What did Spot Hide? A Question-Answering Game for Preschool Children

Anuj Tewari[1,2] & John Canny[1]
[1]Computer Science Division and Berkeley Institute of Design,
University of California, Berkeley, USA
[2]GE Global Research, 2623 Camino Ramon, San Ramon, California, USA
{anuj, jfc}@cs.berkeley.edu

ABSTRACT
Early literacy is critical to child development, and determines a child's later educational and life opportunities. Moreover, preschool children are incessantly inquisitive, and will readily engage in question answering and asking activities if given the opportunity. We argue here that question asking/answering technologies can play a major role in early literacy. We describe the design and evaluation of a conversational agent called Spot, with the goal of engaging children in a 20-questions game. Towards this goal, we conducted a feasibility study to determine if children's questions are "on-topic" and suitable for ASR/dialogue systems. We evaluated Spot's performance at conducting a game of 20-questions against that of a human partner.

Author Keywords
Games; Preschool literacy; question-answering; conversational agents.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Many national organizations recognize the essential role of early literacy in a child's later educational and life opportunities [28], [29], [30]. "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." Hart and Risley [18] note that early literacy is an enormous challenge and will require lengthy and regular language experiences for the child.

The greatest impact on child literacy will arguably come from intervention at pre-school ages [18]. Moreover, 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 [8]. These questions are an essential part of language development: they provide primary experience with question construction, statement construction, explanation construction, complex tenses etc. 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 [13]. They will often be met with additional requests for more information, or a universal "why." Children have strong preferences for the form of the answer (a causal explanation vs. fact-based answer), although less so for content. Question-asking, not surprisingly, goes beyond literacy and is an integral part of children's cognitive development [7]. So attention to this activity may have benefits in cognitive development as well.

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, this linguistic engagement can come in the 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.

ENVISAGED SOLUTION
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 that in scenarios that lack context [36]. The virtual character (usually an animal),

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

http://dx.doi.org/10.1145/2556288.2557205

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 1.

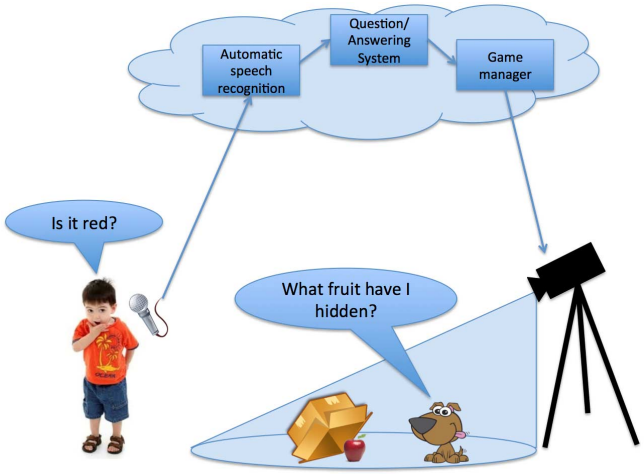

Figure 1: The envisaged solution

CURRENT WORK

In this paper we describe a two-phase study, one phase using a human language partner, and the second using a system which approximates figure 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 12 children studying at the same preschool, playing a 20-questions game with a familiar researcher. It contributed to answering the following research questions:

1. Are children's questions predictable and deterministic, when grounded in an activity like 20 questions?
2. Is the repair required in such a dialogue limited and feasible?
3. 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 (wizard-of-Oz) Spot, an agent that we designed and implemented. Effectively, 6 children played just with the familiar human in both the phases and the remaining 6 first played with the human and then the agent. In phase 3, all the participants played with (non-wizard, fully automated) Spot. Phases 2 and 3 built on phase 1, and answered the following research question:

1. 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. Moreover, literature suggests that vocabulary accrues from language use [18]. It is not feasible to measure vocabulary richness, especially in constrained activities like question answering. Therefore, in this round of research priority was given to the volume of language use. We see the following contributions for this work:

- Testing the feasibility of question-asking/answering behavior in preschoolers, and if technology can be designed to support such behavior.
- Investigation of engaging preschoolers in context of language usage, using technology.
- Using common and effective parenting styles to design and develop a system for preschoolers.

