My Science Tutor: A Conversational Multi-Media Virtual Tutor for Elementary School Science

W. WARD, R. COLE, D. BOLANOS, C. BUCHENROTH-MARTIN, E. SVIRSKY
Boulder Language Technologies, Boulder, CO
AND
S. VAN VUUREN, T. WESTON, J. ZHENG
University of Colorado, Boulder, CO

Submitted for the Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction Applications

This paper describes My Science Tutor (MyST), an intelligent tutoring system designed to improve science learning by students in 3rd, 4th, and 5th grades (7 to 11 years old) through conversational dialogs with a virtual science tutor. In our study, individual students engaged in spoken dialogs with the virtual tutor Marni during 15 to 20 minute sessions following classroom science investigations to discuss and extend concepts embedded in the investigations. The spoken dialogs in MyST are designed to scaffold learning by presenting open-ended questions accompanied by illustrations or animations related to the classroom investigations and the science concepts being learned. In interacting with students Marni applies some of the principles of Questioning the Author, a proven approach to classroom conversations, to challenge students to think about and integrate new concepts with prior knowledge to construct enriched mental models that can be used to explain and predict scientific phenomena. In this article, we describe how spoken dialogs using Automatic Speech Recognition (ASR) and natural language processing were developed to stimulate students’ thinking, reasoning and self-explanations, leading to an improved understanding of science. We describe the MyST system architecture and Wizard of Oz procedure that was used to collect data from tutorial sessions with elementary school students. Using data collected using a WOZ procedure, we present evaluations of the ASR and semantic parsing components, survey results of teachers’ and children’s impressions of MyST.

Categories and Subject Descriptors: I 2.7 [Artificial Intelligence] -- Virtual Tutors - Spoken Dialog Systems - Natural language processing — Speech recognition and synthesis

Additional Key Words and Phrases: Semantic Parsing, Language model, Dialog management, Avatar

1. INTRODUCTION
There is a clear and urgent need to develop accessible and effective learning tools to supplement and improve classroom science instruction for many students in the United States. According to the 2005 National Assessment of Educational Progress (NAEP, 2005) only twenty-nine percent of fourth and eighth grade students in the U.S. performed at or above the proficient level, while only eighteen percent of twelfth graders performed at or above the proficient level. Only three percent of U.S. students attained advanced levels of science achievement in Grades 4 and 8 and only two percent reached advanced levels in Grade 12.

Author Contact Information: W. Ward is Chief Scientist at Boulder Language Technologies and a Research Professor at the University of Colorado, Boulder, CO. E-mail: wward@bltek.com
Acknowledgments - The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070008 and the National Science Foundation Grant DRL 0733323 to Boulder Language Technologies, and DRL 0733322 to the University of Colorado at Boulder. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education or the National Science Foundation.
Since 2007, our research team has been involved in an intensive effort to develop an intelligent tutoring system, My Science Tutor (MyST), to improve science learning by 3rd, 4th and 5th grade children through natural spoken dialogs with Marni, a virtual science tutor. One of the key unanswered questions addressed at the onset of our study was whether state of the art human language technologies could support sustained (15-20 minute) and productive spoken dialogs with a lifelike computer character. MyST requires the integration of automatic speech recognition, robust semantic parsing, dialog modeling and language and speech generation to support conversations with Marni, as well as the integration of multimedia content into the dialogs.

The goal of the MyST project is to help struggling students learn the science concepts encountered in classroom science discussions. Each 15 to 20 minute MyST dialog session functions as an independent learning activity that provides, to the extent possible, the scaffolding required to stimulate students to think, reason and talk about science during spoken dialogs with the virtual tutor Marni. The goal of these multimedia dialogs is to help students think and reason about the science and generate explanations that express their ideas. The dialogs are designed so that over the course of the conversation with Marni, the student is able to reflect on their explanations and refine their ideas in relation to the media they are viewing or interacting with, leading to a deeper understanding of the science they are discussing.

MyST dialogs are linked to the activities, observations and outcomes of classroom science investigations conducted by groups of three to five children. These kit-based science investigations are part of the FOSS (Full Option Science System) program used by over one million students in over one hundred thousand classrooms in all fifty states in the U.S. (FossWeb, 2010). In addition to the science kits that support between 16 and 20 hour-long investigations in each FOSS module (i.e., a specific area of science), the program includes a Teacher Guide, professional development for teachers on how to use the FOSS program to best effect (including helping students organize their predictions, observations and conclusions in science notebooks) a set of science stories that students may read, and valid and reliable standardized Assessments of Science Knowledge (ASK) administered to each student before after each eight to ten week module.

Within a given FOSS module, the initial investigations provide the foundational knowledge for conducting more sophisticated investigations. For example, investigations of magnetism and simple circuits lead to investigations in which children build both serial and parallel circuits, followed by investigations in which they build electromagnets and explore electromagnetism. In our study, a total of 64 different tutorial dialogs were developed to help children explore and explain concepts encountered in the 16 to 20 classroom science investigations in each of the following four areas of science: Variables, Measurement, Water and Magnetism and Electricity.

While there is evidence that student achievement improves in school districts that use high quality inquiry-based science curricula (Valdez, 2001; Klentschy, 2002), there are many students, especially those in low performing schools, who still fail to achieve their potential. In the Boulder Valley School District (BVSD, the site of our study), where elementary school students receive science instruction using the FOSS program, 2006 Colorado standardized (CSAP) science scores of 1,954 fifth graders classified 59.2% as
either proficient (33.5%) or advanced (25.7%). In the top ten scoring schools, the median number of students in the proficient/advanced category was 80% (range: 73-92). In the ten lowest scoring schools, the median numbers of students in the proficient/advanced category was 36% (range: 5-49), with over 60% of students classified as unsatisfactory or partially proficient. BVSD is a diverse school district with many high quality teachers, yet a large number of students still fail to learn science effectively.

