectively "nudge" preschoolers to solve problems without disengaging?

Phase 2 involved the same participants as phase 1. Half of them played the game with the same researcher. The other half played the same with Spot, an agent that we designed and implemented. Effectively, 10 children played just with the familiar human in both the phases and the rest 10 first played with the human and then the agent. Phase 2 built on phase 1, and answered the following research question:

- Using commonly used parenting styles in dialogues; how can we design an agent that can engage preschoolers in a familiar question-answering activity as effectively as a familiar human?

Effectiveness in this case is primarily restricted to question-answering efficiency, flow of communication and affect/engagement.

6.3 Related Works

With the growth of conversational technologies, the possibilities for integrating conversation and discourse in e- learning are receiving greater attention in both research and commercial settings. Conversational agents have been produced to meet a wide range of applications, including tutoring (e.g. [20], [25], [1]), question-answering (e.g. [13], [16], [80]), conversation practice for language learners, (e.g. [18], [71]), pedagogical agents and learning companions (e.g. [43], [70], [3], [14]), and dialogues to promote reflection and metacognitive skills (e.g. [21], [36]). Conversational agents build on traditional education systems, providing a natural and practical interface for the learner. They are capable of offering support for each individual, and recognizing and building upon the strengths, interests and abilities of individuals in order to create engaged and independent learners. However, the current interactive conversational tutors are geared more towards older children, who have a larger set of knowledge or skills than pre-school children and are easier to understand, and also focus on specific skills or domains.

The key difficulty in developing an agent for such a younger audience is maintaining children in their ZPD (Zone of Proximal Developent) [80]. The project CACHET examines the responses young children have to interactive conversational agent using electronic stuffed toys [45]. These toys are designed to speak, respond to touch via sensors, gesture with motors, and be linked to a PC wirelessly to provide support and feedback while a child plays games encouraging number and language learning [45]. Children were able to skillfully navigate through the games, however, and were adept at asking for help when they were aware of and were not irritated with the toy [45]. This technological adroitness suggests high potential for interactive agents for younger children.

There is also some recent reflective work in media psychology [38], education [11] that critically analyses the results of experiments with pedagogical agents in general. Most such work calls for testing with younger audience. Moreover, prior research has shown that projection systems can be effective medium of interaction for younger children. This could either be for remote collaboration with other children [81] or communication via media with parents [82]. Overall, we did not find examples in the literature of specific systems that use projections of virtual agents; to do question-answering based games for preschool children.

6.4 Phase 1: Feasibility Study

6.4.1 Participants

20 children (10 boys, 10 girls) participated in our feasibility study. The participants in the study were 4 and 5 year old children at a preschool in California. Previous research suggested that 3-year-old children would be too young for such an experiment [10]. The preschool that was chosen as the location for the study was a research preschool.

6.4.2 Equipment and setup

The study was conducted in a research room on the preschool premises, reserved for that purpose. The room was equipped with one-way mirror and audio equipment that allowed a visual supervisor to monitor the study at all times. The presence of the visual supervisor was required by the preschool's protocols. During the study two researchers were present for all sessions, in addition to the child. The children could see the researchers, but not the visual supervisor. A video camera recorded the child's and researchers' activity at all times.

6.4.3 Method

The preschool consisted of two classrooms, namely east and west. As per preschool protocol, no child was allowed to be outside the classroom for more than 20 minutes in one session. No child could attend more than three study sessions in a week. Overall all participants attended one such session, and the entire phase 1 took three weeks.

Each classroom had a circle time from 10am every morning, which is when the researchers got to interact with the participants before starting the study. The research team attended several (>5) circle times to become familiar with the study participants. Familiarity was important because a large part of the study involved playing games with researchers on the team.

During the study each child attended a session individually. Before each session a researcher from the team went to one of the classrooms and invited participants to attend a study sessions. Out of the consenting children one was escorted to the study room. As stated above, the study session could not last more than 20 minutes.

Each session comprised of multiple question-answering exchange trials. Each child was shown photographs of two objects. After this both the photographs were shuffled, one was put away and one was retained. The child was told that the purpose of the game is to ask questions and figure out which object is being retained. The session started with a demo trial, where the two objects were cat and ball. In the demo trial, one of the researchers asked questions and another researcher answered. If the child did not understand the demonstration, it was repeated till the child was comfortable in contributing to the questions being asked. After the demo trial, 6 more trials followed. Each child was asked to identify the two objects at the start of each trial. For each question that a child asked, a truthful answer was given. After each question-answer pair, the child was asked if they wanted to ask more questions or were ready to guess. Sometimes the child would just guess without perceivably having enough information to make a guess.

Each child went through 7 pairs of objects, including the demo trial. For each trial, the stimulus pair of objects presented got more difficult. Increased difficulty meant increased similarity in the stimulus pair. For example, a cat and a ball are easily distinguishable, but a bicycle and a car are harder to differentiate. In formal terms, the more difficult stimulus pairs were closer to each other in terms of parts, functions and properties. A list of all the pairs is

Target Item	Low Similarity	Moderate Similarity	High Similarity
Phase I	Book/Banana Elephant/Spoon	Table/Bed Shoe/Hat	Apple/Orange Bicycle/Car
Phase II	Chair/Rose Flower/Kite	Bear/Dog Chicken/Pig	Truck/Bus Clock/Watch

Table 6.1: Object pairs used in the two phases

given in Table 6.1[8]. It should be noted that no object was repeated across two trials. This was done to level the amount of practice children receive with each object.

6.4.4 Data Collection and Analysis

After the study, all the videos were transcribed. Care was taken that critical incidents like questions, explanations, hints and off-topic conversations were recorded. Previous literature suggested that most of these questions fall in one of the three categories: parts of objects, functions of objects and properties of objects [10]. Six individual coders/raters were asked to classify the questions into one of the three categories. There was substantial agreement in the ratings (Fleiss Kappa: 0.71, kappa error = 0.0117, kappa C.I. (95%) = 0.7022, z = 60.5003, p = 0.0001). Analysis of these ratings was done to answer the research questions allocated to this phase.

6.4.5 Results

Children asked questions

Participants in our study seemed to be naturally primed to ask questions. They asked a total of 210 questions, which means an average of 10.5 questions per child in a span of 20 minutes. It also turned out that the number of questions asked by a child was significantly correlated with the number of trials successfully solved ($r(10) = 0.7$, $df = 8$, $p<0.05$). In other words, children who asked more questions solved more problems. This jibes with previous research on the subject [8].

Questions had specific categories

Analysis of the coded data revealed that 80% of the questions asked by our participants were about the part, property or function of objects. The remaining 20% were guesses about what the object could be, example "Is it a cat?". Out of the non-guess questions, 46.5% were property related questions, 30.5% were function related questions and 23% were part related questions. Therefore, it was clear that most questions in such scenarios would deal with one of these three objects. This helped us create the database of possible questions expected by the agent (explained in next section).

Limited need for repeated explanation & tendency to guess without proper inference

Participants in our study needed limited explanation. On average researchers had to intervene only 1.67 times per child. This was generally in cases where a child responded in a manner that represented an inaccurate understanding of the game. In all such cases, one of the researchers would repeat the explanation of the game. All such cases helped us build conversational edge cases into the agent later. This is more formally known as repair/error-recovery in dialogue systems.

Children were adept at inferring the objects. In phase 1, they were able to successfully solve 104 out of the 126 (80%) trials conducted. Sometimes however, they would still try guessing the object without having enough information to solve the problem. This is in addition to the final guess that they would use to solve the problem. Overall 60 such inaccurate guesses were recorded across the 20 participants.

Off-topic dialogues and hint mechanisms

During the study, sometimes a child would talk about objects or contexts that were not grounded in the environment. Any such deviations from the task at hand were counted as an off-topic dialogue (average = 0.83, max =3). It was noticed that approximately 5% of the utterances by the participants were not grounded in the environment, and were counted as off-topic.

Sometimes (41 out of 476 statements, 8%) a child would find it hard to frame a question about two objects. Predictably, more than 70% of such cases happened for objects with high levels of similarity, namely: apple and orange, bicycle and car. In such cases, hint mechanisms were used to help them think of possible questions. Phase 1 was used to come up with effective hint mechanisms. The two types of hints (in that order, if required) that were used were:

- How are <object1> and <object2> different?
- Think about what <object1> has/does that <object2> does not or what <object2> has/does that <object1> does not?

It was important to keep the hints finite and deterministic as they would eventually be built into the agent. It turned out that hints were effective in what they were supposed to do. In about 90% of the cases of a child finding it hard to frame questions, a hint helped them. In cases where both the hints did not work, the trial was considered incomplete and the next trial was started.

Phase 1 helped us determine if building such a system is feasible, if children's questions are deterministic, if the repair required in dialogue is limited, if it is possible to effectively prod children of this age to solve problems without disengaging. The next section describes how the agent/system was built, given the feasibility study.