RELATED WORK

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. [16], [19], [2]), question-answering (e.g. [10], [12], [34]), conversation practice for language learners, (e.g. [14], [1]), pedagogical agents and learning companions (e.g. [23], [33], [4], [11]), and dialogues to promote reflection and metacognitive skills (e.g. [17], [20]). 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) [37]. The project CACHET examines the responses young children have to interactive conversational agents using electronic stuffed toys [25]. 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 [25]. 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 [25]. This technological adroitness suggests high potential for interactive agents for younger children.

There is also some recent reflective work in media psychology [22] and education [9] that critically analyses the results of experiments with pedagogical agents in general. Most such work calls for testing with younger audiences. Moreover, prior research has shown that projection systems can be effective mediums of interaction for younger children. This could either be for remote collaboration with other children [39] or communication via media with parents [38]. Overall, we did not find examples in the literature of systems that use projections of virtual agents to do question-answering based games for preschool children.

PHASE 1: FEASIBILITY STUDY

Participants

12 children (6 boys, 6 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 [8]. The preschool that was chosen as the location for the study was a research preschool.

Equipment and setup

The study was conducted in a research room on the preschool premises, reserved for that purpose. The room was equipped with a 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.

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 session. 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 was 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 given in Table 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.

Target item	Low similarity	Moderate similarity	High Similarity
Phase 1	Book & Banana	Table & Bed	Apple & Orange
	Elephant & Spoon	Shoe & Hat	Bicycle & Car
Phase 2	Chair & Rose	Bear & Dog	Truck & Bus
	Flower & Kite	Chicken & Pig	Clock & Watch

Table 1: Object pairs used in the two phases

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 [8]. 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.

Results

Children asked questions

Participants in our study seemed to be naturally primed to ask questions. They asked a total of 127 questions, which means an average of 11 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.

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 80% of the 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 40 such inaccurate guesses were recorded across the 12 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 offtopic.

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.

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.

System architecture

The system worked through modules shown in Figure 2. The speech recognition component of the system was wizard-of-Oz. Collecting audio data and corresponding transcriptions while we evaluate our system was the motivation for using wizard-of-Oz method [26]. 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 redirections (explained in detail under "Dialogue"). 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. 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.

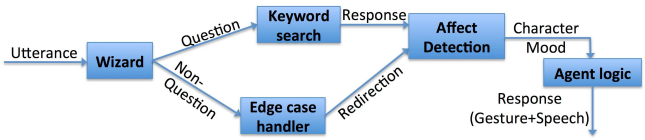

Figure 2: System architecture

Depending on the content of the response (from keyword search and edge case handler), appropriate affect features

were generated and used to predict Spot's gestural response. The audio responses were pre-recorded audio. The voice used to record these responses was that of a female in her early 20s. 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.

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 [21]. He was named "Spot", and will be referred to by that name henceforth.

Previous research also suggested that such agents should have lifelike characteristics [31]. 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 [24]. 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 the child's questions, not just with speech but also through gestures.

Dialogue

Hart and Risely [18] 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 [18] argue that there are three levels of prompts that parents use as discourse functions. 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 [18] argue that this consists of initiations, response and floorholding. 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. Spot's script had a response to any question that was asked, even if it was off-topic. 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 argue that this comprises of prohibitions, approvals and repetitions. 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?" Spot's script also contained affirmative feedback like, "That's right" and "That's great" to encourage further questioning. Spot used repetitions in the context of 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 redirections. Most chatbots use such redirections to direct users to topics within the chatbot's knowledge base. Redirections in Spot's script were encouraging and/or affirmative:

- "I like playing with you! Let's continue!"
- "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. Research suggests that gestures can be of four types [27]: 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 an affirmative response and shook it for a negative response. For an invalid question Spot turned shy and stepped back. For a right guess Spot jumped and for a wrong guess shook his head and looked 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 3 (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.