Our study is based on the assumption that multimedia dialogs with Marni can play an important and beneficial role in helping students achieve a clearer and deeper understanding of the science they encounter in classroom investigations. Conversations with Marni are characterized by two key features: the inclusion of media, in the form of an illustration, animation or interactive simulation throughout the dialog and the use of open-ended questions related to the phenomena and concepts presented via the media. For example, an initial classroom investigation about magnets has students move around the classroom exploring and writing down what things do and do not stick to their magnets. The subsequent multimedia dialog with Marni begins with an animation that shows a magnet being moved over a set of identifiable objects, which picks up some of the objects but not others. Marni then says: “What’s going on here?” If the student says: “The magnet picked up some of the objects,” Marni might say: “Tell me more about that.” To use another simple example, following a classroom investigation about circuits in which children work together to build a circuit using a battery, wires, a switch and a light bulb, the tutorial begins a picture of the circuit components, with Marni asking: “What’s this all about?”

In the remainder of this article, we present the scientific rationale for MyST, describe the system architecture and technologies that support conversations about science with Marni in multimedia environments, describe data collection of a corpus of tutorial sessions, present results of analysis of data collected in a field trial, and present students and teachers impressions of the program. We also describe the corpus of tutoring sessions that we have developed and present results of experiments using the corpus.

2. SCIENTIFIC RATIONALE

MyST is an example of a new generation of intelligent tutoring systems that facilitate learning through spoken dialogs with a virtual tutor in multimedia activities. Intelligent tutoring systems aim to enhance learning achievement by providing students with individualized instruction similar to that provided by a knowledgeable human tutor. These systems support typed or spoken input with the system presenting prompts and feedback via text, a human voice or an animated pedagogical agent (e.g., Graesser et al., 2001; Wise et al., 2005; Lester et al., 1997; Mostow & Aist, 2001). Text, illustrations, and animations may be incorporated into the dialogs. Research studies show up to one sigma gains (approximately equivalent to an improvement of one letter grade) when comparing performance of high school and college students who use the tutoring systems to students who receive classroom instruction on the same content (Graesser et al., 2001; Van Lehn & Graesser, 2001; Van Lehn et al., 2005).
The development of MyST is informed by several decades of research in psychology and computer science. In the remainder of this section we describe theory and research that informed the design of MyST.

**Social Constructivism:** The work of Jean Piaget, Lev Vygotsky and Jerome Bruner gave rise to a theory of cognitive development and knowledge acquisition known as social constructivism, which provides a strong rationale for the use of tutorial dialogs to optimize learning. In social constructivism, learning is viewed as an active social process of constructing knowledge “that occurs through processes of interaction, negotiation, and collaboration” (Palinsar, 1998). Vygotsky (1970; 1978) stressed the critical role of social interaction within one’s culture in acquiring the social and linguistic tools that are basis of knowledge acquisition. “Learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment” (Vygotsky, 1978, p.90). He stressed the importance of having students learn by presenting problems that enable them to scaffold existing knowledge to acquire new knowledge. Vygotsky introduced the concept of the Zone of Proximal Development, “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers.” (Vygotsky, 1978, p. 85). Bruner (1966; 1990) was influential in extending social constructivism to education and learning. He argued that teachers should provide an instructional context in which students are able to scaffold existing knowledge to construct and test new hypotheses. Social constructivism provides the conceptual model for knowledge acquisition in MyST: to improve learning by scaffolding conversations using media to support hypothesis generation and co-construction of knowledge.

**Discourse Comprehension Theory:** Cognitive learning theorists generally agree that learning occurs most effectively when students are actively engaged in critical thinking and reasoning processes that cause new information to be integrated with prior knowledge. Discourse Comprehension Theory (Kintsch, 1988, 1998) provides a strong theoretical framework for asking questions and designing activities that stimulate thinking and construction of deep knowledge that is useful and transferable. This theory provides the foundation for several instructional approaches to comprehension (King, 1991; Beck et al., 1996; Beck & Mckeown, 2006). Comprehension theory holds that deep learning requires integration of prior knowledge with new information and results in the ability to use this information constructively in new contexts.

**Benefits of Tutorial Instruction:** Theory and research provide strong guidelines for designing effective tutoring dialogs. Over two decades of research have demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. In 1984, Bloom (Bloom, 1984) determined that the difference between the amount and quality of learning for students who received classroom instruction and those who received either one-on-one or small group tutoring was 2 standard deviations. Evidence that tutoring works has been obtained from dozens of well designed research studies, meta-analyses of research studies (e.g., Cohen, Kulik & Kulik, 1982), and positive outcomes obtained in large-scale tutoring programs (e.g., Topping & Whitley, 1990; Madden & Slavin, 1989). Cognitive psychology findings suggest that learning is facilitated in tutoring approaches that require students to express
their beliefs and challenge them to think and reason to explain phenomena (Chi et al., 2001; Van Lehn et al., 2003). Benefits of tutoring can be attributed to several factors, of which the following three appear to contribute most:

1. **Question generation**: A significant body of research shows that learning improves when teachers and students ask deep-level-reasoning questions (Bloom, 1956). Asking authentic questions leads to improved comprehension, learning, and retention of texts and lectures by college students (e.g., Craig et al., 2000; Driscoll et al., 2003; King 1989) and school children (e.g., King, 1994; King et al., 1998; Palincsar & Brown, 1984). Nystrand and Gamarond (1991) found that genuine dialogs, although rare in the classrooms studied, were most often initiated by authentic questions asked by students.

2. **Self explanation**: Research has demonstrated that having students produce explanations improves learning (e.g., King, 1994; King et al., 1998; Palincsar & Brown, 1984; Chi et al., 1989; 2001). In a series of studies, Chi et al. (1989; 2001) found that having college students generate self-explanations of their understanding of physics problems improved learning. Self-explanation also improved learning about the circulatory system by eighth grade students in a controlled experiment (Chi et al., 1994). Hausmann and Van Lehn (2007a) note that: “self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom.” Experiments by Hausmann & Van Lehn (2007b) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

3. **Knowledge co-construction**: Students co-construct knowledge when they are provided the opportunity to express their ideas and to evaluate their thoughts in terms of ideas presented by others. There is compelling evidence that engaging students in meaningful conversations improves learning (Chi et al., 1989; King, 1994; 1998; Palincsar & Brown, 1984; Fine & Messer, 2000; Butcher, 2006; Soter et al., 2008; Murphy et al., 2009). Classroom conversations and tutorial dialogs increase the opportunity for occurrences of knowledge co-construction which has been shown to have a significant impact on learning gains (Wood and Middleton, 1975; King, 1994; Chapin et al., 2003; Chi et al., 2001).