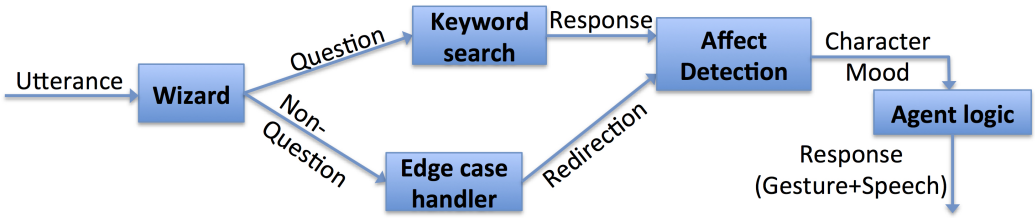

Figure 6.2: System Architecture

6.5 System Design

The system that was built, replicated the same task as in phase 1. An interactive agent in the form of a puppy dog character conducted the game sessions instead of a human. The character introduced the game, showed two objects, hid one and then played a 20-questions game till the child guessed the hidden object correctly. The character followed a script, the design of which is explained in the dialogue subsection.

6.5.1 System architecture

The system worked through modules shown in Figure 7.1. The speech recognition component of the system was wizard-of-Oz. Speech recognition for children is challenging. Collecting audio data and corresponding transcriptions while we evaluate our system, was another motivation for using wizard-of-Oz method [50]. So for everything that the child said, the wizard (a researcher) would transcribe the speech. Using the questions asked in phase 1, templates of possible questions were created. If the incoming utterance's transcription matched any of these templates, it was sent to a keyword search routine, else it was sent to an edge case handler that used re-directions (explained in detail under "Dialogue"). The keyword search routine depending on the matching template and the keywords used, determined the answer/response. Mostly this answer was either "yes, it does" or "no, it doesn't". If the keyword search routine was unsure of the structure, but sure of the semantics of the question, a yes/no response was given. If the question was an invalid (or incomprehensible) question, responses expressing soft disapproval were generated (explained earlier). Moreover, if the child found it hard to frame a question and there was silence for some time, the system would detect this and generate a hint. It is hard to gather enough contextual information in a conversation to be able to detect if the user understands the task at hand. Therefore, in cases where the wizard felt that the participant does not understand the activity, he would send a command directly to the agent logic and Spot would replay the explanation of the game.

Depending on the content of the response (from keyword search and edge case handler), appropriate affect features were generated. These features were used to predict Spot's gestu-

Response	Gesture
Yes/Yes, it does	Spot nods head
No/No, it doesn't	Spot shakes head
Right guess	Spot jumps around
Wrong guess	Spot turns shy and steps back
Idle	Spot sits down, licks foot
All other responses	Moves mouth to look like talking

Table 6.2: Table of verbal responses and corresponding gestures that Spot used

ral response. Table 6.2 summarizes all possible response types and corresponding gestures. The audio responses were pre-recorded audio. The voice used to record these responses was that of a female in her early 20s. She was born and brought up in the same geographical region as the location of the study. The frequency range of her voice was deemed appropriate for Spot. Throughout the duration of the study no child commented on the voice texture, and all of them seemed comfortable.

6.5.2 Interface

In terms of the form that the agent could take, the need was for a character that is gender-neutral, engaging and likeable. A puppy character was employed to be the agent [37]. He was named "Spot", and will be referred to by that name henceforth.

Previous research also suggested that such agents should have lifelike characteristics [60]. To create engaging videos with minimal effort, we used Machinima from the SIMS Pets game. SIMS is a widely used "god" game that supports high-level control of characters, including non-human characters. Machinima is the process of using in-game recording facilities to record segments of game action under high-level control of the player [44]. A puppy character was created in the SIMS create-a-pet tool and machinima videos were recorded (with the inbuilt video capture in SIMS) with various different personality traits set. In the create-a-pet tool, depending on what personality is chosen the pet character responds through gestures. This was ideal for the study, as the character was supposed to respond to child's questions, not just with speech but also through gestures. The personality traits that SIMS create-a-pet tool offers are many and the characters can do many gestures, but videos were only recorded for when the puppy character was: nodding his head, shaking his head, jumping, turning shy, sitting down and licking his foot. This was done because only these gestures were relevant to the question/answering game.

6.5.3 Dialogue

Hart and Risely [23] argue that in particular three features add quality and engagement to such interactions. How we incorporated these features into the script that Spot followed

while talking to children, is discussed in the following subsections.

Discourse Functions

These represent the kind of utterances used by parents. This refers to categorization of utterances in terms of the responses that they can prompt. Hart and Risely [23] argue that there are three levels of prompts that parents use as discourse functions. The first level of prompting is generally a rule that is supposed to be followed (ex. "It's cold, you need to wear a sweater"). The second level is a question (ex. "Can you get a coat?"). The final level is a demand (ex. "Go get your coat"). Therefore to make the child ask a question, we introduced three types of cues into Spot's script: a rule ("if you ask me a question, I will give you the answer"), a question ("can you ask me a question?"), and a demand ("go ahead, ask me a question").

Adjacency Conditions

These represent the relationship between utterances of the speaker and listener. This refers to categorization of sequence of utterances in an interaction. Hart and Risely [23] argue that this consists of initiations, response and floor-holding. Initiations are utterances that start a conversation sequence. Therefore, Spot's script contained initiations that can draw the child's attention if they deviate from the game. This involved saying things like, "Are you still there?", "Do you remember what the two objects were?". It should be noted that these questions are different from the ones mentioned under discourse, in that they are only posed if the child goes off-topic or gets distracted. Responses are utterances produced in reaction to a behavioral demand by the child. Spot's script had a response to any question that was asked, even if it was off-topic. This was deemed necessary to create an engaging experience. Floor-holding is an utterance that helps continue a chain of conversation, without the child taking a turn to speak. Therefore, the instructions on how to play the game were split across multiple audio clips and played after pauses so that it feels like a continued conversation, and doesn't overwhelm the child.

Valence

Valence is the emotional tone given to interactions. It can be both, positive or negative. Hart and Risely [23] argue that this comprises of prohibitions, approvals and repetitions. Prohibitions are utterances that explicitly disapprove a child's words. However, there was a possibility for Spot to be viewed as inappropriate if it explicitly disapproved any child behavior. Therefore, softer disapprovals were included into Spot's script. The script was designed to include things like, "That's not really the right question to ask. Do you want to ask something else?" or "Could you try that again?" Approvals are statements that explicitly approve a child's words. Spot's script also contained affirmative feedback like, "That's right" and "That's great" to encourage further questioning. Repetitions are statements where a part

of child's words are repeated. Spot used these based on the question. Some examples that were included in the script:

- Child: Does it <part/function/property>?
- Spot: Yes it does/No it does not.
- Child: Is it a/an <object>?
- Spot: Yes it is an <object>/ No it is not an <object>.

In terms of the hint mechanism, Spot used the techniques tested in phase 1, if a child found it hard to frame a question. Although children did not deviate from the game frequently during phase 1, any dialogue system is incomplete without certain edge cases or re-directions. Most chat-bots use such re-directions to direct users to topics within the chat-bot's knowledge base. Re-directions in Spot's script were encouraging and/or affirmative:

- "I like playing with you! Let's continue!"
- "I love this game! Want to play more?"
- "This is my favorite game. Let's play my favorite game together!"

Gesture

Spot used a variety of gestures to convey meaning during a game session. These were mappings of the content of the response to specific gestures. This is summarized in Table 6.2. Research suggests that gestures can be of four types [53]: iconic, deictic, metaphoric and beat. Iconic gestures represent object attributes, spatial relationships and actions. Such gestures were used to signal: affirmative or negative response to a question, right or wrong guesses and an invalid question. Spot nodded its head for affirmative response and shook it for a negative response as mentioned in Table 6.2. For an invalid question Spot turned shy and stepped back as shown in Figure 6.4. For a right guess Spot jumped and for a wrong guess shook his head and look away. Deictic gestures involve connecting speech to location of objects, more specifically pointing. Spot used deixis to introduce objects during a trial, as shown in Figure 6.4 (A and B). Spot did not use any metaphoric and beat gestures as they might have been too complex to interpret, given the age of our participants. Moreover, beat gestures are just meant to keep the rhythm of the speech and convey no semantic meaning. In addition to using the aforementioned gestures, Spot also used speech bubbles to grab the child's attention whenever speaking.

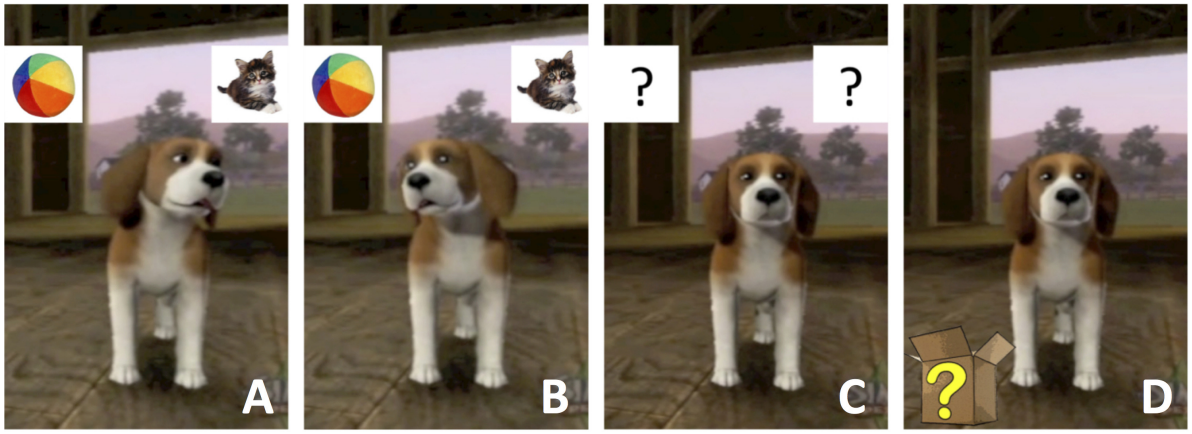

Figure 6.3: A typical game session. Spot first identifies the two objects (A and B), then converts them into question marks (C). After that it hides one object in a box while the other one goes off the screen (D).