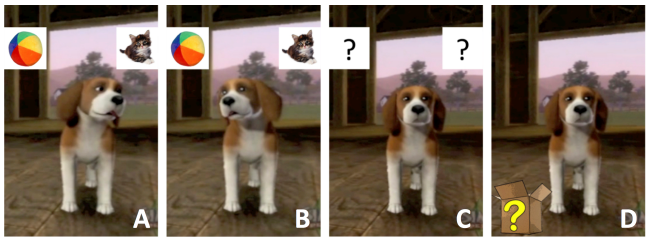

Figure 3: A typical game session. Spot first identifies the two objects (A & B), then converts them into question marks (C). After that it hides one object in a box while the other one goes off the screen (D).

PHASE 2: 20 QUESTIONS (HUMAN + AGENT CONDITION)

Participant

Participants were the same as phase 1. Half the participants (6) were assigned to the human condition, and the other half were assigned to the agent condition.

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 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 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.

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 [5]. 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 4.

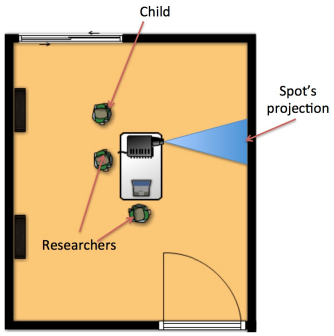

Figure 4: The layout of the research room, during the study session.

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.

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 55 questions and this number was 86 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 [8], 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 the increased number of questions in the agent condition, because children can effectively use questions as a means to solve such problems [8]. Children had limited tendency to guess without proper inference, which is consistent with phase 1. There were 28 such guesses in the human condition and 15 in the agent condition. The 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 [8]. 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 5.

Dialogue flow and quality

Handling redirections and offtopic 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 in the number of offtopic dialogues in the human and agent condition (p = 0.813). Children only went offtopic 6 times in the human condition, and this number was 5 for the agent condition. In all the 5 cases of going offtopic, an edge case handler dialogue (explained above) by Spot was able to bring the child back to the original conversation.

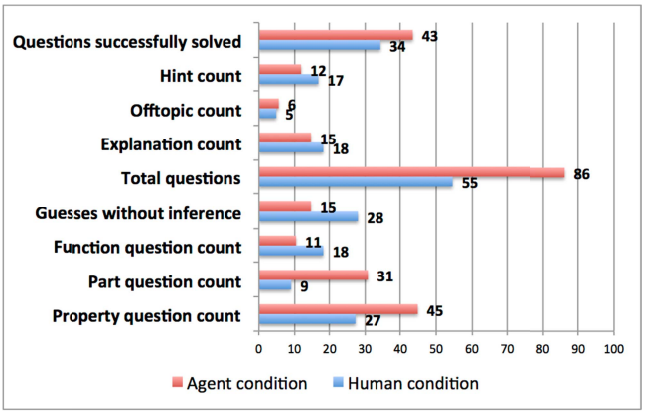

Figure 5: Graph with the total counts for all the measured parameters, for the two groups.

In terms of explanations, children in the human condition needed explanation 18 times and those in the agent condition needed it 15 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 17, and that required by the agent condition group was 12. 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 offtopic dialogue handling. Moreover, in all such edge cases Spot does as well as a familiar human, if not better.

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.

PHASE III: 20-QUESTIONS (FULLY AUTOMATED AGENT ON NEW KNOWLEDGE)

Given that phase 2 had helped understand the engagement that (wizard-of-Oz) Spot offers as compared to a natural interaction, the need was for another study to determine if the fully automated Spot would retain the same level of engagement over a new batch of object pairs (excluding the ones in Table 1).