**Benefits of Social Agency and Pedagogical Agents**: When human computer interfaces are consistent with the social conventions that guide our daily interactions with other people, they provide more engaging, satisfying, and effective user experiences (Reeves and Nass, 1996; Nass & Brave, 2005). Such programs foster social agency, enabling users to interact with them the way they interact with people. Programs that incorporate pedagogical agents, represented by talking heads or human voices, especially inspire social agency in interactive media (Atkinson, 2002; Baylor et al., 2003, 2005; Mayer, 2001; Moreno et al., 2001, Nass & Brave, 2005; Reeves and Nass, 1996). In comparisons of programs with and without talking heads or human voices, children learned more and reported more satisfaction using programs that incorporated virtual humans (e.g., Moreno et al., 2001, Atkinson, 2002; Baylor et al., 2005). A number of researchers have observed that children interact with a virtual tutor as if it were a real teacher and appear motivated to work hard to please or impress it; Lester (1997) termed this phenomenon the “Persona Effect.” We have consistently observed this strong social bonding with the virtual tutor Marni in our own research. Interviews with over 250 kindergarten, first and second grade children who interacted with Marni in Foundations to Literacy, a program designed to teach children to read, revealed that over 80% of the
children reported not only that Marni was smart and that she was a good teacher that helped them learn to read, but also that they trusted Marni and that they believed she cared about them. Results presented below indicate that students who used MyST also enjoyed interacting with Marni, that she helped them learn science, and that they were more enthused about science after working with her.

Benefits of Multimedia Presentations: The design of the proposed tutorials is informed by research on multimedia learning conducted by Richard Mayer and his colleagues (See Mayer, 2001 for a review). Mayer and his colleagues investigated students’ ability to learn how things work (motors, brakes, pumps, lightning) when information was presented in different modalities; e.g., text only, narration of the text only, text with illustrations, narrations with sequences of illustrations, or narrated animations. A key finding of Mayer’s work is that simultaneously presenting speech (narration) with visual information (e.g., a sequence of illustrations or an animation) results in the highest retention of information and application of knowledge to new tasks. Mayer argues that in a narrated animation, a student’s auditory and visual modalities are processed independently but are integrated to produce an enriched mental representation.

3. MULTIMEDIA DIALOGS

Students learn science in MyST through natural spoken dialogs with the virtual tutor Marni, a lifelike 3-D computer character that is “on screen” at all times. In general, Marni asks students open-ended questions related to illustrations or animations displayed on the computer screen. The spoken dialog system processes the student’s speech to assess the student’s understanding of the science under discussion, and produces additional actions (e.g., a subsequent question that may be accompanied by a new illustration) designed to stimulate thinking and reasoning that can lead to accurate explanations, as described below. We call these conversations with Marni multimedia dialogs, since students simultaneously listen to and think about Marni’s questions while viewing illustrations and animations or interacting with a simulation.

Marni produces accurate movements of the lips and tongue in synchrony with either recorded or synthetically generated speech. Marni’s visual speech is produced fully automated by the CU Animate system (Cole et al., 2003; Ma et al., 2004) from an input text string and acoustic waveform of the spoken words in the text string. During the initial development and refinement of the MyST system we used high quality text-to-speech (TTS) synthesis rather than recorded speech; since dialogs were constantly evolving, it was far more efficient and cost effective to use text-to-speech synthesis rather than record new utterances each time we changed the dialog. In addition, using TTS allowed human tutors to type in the text they wanted Marni to speak in real time while students were conversing with Marni; this Wizard of Oz procedure used during development of spoken dialogs is described below. At the conclusion of the development phase of each module, a human tutor recorded each of the prompts produced by Marni, enabling her to speak with a human voice that produced appropriate emotional expression, such as enthusiasm when reinforcing the student for accurate and complete explanations. Results of student surveys presented below suggest that students greatly enjoyed interacting with Marni even with her synthetic voice.
Questioning the Author Approach to Tutorial Dialogs

The design of spoken dialogs in MyST is based on a proven approach to classroom discussions called Questioning the Author, or QtA, developed by Isabel Beck and Margaret McKeown (Beck et al., 1996; McKeown & Beck, 1999; McKeown et al., 1999). QtA is a mature, scientifically-based and effective program used by hundreds of teachers across the U.S. It is designed to improve comprehension of narrative or expository texts that are discussed as they are read aloud in the classroom. The program has well established procedures for training teachers to interact with students, for observing teachers in classrooms and for providing feedback to teachers. In recent studies (Murphy & Edwards, 2005; Murphy et al., 2009), QtA was identified as one of two approaches out of the nine examined that is likely to promote high-level thinking and comprehension of text. Relative to control conditions QtA showed effect sizes of .63 on measures of text comprehension and of 2.5 on researcher-developed measures of critical thinking/reasoning (Murphy & Edwards, 2005). Moreover, analysis of QtA discourse showed a relatively high incidence of authentic questions, uptake, and teacher questions that promote high-level thinking—all indicators of productive discussions likely to promote learning and comprehension of text (Nystrand & Gamoran, 1991; Soter & Rudge, 2005; Soter et al., 2008).

Questioning the Author is a deceptively simple approach. Its focus is to have students grapple with and reflect on what an author is trying to say in order to build a representation from it. Because the dialog modeling used in QtA is well understood, can be taught to others (Beck & McKeown, 2006), and has been demonstrated to be effective in improving comprehension of informational texts, we decided to incorporate principles of QtA into tutorial dialogs within MyST. Tutors in our research study, all former science teachers were trained in the QtA approach by one of its inventors, Dr. Margaret McKeown. Following an initial workshop in which the project tutors learned about, discussed and practiced QtA dialogs, Dr. McKeown reviewed transcriptions of tutoring sessions and provided constructive feedback to the project tutors throughout the development phase of the project. The tutorial dialogs in the final MyST system evolved from the iterative process of testing and refining these QtA-based multimedia dialogs.