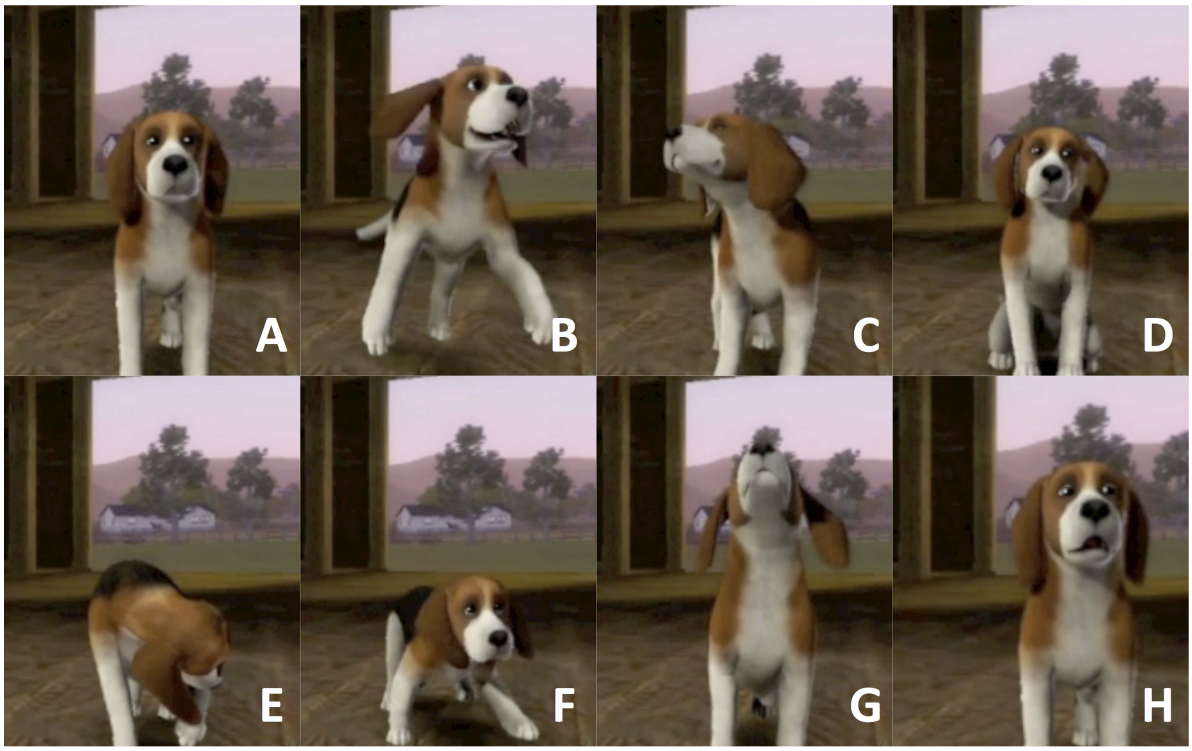

Figure 6.4: Spot's gestures: A) Still, B) Jumping, C) Shaking head, D) Idle, E) Idle, F) Shy, G) Nod, H) Talk

6.6 Phase 2: 20 Questions (Human + Agent Condition)

6.6.1 Participant

Participants were the same as phase 1. Half the participants (10) were assigned to the human condition, and the other half were assigned to the agent condition.

6.6.2 Experiment

Phase 2 of the study followed a between subjects design. Participants from phase 1 were randomly assigned to the human and the agent condition. To make sure that two groups were not different to each other a priori, we conducted individual two-tailed t-tests on the data from phase 1. We tested the two groups for any significant difference in terms of the following parameters: total questions on property, total questions on part, total questions on function, total questions overall, total explanations required, total off-topic count, total hint, total questions successfully solved. None of the t-tests suggested significance, so there was no reason to believe that the two selected groups were different from each other in a priori question answering behavior (this was just a sanity check, as randomization should have already ensured this). Moreover, just like phase 1 in a particular session a child went through 7 trials including the demo trial. The object pairs used in the trials for phase 1 and phase 2 were completely different, and are summarized in Table 6.1.

The procedures in the human condition were the same as phase 1. The only change was that the human conducting the session tried to stick to the dialogue script designed for Spot, and deviated only if the child went off-topic or got confused. In simple words, the script was supposed to be overruled only if the conversation needed "repair", despite the strategies used in the script. The adherence to script was done so that the verbal exposure in the two cases (human and agent) is comparable. The deviations were allowed because parents in practice use a lot of strategies to engage children and technology cannot replicate all of those. Phase 2 was therefore designed to compare and contrast Spot's limited, deterministic but organized strategies against that of an actual person. It should be noted that even though the researchers conducting the study were not teachers or parents, they were significantly familiar to the participants because of exposure during circle time (explained above) and also during phase 1.

The procedures in the agent condition were very similar to the human condition, except that Spot was supposed to conduct the entire game session. As mentioned already Spot used a script to go through its dialogues. The part of setting up the game remained the same for all users. In each game session spot introduced itself, explained the rules of the game step-by-step and then went through the trials with different object pairs. In any given trial, Spot would first show the two items, identify both of them, then convert them into question marks. After this, through some animation, one question mark would leave the screen and the other one will go into a box. All of this is depicted in Figure 6.3. Once the object was

hidden, the child was expected to ask questions to figure out the hidden object. After this, the game took different conversational routes for different participants.

6.6.3 Environment and setup

The environment and setup for the human condition was the same, as phase 1 because the premises and the room used remained the same. However, for the agent condition we decided to project Spot on a wall using a projector. This was done because of multiple reasons. Firstly, having the character on a wall made sure that mona lisa effect exists and gets preserved in the room [4]. Secondly, research suggests children this age have higher depth of search for interactions that are less manual (touchscreens) [6]. Thirdly, projector offered the advantage of larger size and form-factor. The layout of the room can be seen in Figure 6.5.

6.6.4 Data collection and analysis

In the same fashion as phase 1, all the video data from the study was transcribed. Again, six individual coders/raters were asked to classify the questions into one of the three categories: parts of objects, properties of objects and functions of objects. There was substantial agreement in the ratings (Fleiss'es (overall) kappa = 0.79, kappa error = 0.0117, kappa C.I. (95%) = 0.7840, z = 67.5126, p = 0.0001). An analysis of the ratings is contained in the next subsection.

6.6.5 Results

Question-answering experience

It was found that Spot was able to engage children in short conversational sequences, sometimes even better than the human condition. Across the 7 trials in the session, children in the human condition asked 78 questions and this number was 123 for the agent condition. A two-tailed t-test for the total number of questions asked, pointed towards statistical significance (p-value = 0.03, Cohen's d = 0.94). Analyzing the individual question categories, it was found that children asked more property (p-value = 0.046, Cohen's d = 0.83) and part questions (p-value = 0.01, Cohen's d = 1.49) in the agent condition. The numbers of questions on functions of objects were not significantly different. It should be noted that children asked more property questions for objects of increased similarity, and asked more part questions for objects that were easy to distinguish. This trend is predictable given prior research [10], and did not differ across conditions. However, children in the agent condition successfully solved more problems than the ones in the human condition (p-value = 0.03, Cohen's d = 1.3). We hypothesize that this is because of increased number of questions in the agent condition, because children can effectively use questions as a means to solve such problems [10]. Children had limited tendency to guess without proper inference, which is consistent with phase 1. There were 40 such guesses in the human condition and 21 in the agent condition. The

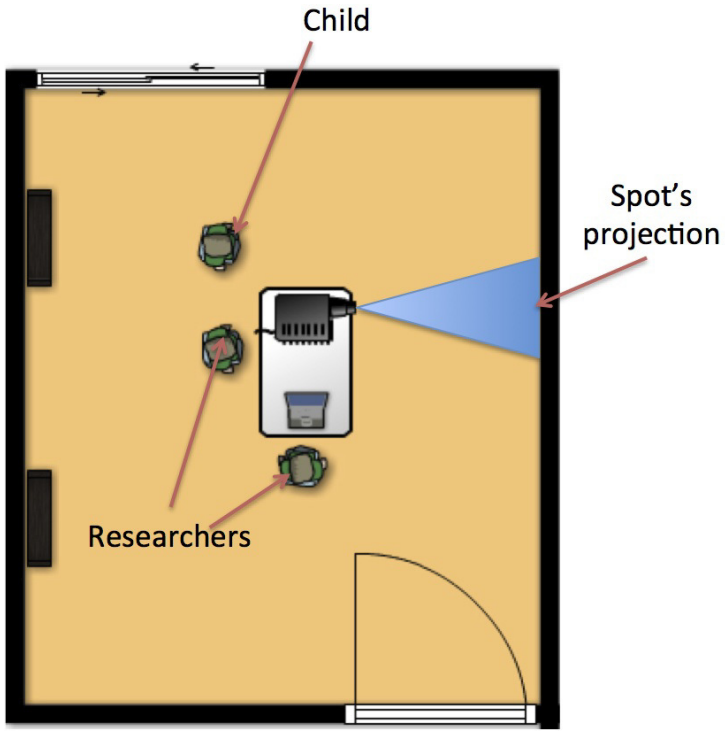

Figure 6.5: The layout of the research room, during the study session

individual numbers of guesses in both conditions were not significantly different from each other (p-value = 0.25). It should be noted that the results from phase 1 were compared to the results from phase 2 for the human group. We did not find any significant difference in performance (across all recorded parameters) of the group assigned to the human condition, in phase 1 and phase 2. The participants did not exhibit any significant learning effect, which is reasonable because the object pairs used in phase 2 were completely different from phase 1 and the activity at hand was not something that is unfamiliar to preschoolers [10]. Therefore, it is reasonable to assume that any change in the performance of the agent group across the two phases was caused by the introduction of the agent. Since the aim of this research was to design an agent that can produce question-answering experience that is comparable to interacting with a familiar human, these results are highly encouraging. All totals have been plotted in Figure 6.6.