Participant

Participants, procedures and environment setup were the same as phase 1 (and 2), except with a fully automated agent and new set of object pairs. It should be noted that all the 12 participants were recruited for this phase. The reason we did not do a human condition for this phase was to collect more data, and because we have already established the effectiveness of Spot's interaction against that with a real person. The contribution of phase 3 was to demonstrate that even a fully automated (non-wizard) system would still give similar performance. Moreover, we did not expect any learning effects, as the activity was a familiar one and something that preschoolers do on a daily basis [8]. Given that no learning effect was observed in phase 2 (stated above), there was no reason to believe in its existence for phase 3.

System

The system, as discussed so far was developed in a way that even if Spot doesn't understand the exact utterance of a participant, using edge case handling and hint mechanisms it would still be able to drive the conversation along the context, which in the case of 20-questions is constrained. For this phase, we substituted the wizard component in the system design with Google™ speech recognition. Given the simplicity of children's speech vocabulary, and the large spectrum of voice features our system was expected to work with, this was the most logical choice. Standard techniques like error lookup tables and Levenshtein distances for query matching were used to match incoming transcription to question templates, but discussion on those are outside the scope of this paper. However, such discussions have been reported by some related works [35].

Data collection and analysis

Just as in the case of the previous phases, all the video data from the study was transcribed. We observed substantial agreement in the ratings (Fleiss'es (overall) kappa = 0.85, kappa error = 0.012, kappa C.I. (95%) = 0.7650, z = 65.5126, p = 0.0001).

Results

Question-answering experience

It was observed that the automated Spot had very similar performance as the wizard-of-Oz Spot. If we cross-compared the human condition (phase 2), the agent condition (phase 2) and the automated agent condition (phase 3), against each other, the fully automated Spot's performance was not significantly different from the wizard-of-Oz agent. This seems logical, because Spot was already built to handle mismatches and errors, and Levenshtein distances in query matching, mostly handled any additional errors that speech recognition introduced. For example if the object in context was a cat, and the question asked was "does it meow?", the recognition routine would transcribe it as "does it mean now?". However, because of the use of edit distance, the keyword search routine would assume that the speaker intended to say, "meow". After transcription, it was also observed that 87% of the incoming queries were matched correctly with an intended response. In simple words, in 87% of the cases, the automated Spot was able to produce a response that the wizard would have generated. Moreover, in terms of the measured parameters, there was no significant statistical difference between the performance of the fully automated Spot and the wizard-of-Oz Spot from phase 2. Given that the object pairs were different, this may not be a fair comparison, but the contribution of phase 3 was to replicate performance similar to the agent condition in phase 2, for new object pairs, which clearly happened. To give a visual understanding of the results we have plotted the average values of the parameters for the two conditions from phase 2 and for the automated Spot from phase 3. We plot averages because there were 6 participants in each condition in phase 2 and all of them (12) were involved in the evaluation of the automated Spot in phase 3.

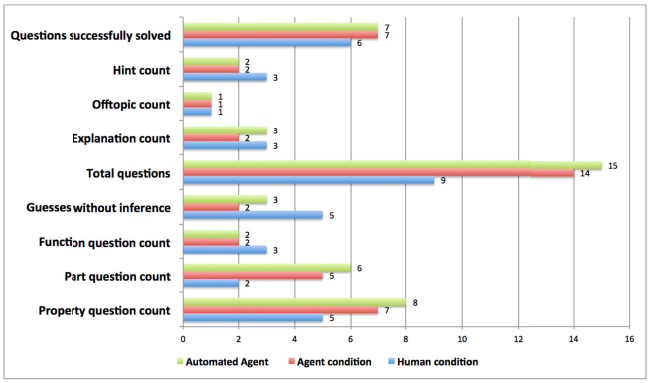

Figure 6: Graph with the averages for all the measured parameters, for the two groups from phase 2, and phase 3.

From Figure 6, it is easy to observe that the fully automated Spot replicates the performance of the wizard Spot from phase 2. And both these conditions are at least at par, if not better than the human condition.