We note that, in the context of an inquiry-based science program, the perspective of the “author” in “Questioning the Author” moves from questions about what a specific author is trying to communicate, to questions about science investigations and outcomes. In a sense, in a science investigation the "author" is Mother Nature, and the “texts” are the observations that students make and the data sets they enter into their science notebooks. During multimedia dialogs, students are able to review, recall, revisit and revise their ideas about the investigation by viewing illustrations and interacting with simulations while producing and evaluating the accuracy of their self explanations during their conversations with Marni.

Use of Media in MyST Dialogs

MyST dialogs typically incorporate one of three types of media: 1) static illustrations, 2) simple animations and 3) interactive investigations. Although they sometimes overlap in the content presented, each media type plays a unique and important role in science learning in MyST dialogs.
Static Illustrations: Static Illustrations are inanimate Flash drawings. We have found that Static Illustrations are a good way to initiate discussions about topics. They provide the student with a visual frame of reference that helps focus the student’s attention and the subsequent discussion on the content of the illustration. For example, each of the Illustrations in Figure 1 can be presented with questions like: “So, what’s going on here?” or “What’s this all about?”

Figure 1: Example Static Illustrations

The sequence of questions presented by the virtual tutor starts with indirect, open-ended questions about the illustration and then moves to increasingly more directed questions contingent on student responses. A series of questions for the first Illustration in Figure 1 might be:

- What are these things all about?
- You mentioned making a circuit. Tell me more about a circuit.
- Great thinking! What’s important about the components in a circuit?
- You said something interesting about components in a circuit having contact points. What are contact points all about?

A visual like the graph in Figure 1 could be very helpful when working with a student that grasps what they are looking at, but not how to interpret it. A QtA inspired sequence about the graph might be:

- T: What do you think this is about?
- S: I think it’s a graph of something.
- T: Good observation. It is a graph of something. Tell me more about the graph.
- S: Umm, I’m not really sure. It has something to do with washers picked up and wraps on an electromagnet, but I can’t tell any more than that.
- T: Great, this is a graph about the number of washers an electromagnet can pick up and how many wraps it has. What happens to the number of washers picked up when the number of wraps changes?
- S: Hmm, I think it, well, I think it doesn’t change? I guess I don’t really know.
- T: Okay, one good way to tackle a graph is to look at the data points on the graph. Here the data points are the green dots. What do you think the first data point, all the way to the left, is telling us?

At any point that the student expresses a grasp of what a graph is, the tutor moves on to the next point.
Simple Animations: Simple Animations are non-interactive Flash animations. Simple Animations can provide additional information and help students visualize concepts that can be difficult to capture in Illustrations. Figure 2 describes several simple animations, such as the flow of electricity in a circuit and creation of a temporary magnet.

In Figure 2a, the direction of the flow of electricity is represented by blue dots moving through the wires and bulb and back to the D-cell. The animations enable questions to elicit explanations about what is being shown. As with other concepts and media, the questions become increasingly specific if the students are not expressing an understanding of the point. The animation can also be used to support dialogs in which the student produces an accurate explanation for the events shown; e.g., “You got it! The electricity is flowing from through the circuit from the negative to the Positive side of the D-Cell.”

Interactive Animations: Interactive Animations (or simulations) allow students to interact directly with the Flash animation through mouse clicks or by using the mouse to move objects on the screen. For example, clicking on the switch in a circuit will open or close the circuit, resulting in a motor running or stopping, or an electromagnet picking up or dropping iron objects (Figure 3). Interactive animations can be used to present relatively simple concepts (e.g., a switch) or to provide students with the opportunity to conduct complete virtual science investigations and graph the results. During multimedia dialogs, as students are interacting with a simulation, the tutor can say things like What could you do to ...? What happens if you ...?
4. DEVELOPING TUTORIAL DIALOGS

Creating natural and effective interactions between Marni and the student is the overarching goal of the development process. It is necessary to design dialogs that 1) engage students in conversations that provide the system with the information needed to identify gaps in knowledge, misconceptions and other learning problems and 2) guide students to arrive at correct understandings and accurate explanations of the scientific processes and principles. A related challenge in tutorial dialogs is to decide when students need to be provided with specific information (e.g., a narrated simulation) in order to provide the foundation or context for further productive dialog. Students sometimes lack sufficient knowledge to produce satisfactory explanations, and must therefore be presented with information that provides a supporting or integrating function for learning. This is the process of scaffolding learning discussed above.

A major challenge of the MyST project was how to design the spoken dialogs and media in a principled way to optimize engagement and learning. To meet this challenge, we developed an iterative approach to dialog design, informed by theory and research on learning, tutoring, and multimedia learning, in which dialogs were designed and refined through a series of design-test-refine cycles. The stages of dialog design proceeded from: a) initial human tutoring sessions using a set of illustrations, b) human tutoring with computer-based illustrations, animations and interactive stimulations, c) Wizard of Oz studies (described below), in which students interacted with Marni independently, while remote human tutors (the Wizards) monitored the session and could take control of the system when needed. At each step of the development process, sessions were recorded, transcribed and analyzed, leading to refinements and subsequent testing.

As noted, the concepts addressed in MyST tutorial sessions are aligned with FOSS science investigations. Each FOSS Module is composed of four investigations, and each investigation consists of a series of four parts. Each of our tutorials is designed to address the key concepts encountered in the individual classroom science investigations for a part of a FOSS investigation. So a FOSS module would have a series of 16 tutorial sessions associated with it (4 investigations of 4 parts each).