Dialogue flow and quality

Handling re-directions and off-topic conversations is an important characteristic of any question-answering system. Interactions with Spot, just like interactions in phase 1 did not deviate from the topic much. Children generally stayed focused and there was no significant difference

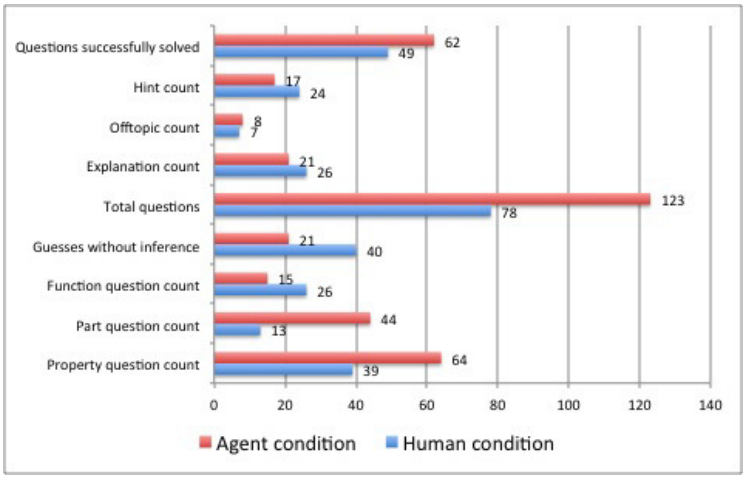

Figure 6.6: Graph with the total counts for all the measured parameters, for the two groups.

in the number of off-topic dialogues in the human and agent condition ($p = 0.813$). Children only went off-topic 8 times in the human condition, and this number was 7 for the agent condition. In all the 7 cases of going off-topic, an edge case handler dialogue (explained above) by Spot was able to bring the child back to the original conversation. In terms of explanations, children in the human condition needed explanation 26 times and those in the agent condition needed it 21 times. Moreover, there was no significant difference in the number of times explanation was required by a child ($p = 0.5$). The total number of hints required by the group assigned to human condition was 24, and that required by the agent condition group was 17. However, there was no significant difference in the number of hints required individually ($p=0.69$). It is clear from these numbers that children follow a similar pattern as phase 1 in talking to Spot, and there was a limited need for explanation, hints and off-topic dialogue handling. Moreover, in all such edge cases Spot does as well as a familiar human, if not worse.

Child's subjective experience

During the session, children appeared to enjoy the interaction with the agent. Some of them said things like "I want to ask more questions", "I really like when he jumps", "Spot is clever, but he lets me win." Three participants in agent condition wanted to go through more trials when the session ended. After the game session, children were interviewed for any immediate experience that they might want to share. All children who were a part of the agent condition, said they enjoyed the session. This observation was the same for human condition. We acknowledge that there might be participant response bias, but the ratings do point towards a positive user experience in the agent condition. It should also be noted that 8 out of the 10 participants in the agent condition said they want to play again. The rest said

they want to play a different game with Spot. Some of these post-game session interview accounts also explain the quantitative results we have seen so far. Two out of the ten children in the agent condition reported that they wanted to know how much Spot knows and therefore asked a lot of questions, even when they were sure of what the hidden object was. Novelty factor could also have played a role in motivating children to ask more questions.

While playing the game, two children said things like "hey, when he says another one, I am doing it!" and "when he says play again, I play again!" Two children demanded playing more games with Spot, asking things like "can he play more games?" Three children said they liked Spot and the fact that he could talk. None of the participants paid attention to the fact that there was a wizard transcribing things that they said. Four children did wonder how the projection on the wall was being executed. During video transcription, we recorded and appropriately tagged any incidents of distraction. In both the human and the agent condition, children had limited tendency to look away or get distracted from the activity at hand. We hypothesize that the activity in itself was engaging to them, and they found it even more engaging when Spot helped them through it.

6.7 Discussion

Given the importance of question-answering in the early years of a child's life, our initial problem was to design a question-answering agent that could help preschool children with short question-answering sequences. Phase 1 of our study was dedicated to studying the feasibility of such a system, given how open-ended dialogues with preschoolers can be. Using 20-questions as an activity that can constrain question-asking behavior, we reached the following conclusions. Firstly, children's questions are predictable and deterministic, when grounded in an activity like 20 questions. Secondly, the repair required in such a dialogue is limited and feasible. Thirdly, it is possible to effectively prod preschoolers to solve problems without disengaging. In phase 2 of the study we designed and implemented a question-answering agent using commonly used parenting styles (in dialogues) and machinima videos. We found that children asked more questions and solved more problems in the 20-questions game with the agent, than if they play with a familiar human. Prior education research [10] suggests that the two observations might not be independent, as children use questions as tools to solve problems. Nevertheless, the results are highly encouraging. Moreover, we found that the tendency to deviate from the task at hand was no more in the agent condition, than the human condition. And even in case of deviations, some standard edge case handlers built into the agent's script were able to take care of the redirection to original dialogue. We also found that children did not need significantly more hints or explanations in the agent condition, than they did in the human condition.

Even though our results were positive and demonstrated strong effects, there were a few limitations to our study. All the findings mentioned in the chapter are restricted to a game of 20-questions and not to open ended question-answering. Open-ended conversational sequences are much harder to handle given the current technological capability. The use of

the wizard-of-Oz for the speech recognition component of the system is another limitation. Recognition errors can add significant complexity to the design of an engaging experience. Moreover, our participants were part of a research preschool and therefore had experience being a part of research studies. That said, all these studies were education or psychology studies and children were not a part of any study that involved technology. The preschool's policies also made sure that all such confounds are avoided. However, some of these limitations are essentially avenues for future investigation in this domain. These are discussed in the next section.

6.8 Future Work

The research presented in this chapter is early work in an area that holds potential. We hope that contributions made in this chapter can help shape the next steps in building question-answering systems (or more generally dialogue systems) for younger children. On the research front, we hope to explore the following few ideas. We think that another solution of the initial envisaged problem could be a "character" toy (instead of a projection) with speech recognition and synthesis capabilities. This could take the form of a plush toy or a favorite TV character. This could be a common internal platform (the electronics, microphone, speaker and batteries) that can be used with various external skins. Future work could compare the individual and relative effectiveness of these solutions against each other, in similar activities. An important next step is to automate the speech recognition. We chose the particular game for this study to be "ASR-friendly". i.e. the dialog is highly contextualized around the objects that the child has seen. The question vocabulary is quite limited, and questions follow a limited number of templates. Furthermore, there is hope that even if ASR fails, when the agent tells the child a correct fact about the hidden object, the child can progress through the game.

We also plan to extend the forms of question-answering that our system can support. We believe that child question-answering can be approached with a database of "frequently-asked questions" (FAQ) on particular topics. While the questions may be somewhat open-ended, we are helped by the fact that children's vocabularies (and their set of known objects and concepts) are limited: about 1000 words for three-year-olds [23].

6.9 Conclusion

Despite the hype around how conversational systems (SIRI and S-voice) can make information more accessible, more rigorous research is clearly called for. This chapter presents an important step towards this goal. We have identified need and opportunities for question-answering based conversational games in the everyday lives of preschool children. Based on these insights, we have investigated how preschoolers ask questions in a constrained environment, and how we can build technology that can handle such questions and keep

children engaged. We have demonstrated that such a system, if carefully designed, can at times perform better than a familiar human in the short-term.

Despite the positive results demonstrated in this chapter, a major challenge of building any such system is to work around speech recognition and natural language understanding errors. These components were handled by a wizard so far, but the next chapter is dedicated to the transition to a fully automated system and also presents a computational evaluation of the same. By treating the data recorded in the sessions so far, as the "gold standard", we evaluate how the automated system would have behaved in similar situations. Rules to work around common errors in speech recognition and language processing, specific to our audience, are also presented.