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 and 3 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 [8] suggests that the two observations might not be independent, as children use questions as tools to solve problems. 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 paper 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 current technological capability. Moreover, our participants were part of a research preschool and therefore had experience being a part of research studies. However, some of these limitations are essentially avenues for future investigation in this domain.

FUTURE WORK

We hope that contributions made in this paper can help shape the next steps in building question-answering systems (or more generally dialogue systems) for younger children. 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. 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. This could be extended to other forms of question-answering, in the future.

CONCLUSION

Despite the hype around how conversational systems (SIRI [3] and S-voice [32]) can make information more accessible, more rigorous research is clearly called for. In this paper, 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. We hope that experiences and the challenges we have encountered will be beneficial to our colleagues who seek to conduct more rigorous, longer-term evaluations of such conversational applications under naturalistic conditions in similar environments.

ACKNOWLEDGEMENTS

We would like to thank the Early Child Education Program, the Child Study Centre (Berkeley) in particular for providing the necessary setting to conduct such a research experiment. This work would not have been possible without help from volunteers. We would like to thank Timothy Brown, Samantha Dove and Peiyi Ko, for their help in conducting pieces of the study.

REFERENCES

1. Abu Shawar, B. and E. Atwell, Fostering language learner autonomy via adaptive conversation tutors, in Corpus Linguistics. 2007: Birming- ham, UK.
2. Aleven, V., K. Koedinger, and K. Cross, Tutoring Answer Explanation Fosters Learning with Understanding, in Artificial Intelligence in Education. 1999, IOS Press: Amsterdam. p. 199-206.
3. Apple Siri: http://www.apple.com/ios/siri/. (Last accessed: Sept 19, 2012)
4. Baylor, A.L. and Y. Kim. Simulating Instructional Roles through Pedagogical Agents. International Journal of Artificial Intelligence in Edu- cation, 2005. 15(2): p. 95-115.
5. Boyarskaya E, Hecht H, "The Mona Lisa effect: Is it confined to the horizontal plane?" Perception 38 ECVP, 2009
6. Byrd D., van der Veen T., McNamara J. & Berg W. Preschoolers Don't Practice What They Preach: Preschoolers' Planning Performances With Manual and Spoken Response Requirements, Journal of Cognition and Development, 5:4, 427-449, 2004
7. Callanan M. and Oakes L. Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. Cognitive Development, 7:213-233, 1992.
8. Chouinard, M. M. Childrens questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development, 72 (1, Serial No. 286), 2007.
9. Clarebout, G., Elen, J., Johnson, W.L., and Shaw, E. Animated pedagogical agents. An opportunity to be grasped? Journal of Educational Multimedia and Hypermedia, 11, 267286, 2002.