**Tutorial Strategy**

Each tutorial session in MyST is designed to cover a few main points (2-4) in a 15 to 20-minute session with a student. The tutorial dialog is designed to get students to articulate concepts and be able to explain processes underlying their thinking. Tutor actions are designed to encourage students to share what they know and help them articulate why they know what they know. For the Virtual Tutor (VT), the goal of a tutorial session is to elicit responses from students that show their understanding of a specific set of points, or more specifically, to entail a set of propositions. The VT attempts to elicit the points by encouraging self-expression from the student. Question the Author (QtA) influences the strategies we use to get students to share what they know. QtA is very effective at getting students to think more deeply about a concept. Two of the strategies that it utilizes that are employed by MyST are marking and revoicing. These two techniques require the ability to identify the student’s dialog content (referred to as marking it) followed by repeating (revoicing) the question back to the student using similar phrasing; e.g., *You*
My Science Tutor

mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?

The interactions for a concept typically begin with open-ended questions about the concept. Further sequences are written in such a way that they proceed from more general open-ended questions (What’s this all about? to more directed open-ended questions Tell me more about the flow of electricity in the circuit). Initially, students are prompted to consider a concept in terms of their recent experiences in class.

Implementing Tutorial Sessions

The behavior of the Virtual Tutor in a dialog with a student, including the presentation of media within dialogs, is controlled by a task file. The task file contains the definition of the task frames to be used by the application. A task frame is a data object that contains all of the information necessary (or at least available) to interact about the frame:

- Frame Elements – as in parser frames.
- Templates for generating responses
- Pattern-Action pairs, called Rules, for generating responses contingent on certain conditions in the context.

By default, Marni will attempt to elicit speech to fill the Frame Elements representing the propositions of a frame. A sequence of interface actions is generated to elicit a response. The set of interface actions used are those supported by the Hub, flash(), movie(), show(), clear(), speak() and synth(). An example action sequence would be flash(Components); synth(Tell me about that.). This sequence would run the Flash file Components and would synthesize the word sequence and have the character speak it. In order to elicit speech to fill a frame element, the developer specifies a list of action sequences for the element. During a session, the Dialog Manager (DM) keeps count of how many times each element has been prompted for and uses the next action sequence in the list. Once it has exhausted the list, it gives the element the value FAIL, and will move on.

The tutorial developer may also specify a set of Rules for the frame. Rules are pattern-action pairs that can be used to generate action sequences conditioned on features of the context. Rule pattern definitions are Boolean expressions based on element values in the context. If the rule evaluates to true, one of the action sequences following it are sent to the interface manager. Like when prompting for an element, the system keeps count of the number of times a rule has been used and uses the next sequence each time. Figure 6 shows an example frame with a rule. The tutor would initially try to elicit information about flow direction by showing an animated Flash file named Flow and having the agent say Tell me about what’s going on here. If the student responded with it goes from plus to minus where the direction of electrical flow reversed, the parse would be [Flow]: [DirFlow].[Origin]: positive [DirFlow].[Destination]: negative

The mapping of plus and minus to the canonical forms positive and negative is done by the parser. When the parse is integrated into context, the rule would fire and the tutor would continue to show the flash animation Flow, and the avatar would say “Tell me again about the flow”.

ACM Trans.
Rules are useful for marking and revoicing what students have said. They are used to mark and encourage students to go forward, question students if they get a relationship incorrect, and reward them when their efforts result in responses that accurately express conceptual understandings.

**Frame: FlowDirection**

<table>
<thead>
<tr>
<th>[Flow]</th>
<th>[DirFlow]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action: flash(Flow); synth(Tell me about what's going on here.)</td>
<td></td>
</tr>
<tr>
<td>Action: synth(What do you notice about the flow?)</td>
<td></td>
</tr>
</tbody>
</table>

[DirFlow].[Origin]

| Action: flash(Flow); synth(which side of the battery is the electricity coming from) |

[DirFlow].[Destination]

| Action: flash(Flow); synth(which side of the battery is the electricity going to) |

**Rules:**

# Got direction backward

- 
  ([DirFlow].[Origin] == “positive”) || ([DirFlow].[Destination] == “negative”)
  
  | Action: flash(Flow); synth(Tell me again about the flow?) |
  | Action: flash(Flow); synth(What direction is it going?) |

Figure 6 – Example Task Frame

The Dialog Manager (DM) uses a stack driven algorithm for flow control. It maintains two frame stacks, 1) *current* – the set of currently active frames and 2) *history* – the set of completed frames. The DM tries to complete the frame on top of the *current* stack. If the frame on top is complete, it is moved to the *history* stack and the new top frame is completed. In attempting to complete a frame, the Rules are checked first. If a rule expression evaluates True and it has not been marked *FAIL*, the next action sequence for the rule is used. If no sequence was generated by checking the Rules, the DM determines the first unfilled frame element that has an associated action sequence. If all required elements are filled, the frame is moved to the *history* stack, and the system attempts to fill the new top frame. The action sequences for both Rules and Frame Elements can cause new frames to be pushed onto the *current* stack, or old frames to be moved off to the *history* stack.

The development of dialogs, as noted above, represented in the *task* files proceeds through the iterative design, test, and revision process. As new data are received from student sessions, they are analyzed for features like: aspects of the flow of the tutoring session; details of the prompt generation; the use and utility of visuals; and the general completion of frames. This information is used to modify task files to streamline prompts, refine rules, and further design graphics and interactive animations to support or clarify concepts and eliminate misconceptions.
Wizard-Of-Oz Interface

Our development strategy is to model spoken dialogs from tutoring sessions of the type we would like to emulate. In order to gather and model data from effective multimedia dialogs of the sort we would like to create, we developed an interface to MyST that allows a human tutor to be inserted into the interaction loop. In this mode, the student interacts with the MyST Virtual Tutor (VT), while the human tutor can monitor the student’s interaction with the system and take control of the system when desired. This type of data collection system is often referred to as a Wizard-Of-Oz system. The WOZ gives a remote human tutor control over the Virtual Tutor system. At each point in a dialog when the system is about to take an action (e.g., have Marni talk; present a new illustration) the action is first shown to the human wizard who may accept or change the action. For all WOZ data collected, sessions were monitored by project tutors (former science teachers) who served as the Wizards. The data from WOZ sessions was used to improve system coverage concepts and to gain insights into MyST dialog behaviors based on intervention by the Wizards.