Chapter 7

Behind Spot: Dialogue Driven Non-linear Machinima

7.1 Introduction

This chapter builds on the work discussed in previous chapters. We move from the wizard-of-Oz Spot to a fully automated version of Spot. Having the prior domain knowledge and experience from the studies so far, helps us in the process. In this chapter we contribute to the following directions:

- We propose a system architecture for dialogue controlled non-linear machinima (a concept discussed previously).
- We prototype Spot using the proposed system architecture.
- We use the wizard-of-Oz study from last chapter as the "gold data" and evaluate our system's performance in view of that data.

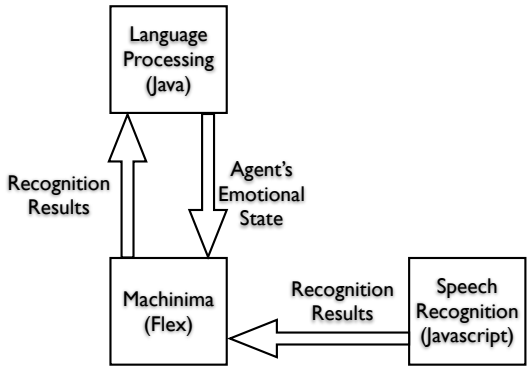

Figure 7.1: System Architecture

7.2 System Architecture

7.2.1 Speech Recognition

This component of the system used Google Web Speech API [68]. The reason for using Google's Speech API and not training our own recognizer was that it is most accurate open-ended speech recognition available [2], and also considerably reduces development time. Also advancing the state-of-the-art for recognition of children's speech is beyond the scope of our work. Moreover, this component was written in Javascript and communicated with the Machinima component of the system using web-sockets. The JavaScript page was hosted by a local Apache server. This component ran the speech recognizer continuously, waiting for long pauses in speech. As soon as a pause was encountered, the recognition results were sent to the Machinima component for processing and response generation.

7.2.2 Non-linear Machinima

The speech recognition component exchanged information with Flex (Actionscript) code. The Flex code was built to run an AIR (Adobe Intergrated Runtime) socket server, and to do script-controlled Machinima. The AIR server was built to send information received from the speech recognition component to the language processing component. The script-controlled machinima code was built to use this information to choose the video to be played. The machinima code was built to run videos in a loop. Each time the video started, the Flex code would check the "emotional state" of the system. This "emotional state" of the system was determined by the language processing component. The "emotional state" is just a deterministic state that the language processing unit generates. For example, keywords like "happy", "sad", "idle", "talking" could be used to denote such a state. This mapping between the "emotional state" generated by the language processing unit, and the machinima video played by the system, can be supplied as an XML file. In terms of audio, the machinima component was built to maintain an array of pre-recorded audio, and read the mapping between video and audio as an XML file. However, for a more open-ended system this could be substituted with a speech synthesizer.

7.2.3 Language Processing

The language processing component was designed to take the incoming transcription of the user's speech, and generate an "emotional state" of the system (defined above). This particular component could be as complex as the developer wants it to be. For example, for a simple question-answering system (as discussed in next section), this component would just generate a "yes" or "no" response. For a fully open question-answering system, this component would simulate a chat-bot. However, most visual chat-bots only perform a few visual acts, therefore our machinima architecture could still be re-used. The entire language processing code was built in Java, and communicated with the machinima component, over a TCP socket connection.

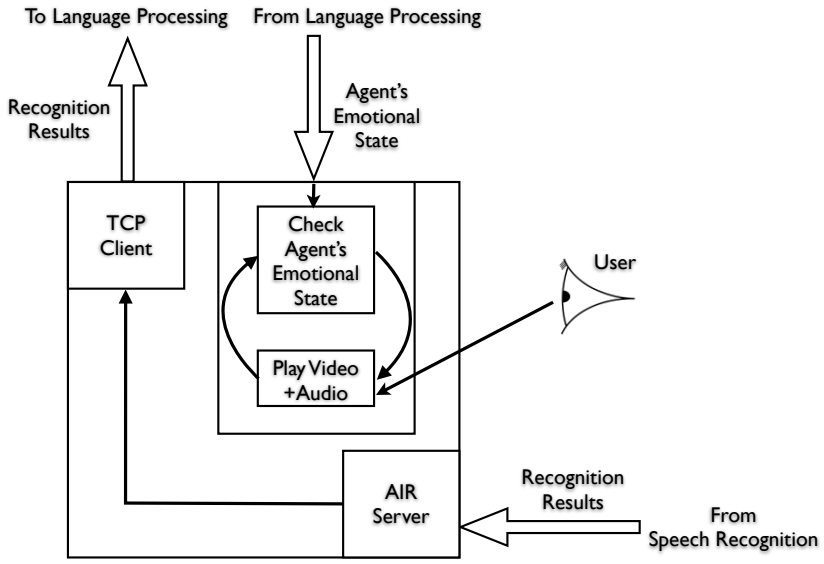

Figure 7.2: Internal architecture of the Machinima component

The language processing also allowed the developer to override the output generated by it, and send a command to the machinima component. This was done with the knowledge that all such automated systems have their restrictions, and sometimes a wizard-of-oz setup is required.

7.3 Question-Answering Agent

Spot and its design has already been discussed in previous chapters, so we will just talk about the language processing components in this chapter.

7.3.1 Question Analysis

Determining whether or not a statement is a question

In a typical case, the verb precedes the subject of the sentence. In the context of the activity at hand, most of the questions are simple in nature. Since the object in the box is unknown, the subject of the sentence is typically referring vaguely to the object as 'it' or to an action that someone can do with or to the object, so the subject is along the lines of 'you' or 'I'. The types of verbs that we see at the beginning of these questions are a small subset of verbs and along the lines of 'can', 'does', or 'is'. Looking for the index of the verb and subject of the sentence allows for a simple comparison of which one occurs first in order to see if it is a question in the general case, with the case where the index of the subject is greater than the index of the verb being a valid question.

However, there are some exceptions to the discussion above. Commands such as 'do this' will be seen as questions with the above technique so the case of 'do this' or 'do that' is checked for and outputs that it is in fact not a question. The five W's (who, what, when, where, and why) also cause complications for the technique for the general case. Although they are considered question words, if the subject follows them immediately then they are in fact sentences. (Example: what we found, was the following) Both of these exceptions are handled but probably unlikely to occur given the context of the game. A child is most likely not going to try to issue verbal commands to the game such as 'do this' and using one of the five W's in a sentence as opposed to a question requires a more complex sentence structure that preschoolers are not experienced in.

Extracting the content of the question

The main content is referring to the meat of the question. It is the property of the object that the question is referring to. To extract the main content, the prepositional phrase is removed if there is one by locating the preposition from a set of common prepositions in questions and removing the content after it. The main content is found by taking all of the words between a start and end index that are not filler words (a, an, the, have), one of the common question verbs, or a negation.

- If one of the five W's occurs before the index of the subject then we start at the index of the W and end at the index of the verb
- In general we start at the index of the subject and end at the end of the sentence without the prepositional phrase

Moreover, when looking at the transcripts of the questions asked while the game was played, questions that contained prepositional phrases referred to a location. This location was typically in reference to where something belonged, like a hat on your head or an elephant in a zoo. Gaining the main content of these phrases was done by finding the preposition and taking the content after that if the word was not a filler, common subject, or a possessive.

7.3.2 Question Answering

Object Properties and Locations

An XML document was created that contained properties of the object, which in this case refers to physical attributes along with actions that can be done to or with an object. A locations category was also made which contains places where the object is commonly found. There is potential to automatically extract this information from online resources like OMCS [74], but outside the scope of this thesis.

Matching Content

Content is considered a match if it is within a Levenshtein distance of the length of the content string divided by three. This is done so that you do not need to enter every form of a word. For example, listing bouncy as a property will also answer the question "does it bounce?" correctly as well even though bounce is not listed. Once a question is asked the content of that question is extracted. If the main content of the question matches one of the properties in the XML document under that object then there are two scenarios that follow:

- There is no prepositional phrase so a positive response is generated.
- There is a prepositional phrase.
    1. If the content of the prepositional phrase matches one of the locations, a positive response is generated.
    2. If the content does not match a location, then a negative response is generated.

If the main question content does not match any of the properties then a negative response is generated.

Handling Negations

When doing the analysis on the question, the number of negations in the question is counted and taken modulo two. If the output of the above is a one then the positive outputs are switched to negative ones and negative outputs are changed to positive ones.

Reveal

If the question is whether the hidden object is one of the two possibilities and the player gets it right, a response of 'reveal' is generated instead of the usual yes/no response.

Dialogue Strategies

A few dialogue strategies were also built into Spot, to recover from some basic problems. For example, in case an incoming utterance is detected as a question, but our system is unable to extract the content out of it, Spot asks the user to repeat the question or ask a different one. On multiple such errors, the user is given feedback that can help remove mis-recognitions. If such error still keeps happening, Spot ends the current trial and moves to the next one. Spot is programmed to return to skipped questions later. Spot has been developed in a way that if there are any loops during dialogue, Spot uses the tricks mentioned above to break the loop and move to other content.