10. De Pietro, O. and G. Frontera, TutorBot: an application AIML based for Web-Learning. Advanced Technology for Learning, 2005. 2(1): p. 29-34.
11. Dillenbourg, P. and J. Self, People Power: A Human-Computer Collaborative Learning System in Intelligent Tutoring Systems. 1992, Springer-Verlag: Berlin-Heidelberg. p.651-660.
12. Feng, D., E. Shaw, J. Kim, and E. Hovy, An Intelligent Discussion-Bot for Answering Student Queries in Threaded Discussions, in 2006 International Conference on Intelligent User Interfaces. 2006, ACM Press: Sydney, Australia. p. 171-177.
13. Frazier B., Gelman S., and Wellman H.. Preschoolers' search for explanatory information within adult-child conversation. Child Development, 80:1592-1611, 2009.
14. Fryer, L. and Carpenter R., Bots as Language Learning Tools. Language Learning and Technology., 2006. 10(3): p. 8-14.
15. Goffin V., Allauzen C., Bocchieri E., Hakkani-Tr D., Ljolje A., Parthasarathy S., Rahim M., Riccardi G., and Saraclar M., "The AT&T WATSON speech recognizer," in Proceedings of IEEE ICASSP-2005, Philadelphia, PA, USA, 2005.
16. Graesser, A.C., N.K. Person, and D. Harter, Teaching Tactics and Dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 2001. 12: p. 23-39.
17. Grigoriadou, M., Tsaganou G., and Cavoura T., Dialogue-Based Reflective System for Historical Text Comprehension, in Workshop on Learner Modelling for Reflection at Artificial Intelligence in Education. 2003.
18. Hart, B. and Risley, T. Meaningful Differences in the Everyday Experience of Young American Children, Paul H. Brookes, 1995.
19. Heffernan, N.T., Web-Based Evaluations Showing both Cognitive and Motivational Benefits of the Ms. Lindquist Tutor, in Artificial Intelli- gence in Education. 2003, IOS Press: Amsterdam. p. 115-122.
20. Kerly, A., Ellis, R. and Bull, S. CALMsystem: A Conversational Agent for Learner Modelling. Knowledge Based Systems, 2008. 21(3): p. 238-246.
21. Koda T. and Maes. P. Agents with faces: The effect of personification. In the 5th IEEE international workshop on Robot and Human Communication, 1996, pp. 189-194.
22. Kramer, N.C: Psychological Research on Embodied Conversational Agents: The Case of Pedagogical Agents. Journal of Media Psychology 2010. Vol 22(2): p. 47-51
23. Lester, J.C., B.A. Stone, and G.D. Stelling, Lifelike Pedagogical Agents for Mixed- Initiative Problem Solving in Constructivist Learning Environments. User Modeling and User-Adapted Interaction, 1999. 9: p. 1-44.
24. Lowood, H. High-performance play: The making of machinima. In A. Clarke & G. Mitchell (Eds.), Videogames and art: Intersections and interactions. London: Intellect.
25. Luckin R., Connolly D., Plowman L., and Airey S. With a little help from my friends: children's interactions with interactive toy technology. Journal of Computer Assisted Learning, Special issue on Children and Technology:165-176, 2003.
26. Makela K., Salonen E. and Raisamo R. 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, pages 115–119, 2001.
27. McNeil D.l, Hand and Mind: The University of Chicago Press, Chicago IL, 1992.
28. NELP: National Early Literacy Panel (2008) Developing Early Literacy, National Institute for Literacy (NIFL), available from site www.nifl.gov
29. NFCL: National Family Literacy Organization, main site www.famlit.org
30. NIH: Clear Communication: An NIH Health Literacy Initiative http://www.nih.gov/clearcommunication/healthliteracy.htm
31. Pérez-Marín, D., Pascual-Nieto, I.: Conversational Agents and Natural Language Interaction. Techniques and Effective Practices, IGI Global, 2011
32. S-voice: http://www.samsung.com/global/galaxys3/ (Last accessed: Sept 19, 2012)
33. Shaw, E., W.L. Johnson, and R. Ganeshan, Pedagogical Agents on the Web, in International Conference on Autonomous Agents. 1999, ACM Press: Seattle, WA, USA. p. 283-290.
34. Taylor, K. and Moore, S., Adding Question Answering to an E-Tutor for Programming Languages, in AI-2006, 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. 2006, Springer: Cambridge, UK. p. 193-206.
35. Tewari, A., Brown, T., & Canny, J. (2013, January). A Question-answering Agent Using Speech Driven Non-linear Machinima. In *Intelligent Virtual Agents*(pp. 129-138). Springer Berlin Heidelberg.
36. Watson, AT&T: http://www.research.att.com/projects/WATSON. (Last accessed: Sept 19,2012)
37. Wood, D. Scaffolding, contingent tutoring and computer-supported learning. International Journal of Artificial Intelligence in Education, 12:280-292, 2001.
38. Yarosh S., Cuzzort S., Muller H., and Abowd G. 2009. Developing a media space for remote synchronous parent-child interaction. In Proceedings of the 8th International Conference on Interaction Design and Children (IDC '09). ACM, New York, NY, USA, 97-105
39. Yarosh, S, Inkpen, K.M., Brush, A.J. "Video playdate: toward free play across distance." Proc. CHI 2010, 1251-1260.