During the second and third years of the project, students have independently interacted with MyST in their schools, while Wizards at Boulder Language Technologies offices have monitored the tutoring sessions remotely. One project tutor will go to the school to set up the computers, retrieve students from classrooms, bring them to a computer and initiate the session. The Wizard then connects to a student’s MyST session via the internet.

The WOZ interface is a pluggable MyST component. If the Wizard is not connected, MyST sends the output straight to the user. If the Wizard connects to the session, MyST automatically sends actions to the Wizard for approval or revision. If the Wizard disconnects from the session, the system switches automatically to independent mode. The WOZ system supports both independent use by a student and the ability of a human wizard to connect to any given session. Over the course of the data collection, we have observed the expected pattern that Wizards intervene less and less as the tutorial matures during the development process.

Wizard display

Since the WOZ interface connects to the Virtual Tutor over the internet, the wizard can be at a remote site. The wizard can see everything on the student’s computer, and hear what the student is saying, but can only communicate with the student through the MyST WOZ interface. Figure 7 shows the layout of the Wizard display.
The wizard screen displays:

- A screenshot of the Virtual Tutor screen that the student sees
- The action the Virtual Tutor is about to take
- The frame in focus, including all action sequences associated with elements of the frame.
- A list of all frames in the task file for the session
- A set of command buttons
  - stop agent
  - clear screen
  - end session
- An input history list that can be recalled, to see what has been done and to allow cutting and pasting new responses.

When the Virtual Tutor suggests an action, it is displayed in the top-center screen. Wizards can choose to:

- Accept the proposed action
- Select a new action from the current frame
- Switch to a new frame, and have the system generate a new proposed action
- Generate a new response manually by selecting system content and typing in strings for the agent to speak.

The Virtual Tutor system keeps a log of time-stamped events occurring during the session, any wizard generated actions are included in this log. The log records whether the wizard accepts each proposed system action, or how they changed it. Throughout the project, we used WOZ collected data to train speech recognition acoustic and language models, and to develop grammars for parsing. Analysis of log files from WOZ sessions gives insight into problems with tutorials and can lead to development of additional
multi-media resources or modification of the task file to cause the Virtual Tutor to behave more like the wizards.

4. MYST SYSTEM OVERVIEW

Student Interface
The student’s computer shows a full screen window that contains the virtual tutor Marni, a display area for presenting information and a display button that indicates the listening status of the system. The agent’s lips and facial movements are synchronized with her speech, which may be played back from a recording or generated by a speech synthesizer. Some displays are interactive and the student is able to use the mouse to control elements of the display.

When the student is not speaking, the listening status icon says “OFF” and is dimmed. MySt uses what is known as a “Push-and-Hold” paradigm, where the student holds down the space bar while speaking. When the space bar is released, the Listening Status indicator returns to “OFF” and the system responds to the student utterance. Push-and-Hold system work well with children and in environments with background noise. Having the hard indication that the user is talking to the system, as compared to an “open mike”, provides useful constraints for the recognizer. In interviews with students following the tutoring sessions, all students reported that they found holding down the space bar was easy to do. This procedure encouraged students to spend time thinking about their spoken responses (while Marni waited “patiently” in a state of idle animation, with natural head movements and eye blinks) before responding. It is likely that performance of the speech recognizer was also improved by having the interval of speech indicated by the student.

Dialog Interaction
The tutor takes a series of actions and then waits for input from the student. A typical sequence of actions would be to introduce a Flash animation (“Let’s look at this.”), display the animation, and then ask a question (“What’s going on there?”). Depending on the nature of the question and the media, the student may interact with content in the display area, watch a movie, or make passive observations. When ready to speak, the student holds down the space bar. As the student speaks, the audio data is sent to the speech recognition system. When the space bar is released, the single best scoring word string is sent to the parser, which returns a set of semantic parses. The set of parses is sent to the dialog manager which selects a single best parse given the current context, integrates the new information into the context and generates an action sequence given
the new context. The actions are executed and the system again waits for a student response.

Each tutorial dialog is oriented around a set of key concepts that the student is expected to know based on the content, instructional activities and learning objectives of each science investigation in each FOSS module. The development process benefits greatly from the material provided by FOSS, which describes the key concepts in the investigations and identifies the learning objectives. The key points of the dialog are specified as propositions that are realized as semantic frames. The tutor attempts to elicit speech from the student that entails the target propositions. Following Q&A guidelines, a segment begins with an open-ended question that asks the student to relay the major ideas presented in a science investigation. Follow-up queries and media presentations are designed to draw out important elements of the investigation that the student has not included. The follow-up queries are created by taking a relevant part of the student’s response and asking for elaboration, explanation, or connections to other ideas. Thus the follow-ups focus student thinking on the key ideas that have been drawn from the investigation.

Throughout a dialog, the system analyzes utterances produced by a student and maintains a context that represents which points have been addressed by the student, and which have not. In analyzing a student’s answer, the dialog system tests whether the correct entities are filling the correct semantic roles. The dialog manager then generates questions about the missing or erroneous elements to attempt to elicit information about them. The tutor will continue to try to elicit student explanations about an element until the element is filled or the associated prompts are exhausted.

5. MYST SYSTEM ARCHITECTURE
MySt was developed using the BLT Virtual Human Toolkit (VHT). The VHT is a resource for designing and experimenting with multimedia programs that support real-time conversational interaction with virtual humans. The Virtual Human Toolkit provides a general purpose platform, a set of technology modules, and tools for researching and developing conversational systems using natural mixed initiative interaction with users in specific task domains. In mixed-initiative dialogs, either the user or the system can seize the initiative and take control the dialog. The toolkit consists of an integrated set of authoring tools and technologies for developing applications that incorporate virtual humans in applications. It provides authoring tools for presenting and interacting with media (text, images, audio, video and animations), designing and controlling lifelike 3D computer characters, and designing natural spoken dialogs with the virtual agent.