7.4 Performance Evaluation

To evaluate the performance of the system we built, we needed to have a reasonable "gold dataset" that could be used to compare with the responses that our system generated. This exercise is a necessary one before we directly use our system with children, as this would help come up with recovery mechanisms and dialogue repair required. A faulty piece in most dialogue system can be the speech recognition component. Therefore, we use the wizard-of-oz study in which we play the role of the speech recognition and language processing component, from previous chapter. In simple words, instead of these components, a researcher sends direct commands to the machinima component based on what the child (user) says. Obviously, most of the errors in such a system would come from recognition errors or natural language understanding errors, so the wizard-of-oz study was ideal for producing the "gold data". The efficiency of the automated question-answering system could then be seen as a comparison of the "gold data" with the output of the built system. The following subsections explain the results from the performance evaluation.

7.4.1 Computational experiments

Once the data through the wizard-of-oz study had been collected, the next step was to run it by the automated system and measure the performance exhibited. Therefore, the recorded audio was divided into various audio snippets. After this they were converted into FLAC (Free Lossless Audio Compression) format and sent to Google's recognition endpoint that returned JSON objects as recognized hypothesis. Once the recognized transcripts were obtained, they were turned in as inputs to the language processing component and response generated by the system was recorded. If the generated response was the same as the one generated by the researchers during the wizard-of-study, that was considered a perfect response.

Speech recognition errors

A major component of errors in dialogue systems come from speech recognition errors. It is also hard to train a speech recognizer for children's voices. However, with advancement of cloud based speech recognition systems and Google releasing its web speech API, it has become easier for developers to include speech recognition into their applications. As explained before, we used the Google speech recognition system as a part of our system, instead of training our own recognizer. Therefore, it was critical to evaluate the performance of speech recognition alone, on children's voices. Out of a total number of 346 utterances, the speech recognition was unable to produce a correct transcription for only 49 utterances, which is just in 14% cases. In these 14% cases, the average Word Error Rate (WER) was 0.21. This means that even for the mis-recognized transcriptions, the amount of mis-recognition was low. Looking at the errors in more detail, the following were some recurring cases:

1. Some errors were because children used colloquial like "does it meow?" and "does it make oink oink sounds?", while asking questions.

2. Some errors were just language model based errors, like "fur" getting recognized as "for" and "twirly tail" getting recognized as "great detail".

However, in spite of these errors, the performance of the recognition system is quite impressive by any standard, and the error rates beat any error rates reported by a system trained for children's speech.

Natural language understanding errors

The most common reason why the algorithm output incorrect response in these questions was that the property was not listed in the XML file. If a property is not listed, then the algorithm thinks that the item being distinguished cannot do a certain action or that it has a certain characteristic. In addition to that, if a location where an action can be performed is not listed, the response will be incorrect. This can be seen in the case of "Can you pedal it to the library?" where the output is a "no" for a bicycle simply because it does not know that it can be pedaled specifically to a library. There are also cases where the property listed in the XML file results in the wrong answer, such as listing "flow" as a property for kite caused a question asking if it was a flower to be answered as "yes, it does" solely due to the fact that flower and flow are within an acceptable edit distance. Also the context of the question is not captured so asking "Can you eat it?" and "Do you eat at it?" to both be seen as questions about "eat" which will lead to errors such as saying yes to a question about being able to eat a table. Also the algorithm is only able to produce answers to well formed questions and fails on something like "a watch?" which could be a question or a simple statement. There was also a case in the other extreme where the question was too complicated and did not fit the general form that we were looking for. Using "and" and "but" will in general throw off the algorithm at this point in time. Also adding adverbs and uncommon adjectives to properties will throw off the system because they are not listed in the properties XML file word for word.

In various cases, the difference from the actual transcript and what the software produced sounded alike but was out of the acceptable edit distance range so it was not recognized as a valid property of the object. Sounds were commonly misrepresented. By adding a lookup table for commonly mistaken phrases and what they actually should be, we were able to increase the accuracy of the algorithm.

However, in spite of the above restrictions, the question-answering algorithm had good matching accuracy. For a total of 346 utterances generated by the children in our study, the algorithm was able to generate the correct response in 297 cases. Which is a matching accuracy of 86%. In simple words, for 86% cases of the child talking to the system, the system was able to generate a response that the wizard generated during the data collection study. Also, when the lookup table of most common errors was added, the accuracy went up to 89%. If the errors of the speech recognition component were ignored and the question-answering component was given 100% accurate transcripts, the matching error rate was 94%. This means that there was only a drop of 5% accuracy because of speech recognition errors.

7.5 Conclusion

In this chapter we proposed a system architecture that could be used to do dialogue controlled non-linear machinima. In view of the proposed architecture, we built the "brain" behind Spot. The wizard-of-oz study to collect data, helped in optimizing the system for better performance. It should be here that the experiments in this chapter are fairly ideal and we have derived heavily from previous knowledge to build a system that performs well for situations contained in that knowledge. However, the coverage for "out of brain" situations and content remains open to exploration. We do expect a major portion of the system to remain "as is". For new sets of objects, the property XML needs to be modified. Such a process of knowledge creation could rely on crowd-sourcing. It will also be interesting to explore in more detail the engagement and redirection strategies that such system could use to gain a child's attention back after recovering from an error.[1]

7.6 Future Directions

Future advancements to this research could investigate migration to more ubiquitous and mobile scenarios, and also to experiment with language activities built around question-answering, that engage pre-school children on a long term basis. A few ideas are discussed in next few subsections:

7.6.1 Language Games

Games provide an excellent mixture of self-directed activity, exploration, regular rewards, positive affect and competition or cooperation. They can engage children for hours. Cell-phone language games have shown significant improvements in vocabulary for children in India, and the design space is extremely rich i.e. games can be designed for all of the competencies that comprise literacy. It is somewhat more challenging to design games for a screenless setup, but there are still many possibilities, especially for question answering. Researchers can also work around the absence of a screen by using pico-projectors to project in the child's environment. That would help present visual information in a more natural modality, ideal for a game. Games provide further benefits by integrating fine-grained evaluation. That is, the child's competencies are being continually assessed as a natural part of game play. Games provide a variety of feedback and incentives to encourage play and to make it an enjoyable experience.

[1] Experiments pertaining to coverage and long-term effects were being conducted as this thesis was being written, and will appear as followup publications on this work.

7.6.2 Reading Activities

Storytelling from texts where the toy reads from an internal text without a book and interacts with the child about it, is a possible application. While this is not as versatile as shared reading with a book, there is evidence that it should still be very effective for early literacy. First of all, shared reading is valuable to preschoolers long before they are able to decode words. In place of pictures in the book, the read story may contain sound effects that capture similar context to allow the child to remember where they are in the story. Shared book reading in preschool years has a strong correlation with oral literacy in the early years of school, but little correlation with written literacy. While children will eventually learn to decode a written text from a book, they must first learn the structure of language used in written texts, which is quite different from spoken language. Finally, oral storytelling, where a parent recounts stories from their own memory without a text, shows many of the same early literacy benefits as shared reading. In this mode, the child could ask questions while the story is being read thereby clearing doubts and increasing knowledge base. The system would also ask occasional questions, to keep the child involved.

For content, research could leverage the International Children's Digital Library, a large corpus of open-source and licensed children's picture books. By including the text of each story, the character toy can read the story using its own internal copy, asking the child questions about the characters and what is coming, and entertaining questions from the child.

Chapter 8

Conclusion

Challenges in achieving reasonable levels of proficiency in English as a language exist across various different populations and segments in the world. It is also widely known that the level of proficiency in the language of the service economy has effects on levels of poverty. There is no uni-dimensional solution to poverty, however, literacy forms a part of it.

While literacy in English is important, there is also a growing opportunity in using technology to bridge this gap. Prior research has argued and demonstrated merit in the use of technology (especially mobile) towards language learning [35] in developing regions of the world. However, the issues in language learning are fairly ubiquitous in the world. These issues transform from one kind to the other, but don't seize to exist. As mentioned before, for example while teacher absenteeism is a big problem in the developing regions, lack of motivation to speak and practice English seems to be a challenge for children in developed countries [78].

Moreover, in this thesis we explored how conversational and spoken language technologies could contribute to the domain of acquisition of English as a second language. Towards this end, we talked about two major lines of work, one in pronunciation feedback for Hispanic children and the other one being question-answering technology for preschoolers. We will discuss the two separately and then argue for a common conclusion.

8.1 Pronunciation Feedback Technology

We designed and developed games that could do real-time pronunciation monitoring and feedback for children, so as to create an environment of engagement and productive practice, that could have benefits for literacy levels. Through controlled studies we demonstrated value to this approach, and then through linguistic evaluation and optimization, we established grounds for importance of intelligibility in speech. We also reflected these perceptual characteristics in our system, and computationally demonstrated merits to it. Future research in this area, could pick up from where we left of, and choose to evaluate these claims on a longitudinal basis. Another interesting research direction could be to explore other types

of game design and pair with method of teaching, to garner long term retention of course material. However, the work we have done should not be seen as an argument for replacement of classroom instruction. There is significant merit to that, and the studies we have conducted, have focused primarily on out-of-school learning that could add to the existing learning process. The idea of this research has been to create an environment where children feel free to engage in language games without any peer pressure, and thereby learn through productive practice.

8.2 Question Answering Technology

In the second part of the thesis we investigated the context of conversational agents for preschoolers. After some qualitative and quantitative evaluations using the CHILDES database [47], we concluded that context and focus is important to any activity that engages preschool children. Therefore, we designed and developed an agent called Spot, that engages preschoolers in short question-answering game, very similar to 20 questions. In this process, we did a feasibility study to determine the predictability in a child's conversations. Moreover, we also did a wizard-of-Oz study with Spot, where we mimicked the speech recognition and language processing components of the system. Furthermore, we transitioned to an automated system and evaluated it using the data collected during the wizard-of-Oz study. Children found Spot to be engaging and qualitative and quantitative results corroborated the same. At times Spot turned out to be more engaging than a session with a familiar human (researchers). Again, the argument is not in support of replacing traditional teaching methods with automated agents, but to demonstrate merit to exploration of technological solution in the space preschoolers' conversations. Given the unpredictability in speech and the heavy dependency on context that is common in a child's speech, this was a fairly challenging begin with and the feasibility of it was always under question. However, through a user-centered design process and using context knowledge in a constrained activity we were able to create an engaging system that engages on a short-term basis and has potential for future advancement.

8.3 Way Forward

In this dissertation we have explored, designed and developed systems that use spoken-language technology for language learning. Through multiple short-term studies we have demonstrated value to such systems in a variety of contexts and age groups. We hope that this work can not only inspire future research, but could also be used as an argument in favor of technology augmented language learning. However, there is no uni-dimensional solution to a problem like literacy gaps and poverty. The probes we have presented in this dissertation are ones that could potentially be used by future researchers to build on, from a technical and a design standpoint. The real change in state of affairs will come from a wider adoption of

these technologies in combination with classroom teaching. The structure of this combination and the long-term effects of it, remain open for exploration.

Appendix A

Data

A.1 Datasets Used

There are two types of dialogs that we are interested in: free-interaction between adults and children and specific activities between parents and children done in a laboratory setting. Free-interaction dialogs are the window into analyzing what children naturally care about in their everyday world, how adults react to their children's questions, and what interactions naturally engage children the most. Conversely, the motivation behind this analysis research is to discover what types of activities engage children the most and best place them in an inquisitive mindset. Furthermore, these dialogs tell us what type of questions children ask when actively engaged in specific activities. We select datasets fitting these descriptions from the CHILDES database and describe them below [48],[49].

A.2 Free-interaction Datasets

For free-interaction, we choose a variety of datasets from CHILDES where the children were recorded between the ages 3 -5, and were recorded in either normal family interactions or interactions with family members engaging the child in a natural way. We use 5 different corpuses of transcripts for free interaction: Brown, Kuczaj, Macwhinney, Sachs, and Warren [5], [39], [47], [66], [23]. In the Brown corpus, we use the longitudinal study of two children, Adam (2;3 : 4;10) and Sarah (2;3 : 5;1), who were observed by Brown and his students between 1962 and 1966 on a bi-weekly basis. Adam's father was a minister and elementary school teacher. He came from an African American family, but their family spoke standard American instead of African American Vernacular English (AAVE). Sarah was from a working class family. In the Kuczaj corpus, we use the longitudinal study of the child Abe (2;3 : 4;1), collected by the boy's father, a psychology professor. Spontaneous speech was recorded twice a week in 30 minute increments until age 4;1 and recorded once a week onward. The Macwhinney corpus is the longitudinal study of two children Ross (0;6 : 8;0) and Mark (0;7 : 5;6), collected by the boys' father, a psychology professor at different intervals. Ross is

approximately 2 years older than Mark and the records are of natural family interactions. We use the longitudinal study of the child Naomi (1;1: 5;1) in the Sachs corpus, collected the girl's mother, a psychology professor. Lastly, in the Warren corpus, we use the studies of 10 children (4;6 : 6;2), who interacted with their mother and father separately 15-30 minutes each, and were recorded in free interaction. The families were White middle-class nonprofessionals, and the sessions took place at home with the child's own toys or books. The parents were instructed to engage the child as naturally as possible to promote conversation, with the restriction that they did not read to each other.

A.3 Laboratory Datasets

We use the Gleason dataset, which includes transcripts of 24 different children- 12 boys and 12 girls-in various activities with their father and mother separately. In the lab, the child and parent engaged in three activities: playing with a toy auto, reading a picture book, and playing store (also referred to as "pretend"). The parents were encouraged to divide the time evenly among the activities, and the activity order and parent order were randomized [52].

Bibliography

[1] Vincent Aleven, Kenneth R Koedinger, and Karen Cross. "Tutoring Answer Explanation Fosters Learning with Understanding Understanding". In: *Proceedings of the 9th International Conference on Artificial Intelligence in Education*. 1999, pp. 199–206.

[2] Michiel Bacchiani et al. "Deploying GOOG-411: Early lessons in data, measurement, and testing". In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE. 2008, pp. 5260–5263.

[3] Amy L Baylor and Yanghee Kim. "Simulating instructional roles through pedagogical agents". In: *International Journal of Artificial Intelligence in Education* 15.2 (2005), pp. 95–115.

[4] E Boyarskaya and H Hecht. "The Mona Lisa effect: Is it confined to the horizontal plane?" In: *Poster presented at ECVP* (2009).

[5] Roger Brown. "A first language: The early stages". In: *London: George Allen* (1973).

[6] Dana L Byrd et al. "Preschoolers don't practice what they preach: Preschoolers' planning performances with manual and spoken response requirements". In: *Journal of Cognition and Development* 5.4 (2004), pp. 427–449.

[7] Maureen A Callanan and Lisa M Oakes. "Preschoolers' questions and parents' explanations: Causal thinking in everyday activity". In: *Cognitive Development* 7.2 (1992), pp. 213–233.

[8] Pew Hispanic Center. *Statistical portrait of Hispanics in the United States, 2007*. Pew Hispanic Center, 2009.

[9] Colin Cherry and Shane Bergsma. "An expectation maximization approach to pronoun resolution". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Association for Computational Linguistics. 2005, pp. 88–95.

[10] Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. "Children's questions: A mechanism for cognitive development". In: *Monographs of the Society for Research in Child Development* (2007), pp. i–129.

[11] Geraldine Clarebout et al. "Animated pedagogical agents: An opportunity to be grasped?" In: *Journal of Educational Multimedia and Hypermedia* 11.3 (2002), pp. 267–286.

[12] Sharon Darling and Laura Westberg. "Family Literacy: Parent Involvement in Children's Acquisition of Reading". In: *The Reading Teacher* 57.8 (2004), pp. 774–776.

[13] Orlando De Pietro and Giovanni Frontera. "TutorBot: An Application AIML-based for Web-Learning". In: *Advanced Technology for Learning* 2.1 (2005), pp. 29–34.

[14] Pierre Dillenbourg and John A Self. "People power: A human-computer collaborative learning system". In: *Intelligent Tutoring Systems*. Springer. 1992, pp. 651–660.

[15] Sanaz Fallahkhair, Lyn Pemberton, and Richard Griffiths. "Dual device user interface design for ubiquitous language learning: Mobile phone and interactive television (iTV)". In: *Wireless and Mobile Technologies in Education, 2005. WMTE 2005. IEEE International Workshop on*. IEEE. 2005, pp. 85–92.

[16] Donghui Feng et al. "An intelligent discussion-bot for answering student queries in threaded discussions". In: *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM. 2006, pp. 171–177.

[17] Brandy N Frazier, Susan A Gelman, and Henry M Wellman. "Preschoolers' search for explanatory information within adult–child conversation". In: *Child development* 80.6 (2009), pp. 1592–1611.

[18] Luke Fryer and Rollo Carpenter. "Bots as Language Learning Tools." In: *Language Learning & Technology* 10.3 (2006), pp. 8–14.

[19] Vincent Goffin et al. "The AT&T Watson speech recognizer". In: *Proceedings of ICASSP*. 2005, pp. 1033–1036.

[20] AC Graesser et al. "Teaching tactics and dialog in AutoTutor". In: *International Journal of Artificial Intelligence in Education* 12.3 (2001), pp. 257–279.

[21] M Grigoriadou, G Tsaganou, and Th Cavoura. "Dialogue-based reflective system for historical text comprehension". In: *Workshop on Learner Modelling for Reflection at Artificial Intelligence in Education*. Vol. 182. Citeseer. 2003.

[22] Aria Haghighi and Dan Klein. "Coreference resolution in a modular, entity-centered model". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 385–393.

[23] Betty Hart and Todd R Risley. *Meaningful differences in the everyday experience of young American children*. ERIC, 1995.

[24] National Institutes of Health et al. *Clear communication: An NIH health literacy initiative*. 2008.

[25] Neil T Heffernan. "Web-based evaluations showing both cognitive and motivational benefits of the Ms. Lindquist tutor". In: *Artificial Intelligence in Education*. 2003, pp. 115–122.

[26] Anna Hjalmarsson, Preben Wik, and Jenny Brusk. "Dealing with DEAL: a dialogue system for conversation training". In: *Proceedings of SigDial*. Citeseer. 2007, pp. 132–135.

[27] J Horowitz et al. *Evaluation of the PBS Ready to Learn cell phone study: Learning letters with Elmo*. 2006.

[28] Adam Janin et al. "The ICSI meeting corpus". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 1. IEEE. 2003, pp. I–364.

[29] W Lewis Johnson. "Serious use of a serious game for language learning". In: *Frontiers in Artificial Intelligence and Applications* 158 (2007), p. 67.