Virtual Human Toolkit

VHT enables system designers to select a set of 3D models, place them on the screen, and control their actions. While a virtual human can be represented as a disembodied voice (i.e., users can converse with a spoken dialog system over a telephone, and voices can be presented in multimedia applications without a talking head), developers can create
multimedia applications that incorporate embodied conversational agents—lifelike 3D
talking heads or full-bodied characters that posture and gesture. Within the toolkit, a set
of ethnically diverse animated agents each produce anatomically correct visual speech
(through movements of the lips, tongue and jaw) synchronized automatically with either
recorded speech (given a text string representing the spoken words) or with synthesized
speech generated by a text-to-speech synthesis program. The CU Animate module
enables authors to produce facial expressions and animation sequences during speech
production, while “listening” to the user, or in response to mouse clicks or other input
modes. Each animated agent can produce accurate facial expressions of six basic
emotions (surprise, joy, sadness, fear, disgust, anger).

The VHT is composed of modules for:
- speech recognition
- speech synthesis
- semantic parsing
- dialog management
- character animation

It also contains a Hub written in Java that implements the application. The organization
of the toolkit is illustrated in Figure 5.

\textit{VHT Hub}

The Hub is a Java program that provides all of the functions necessary to invoke and send
data to all of the modules, manage the user’s input, invoke Flash applications, play media
files and invoke the agent. The Hub timestamps and logs all interactions. The Hub
executes a set of interaction actions requested by a client module consisting of:
- flash(file) – execute the specified Flash file
- movie(file) – play the specified media file
- show(file) – display the specified static file
- clear() – clear the display
- speak(file) – send the pre-recorded file to CUAnimate for the character to
  speak
- synth(word string) – send the specified word string to the TTS then to
  CUAnimate for the character to speak

Any client module that implements the Hub Application Program Interface (API) can
send interaction requests to the Hub. In Figure 5, both the Phoenix Dialog Manager and a
Flash Application are shown sending interaction requests to the Hub. The Dialog
Manager can invoke a Flash application, which can in turn use the Hub services.
The speech recognizer used in the VHT is a large vocabulary continuous speech recognition (LVCSR) system written by Daniel Bolanos (Bolanos et al., this volume), supported jointly by BLT and CU. It uses the general approach of many state-of-the-art speech recognition systems: A Viterbi Beam Search is used to find the optimal mapping of the speech input onto a sequence of words. The score for a word sequence is calculated by interpolating Language Model scores and Acoustic Model scores. The Language Model assigns probabilities to sequences of words using trigrams, where the probability of the next word is conditioned on the two previous words. The Language Models were trained using the CMU-Cambridge LM Toolkit (Clarkson and Rosenfeld, 1997). Acoustic Models are based on Hidden Markov Models using Gaussian Mixture Models (HMM/GMM) to estimate the probabilities of the speech vectors. The system uses disfluency models to match the types of disfluencies found in applications. The recognizer can output word graphs, but MyST currently uses only the single best scoring hypothesis.
Semantic Parser

The Phoenix parser (Ward, 1994) maps the speech recognizer output, or any text, onto a sequence of semantic frames. These frames represent the system’s understanding of an utterance. The type of representation Phoenix uses to extract information from user input is generally referred to as shallow semantics. Shallow semantics represents the entities, events and relations between them important to understanding an utterance. In Phoenix, these are characterized as semantic frames, together with semantic frame elements. An example parse for *Electricity goes from minus to plus* is:

```
Frame: FlowDirection
  [Electricity] (electricity)
  [Flows] (goes)
  [DirFlow].[Origin] Negative (minus)
  [DirFlow].[Dest] Positive (plus)
```

Semantic grammars are used to match word strings against patterns for frame elements. These are Context Free patterns where the NonTerminals are concepts, events and relations important in the domain. Separate grammars are written for each Frame Element (like [DirFlow].[Origin]). In matching Frame Element grammar patterns against the input text, the parser ignores words that do not match any frame element. This allows the system to match expressions relevant to understanding the domain while ignoring extraneous information and disfluencies such as restarts. A Viterbi search is used to find the optimal set of frames and frame elements. The most optimal parse is the one that covers most of the input and is least fragmented. A set of parses of equal score is produced for an ambiguous input. The grammar rules may be written manually or may be trained from an annotated corpus if one is available.

Dialog Manager

The Dialog Manager (DM) controls the system’s dialog interaction with the user and is responsible for:

a. maintaining a context representing the history of the dialog
b. selecting a preferred parse from a set of candidate parses given the context
c. integrating the new parsed input into the context
d. generating a sequence of actions based on the context

The DM also uses the frame representation used by the parser. It also provides a mechanism for developers to specify the behavior of the system. This mechanism will be discussed below in Tutorial Development.

Text-to-Speech Synthesis

A Text-To-Speech synthesizer receives word strings from the natural language generator and synthesizes them into audio waveforms that can be played back to the user. The VHT interfaces to the general-purpose Festival speech synthesis system (Taylor et al., 1998), and to the commercially available Acapela synthesizer.
Use of Spoken Responses

In the tradition of other systems using children’s speech (Mostow & Aist, 1999), MyST does not use the information extracted from students’ responses to grade students, and the system never tells the student that a response is wrong. This is a good strategy for ASR based systems because the recognizer can make mistakes, but it is also the natural strategy used by QtA and that is why this task and interaction style are particularly suited as an ASR application. Student’s utterances are processed and interpreted to help guide the interaction. After each spoken response produced by a student, the system decides whether the current point should be discussed further, whether to present an illustration, animation or investigation accompanied by a prompt, or to move on to another point. In the absence of usable spoken responses, the system will proceed with a default tutorial presentation as specified in the task file for the session. In sessions where the VT is able to accurately recognize and parse student responses, it is able to adapt the tutorial dialog to the individual student. It may move on as soon a student expresses an understanding of a point, and delve more deeply into a discussion of concepts that are not correctly expressed by the student. It may present more background material if the student doesn’t seem to grasp the basic elements under discussion. If the system is unable to elicit student responses that fill any of the semantic roles related to the science concepts in a dialog, it should present a useful tutorial but may stay on individual points longer than necessary, and may not cover all points by the end of the session.