[30] W Lewis Johnson, Hannes Vilhjalmsson, and Stacy Marsella. "Serious games for language learning: How much game, how much AI". In: *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (2005), pp. 306–313.

[31] Matthew Kam et al. "Designing e-learning games for rural children in India: a format for balancing learning with fun". In: *Proceedings of the 7th ACM conference on Designing interactive systems*. ACM. 2008, pp. 58–67.

[32] Matthew Kam et al. "Improving literacy in rural India: Cellphone games in an after-school program". In: *Information and Communication Technologies and Development (ICTD), 2009 International Conference on*. IEEE. 2009, pp. 139–149.

[33] Matthew Kam et al. "Localized iterative design for language learning in underdeveloped regions: the PACE framework". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 1097–1106.

[34] Matthew Kam et al. "Mobile gaming with children in rural India: Contextual factors in the use of game design patterns". In: *Proceedings of 3rd Digital Games Research Association International Conference* (2007).

[35] Matthew Boon Tian Kam, Divya Ramachandran, and John Canny. "Millee: mobile and immersive learning for literacy in emerging economies". PhD thesis. University of California, Berkeley, 2008.

[36] Alice Kerly, Richard Ellis, and Susan Bull. "CALMsystem: a conversational agent for learner modelling". In: *Knowledge-Based Systems* 21.3 (2008), pp. 238–246.

[37] Tomoko Koda and Pattie Maes. "Agents with faces: The effect of personification". In: *Robot and Human Communication, 1996., 5th IEEE International Workshop on*. IEEE. 1996, pp. 189–194.

[38] Nicole C Krämer. "Psychological research on embodied conversational agents: The case of pedagogical agents". In: *Journal of Media Psychology: Theories, Methods, and Applications* 22.2 (2010), pp. 47–51.

[39] Stan A Kuczaj. "The acquisition of regular and irregular past tense forms". In: *Journal of Verbal Learning and Verbal Behavior* 16.5 (1977), pp. 589–600.

[40] Anuj Kumar et al. "An exploratory study of unsupervised mobile learning in rural India". In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 743–752.

[41] Anuj Kumar et al. "Improving Literacy in Developing Countries Using Speech Recognition-Supported Games on Mobile Devices". In: (2012).

[42] William Labov. *Sociolinguistic patterns*. Vol. 4. Philadelphia: University of Pennsylvania Press, 1972.

[43] James C Lester, Brian A Stone, and Gary D Stelling. "Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments". In: *User modeling and user-adapted interaction* 9.1-2 (1999), pp. 1–44.

[44] Henry Lowood. "High-performance play: The making of machinima". In: *Journal of Media Practice* 7.1 (2006), pp. 25–42.

[45] Rosemary Luckin et al. "Children's interactions with interactive toy technology". In: *Journal of Computer Assisted Learning* 19.2 (2003), pp. 165–176.

[46] Hans Peter Luhn. "A statistical approach to mechanized encoding and searching of literary information". In: *IBM Journal of research and development* 1.4 (1957), pp. 309–317.

[47] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk. Transcription, format and programs*. Vol. 1. Lawrence Erlbaum, 2000.

[48] Brian MacWhinney, Catherine Snow, et al. "The child language data exchange system". In: *Journal of child language* 12.2 (1985), pp. 271–296.

[49] Brian MacWhinney, Catherine Snow, et al. "The child language data exchange system: An update". In: *Journal of child language* 17.2 (1990), pp. 457–472.

[50] Kaj Mäkelä et al. "Conducting a Wizard of Oz experiment on a ubiquitous computing system doorman". In: *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*. 2001, pp. 115–119.

[51] Dominic W Massaro. "A computer-animated tutor for spoken and written language learning". In: *Proceedings of the 5th international conference on Multimodal interfaces*. ACM. 2003, pp. 172–175.

[52] Elise F Masur and Jean B Gleason. "Parent–child interaction and the acquisition of lexical information during play." In: *Developmental Psychology* 16.5 (1980), p. 404.

[53] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[54] Suzanne E Mol et al. "Added value of dialogic parent–child book readings: A meta-analysis". In: *Early Education and Development* 19.1 (2008), pp. 7–26.

[55] Jack Mostow. "" 4-Month Evaluation of a Learner-Controlled Reading Tutor that Listens,"". In: *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice, edited by VM Holland and FP Fisher (New York: Routledge, 2008)* (), pp. 201–19.

[56] Jack Mostow et al. "Evaluating tutors that listen: an overview of project LISTEN". In: *Smart machines in education*. MIT Press. 2001, pp. 169–234.

[57] Jack Mostow et al. "Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction". In: *Journal of Educational Computing Research* 29.1 (2003), pp. 61–117.

[58] Kevin P Murphy, Yair Weiss, and Michael I Jordan. "Loopy belief propagation for approximate inference: An empirical study". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 467–475.

[59] Vincent Ng and Claire Cardie. "Improving machine learning approaches to coreference resolution". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2002, pp. 104–111.

[60] Diana Perez-Marin and Ismael Pascual-Nieto. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. Information Science Reference-Imprint of: IGI Publishing, 2011.

[61] Jean Piaget. *Judgment and reasoning in the child*. Taylor & Francis, 1999.

[62] Paul Pimsleur. "A memory schedule". In: *The Modern Language Journal* 51.2 (1967), pp. 73–75.

[63] Robert Poulsen, Peter Hastings, and David Allbritton. "Tutoring bilingual students with an automated reading tutor that listens". In: *Journal of Educational Computing Research* 36.2 (2007), pp. 191–221.

[64] David Powers et al. "PETA: a pedagogical embodied teaching agent". In: *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*. ACM. 2008, p. 60.

[65] K Reeder et al. "The Role of L1 in Young Multilingual ReadersâĂŹ Success With a Computer-Based Reading Tutor". In: *Talk at the Fifth International Symposium on Bilingualism, Barcelona, Spain*. 2005.

[66] Jacqueline Sachs. "Talking about the there and then: The emergence of displaced reference in parent-child discourse". In: *Children's language* 4 (1983).

[67] Jenny R Saffran, Ann Senghas, and John C Trueswell. "The acquisition of language by children". In: *Proceedings of the National Academy of Sciences* 98.23 (2001), pp. 12874–12875.

[68] Johan Schalkwyk et al. ""Your Word is my Command": Google Search by Voice: A Case Study". In: *Advances in Speech Recognition*. Springer, 2010, pp. 61–90.

[69] Timothy Shanahan and Christopher J Lonigan. "The National Early Literacy Panel A Summary of the Process and the Report". In: *Educational Researcher* 39.4 (2010), pp. 279–285.

[70] Erin Shaw, W Lewis Johnson, and Rajaram Ganeshan. "Pedagogical agents on the web". In: *Proceedings of the third annual conference on Autonomous Agents*. ACM. 1999, pp. 283–290.

[71] Bayan Abu Shawar and Eric Atwell. "Fostering language learner autonomy via adaptive conversation tutors". In: *Proc. of Corpus Linguistic (CLâĂŹ07)* (2007).

[72] Thomas R Shultz and Roslyn Mendelson. "The use of covariation as a principle of causal analysis". In: *Child Development* (1975), pp. 394–399.

[73] Robert S Siegler and Martha Wagner Alibali. *Children's thinking*. Prentice-Hall Englewood Cliffs, NJ, 1991.

[74] Push Singh et al. "Open Mind Common Sense: Knowledge acquisition from the general public". In: *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*. Springer, 2002, pp. 1223–1237.

[75] RJW Sluis et al. "Read-It: five-to-seven-year-old children learn to read in a tabletop environment". In: *Proceedings of the 2004 conference on Interaction design and children: building a community*. ACM. 2004, pp. 73–80.

[76] Larry E Smith and Khalilullah Rafiqzad. "English for cross-cultural communication: The question of intelligibility". In: *Tesol Quarterly* (1979), pp. 371–380.

[77] Jesse Snedeker, Kirsten Thorpe, and John Trueswell. "On choosing the parse with the scene: The role of visual context and verb bias in ambiguity resolution". In: *Proceedings of the 22nd annual conference of the Cognitive Science Society*. 2001.

[78] Anuj Tewari et al. "SPRING: Speech and PRonunciation ImprovemeNt through Games, for Hispanic children". In: *Proc. IEEE/ACM International Conference on Information and Communication Technologies and Development* (2010).

[79] Lev Vygotsky, Eugenia Hanfmann, and Gertruda Vakar. *Thought and language*. MIT press, 2012.

[80] David Wood. "Scaffolding, contingent tutoring, and computer-supported learning". In: *International Journal of Artificial Intelligence in Education* 12.3 (2001), pp. 280–293.

[81] Svetlana Yarosh, Kori M Inkpen, and AJ Brush. "Video playdate: toward free play across distance". In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 1251–1260.

[82] Svetlana Yarosh et al. "Developing a media space for remote synchronous parent-child interaction". In: *Proceedings of the 8th International Conference on Interaction Design and Children*. ACM. 2009, pp. 97–105.

[83] Chen Yu and Dana H Ballard. “A multimodal learning interface for grounding spoken language in sensory perceptions”. In: *ACM Transactions on Applied Perception (TAP)* 1.1 (2004), pp. 57–80.