In cases where the system understands the student, it is also able to apply marking and other techniques that use information from the student’s response to be more engaging. The VT does not simply recognize and parrot back keywords spoken by the students. It represents the events and entities in the student’s response, and it also represents the relations expressed between them. The extracted representation is compared to the desired propositions to decide what action to take next.

Using spoken responses in this way provides a robust system interaction. False Negative errors by the system, in which the system misses correct information provided by the student, account for the bulk of Concept errors. In this case, the system simply continues to talk about the same point in a different way rather than moving on. False Accept errors, where the system fills in an element because of a recognition error, are very rare in MyST. When they do occur, the system may move on from a point before it is sufficiently covered. Recapitulations by the system or errors by the student in later frames can catch some of these. Thus, dialogs are designed to use speech understanding to increase efficiency and naturalness of the interaction while minimizing the impact of system errors.

6. CORPUS DEVELOPMENT

One significant product of the MyST project is the development of a corpus of elementary school students interacting with the virtual tutor. The corpus can be used to train and evaluate children’s speech recognition and spoken dialog algorithms. It can also be used to support analyses of the characteristics of children’s speech. We also plan to use the data for modeling tutorial dialogs and study those features which are associated with learning gains.
Data Collection

All data are being collected from sessions at elementary schools in the Boulder Valley School District (BVSD). BVSD is a 27,000-student school district with 34 elementary schools. There is great student diversity across schools, and schools vary from low to high performing on state science tests. During the 2009-2010 school year alone, 211 students in 7 different elementary schools engaged in multimedia dialogs with Marni in Wizard of Oz mode. We administered tutorial dialogs to students in both high performing and low performing schools in order to gauge the potential benefits to a broad range of students. All but 3 Boulder Valley schools teach science using FOSS.

Data are being collected in three basic conditions:
1. HumanTutor – A human tutor conducts a tutorial with a student. The human tutor has access to the visuals and other supplementary materials, but the tutor talks directly with the student and the student does not interact with the virtual tutor.
2. Wizard-Of-Oz – The WOZ interface is used to interact with the student as described earlier. All interactions are saved to a time-stamped log file.
3. Stand-alone Virtual Tutor – Students interact with the MyST system without a wizard being connected. As noted, the data collected during independent sessions are essentially the same as the data collected during WOZ sessions, but there is no wizard present to change the system’s actions. This is the procedure being used to assess the effectiveness of the MyST system in schools.

<table>
<thead>
<tr>
<th>Module</th>
<th>WOZ sessions</th>
<th>Human sessions</th>
<th>Hours of Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetism &amp; Electricity</td>
<td>415</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Measurement</td>
<td>132</td>
<td>331</td>
<td>67</td>
</tr>
<tr>
<td>Variables</td>
<td>234</td>
<td>49</td>
<td>28</td>
</tr>
<tr>
<td>Water</td>
<td>34</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1 – Data collected by Module

Data Processing

Speech Transcription
The speech data is stored in files by student turns, i.e. whatever is said from the time the student pressed the space bar to talk until the bar is released. The speech is sampled at 16KHz, as is typical with microphone speech. The subjects are wearing Sennheiser headsets with noise canceling microphones. The speech data are professionally transcribed at the word level. Disfluencies (false starts, truncated words, filled pauses, etc) are also marked in the transcriptions. Thus far, 136 hours of speech have been transcribed. Table 1 shows the amount of data collected for each module. The Water module was developed last and collection is just beginning. Data collection of stand-alone Virtual Tutor sessions for assessment has also just begun.
Log files
Each Virtual Tutor dialog session produces a log file that contains time-stamped entries for the events that occurred during the dialog. At each point that the student speaks, an entry is written into the log that gives the filename for the associated recorded speech file. After the speech has been transcribed, the transcript is automatically inserted into the log file at the appropriate place. Some additional pieces of information stored in the log file are: speech recognition output, extracted frame elements, current context, frame name and frame element or rule that is generating the VT response, the number of times this frame element or rule has been used, and action sequence generated for the response.

Concept Annotation
The transcript data are annotated to mark the concepts used by the semantic parser. Human annotators highlight word strings in the transcripts and assign the appropriate concept tags. The concept annotations are hierarchical, for example from the positive end would be a [DirFlow].[Origin].[Terminal] concept where the substring positive end refers to a [Terminal] of a battery. This process is essentially finding paraphrases of the ways concepts are referred to. These annotations are used to expand the coverage of the grammar patterns for the parser, to evaluate coverage of the parser, and to provide “gold standard” input for testing other components of the system.

7. COMPONENT EVALUATIONS
Since only a small amount of data has been collected for the Water module, and transcripts for those are not completed, experiments were conducted using data from only 3 modules: Magnetism & Electricity, Measurement and Variables. The data were partitioned by speaker into Training (78% of speakers), Development (11%) and Evaluation (11%) sets. Data from any individual was in only one of the sets. The training set was used to train acoustic models and language models for the speech recognizer and to train grammar patterns for the parser. The development set was used to optimize parameter values such as language model weights. The Evaluation set was used for component level evaluation of the ASR and Parsing components.

Automatic Speech Recognition Performance
The VHT speech decoder uses standard Trigram language models that were trained using the CMU-Cambridge Language Model Toolkit (Clarkson & Rosenfeld, 1997). In creating lexicons and language models, the developer must balance using more specific models created by partitioning the training data into more classes against using less specific models that use fewer classes but have more data to estimate the parameters for each class. We evaluated the performance of the recognizer using three sets of Language Models:

1. Single LM trained by pooling all training data
2. Separate LM for each module
3. Separate LM for each investigation

What we would expect to see is that words related to concepts specific to a topic have a relatively better chance for inclusion in the more specific models, but general word sequences would be less well modeled. The Word Error Rate (WER) for the recognizer
on the Evaluation set is shown in Table 2. The column labeled *Lexicon* is the number of words known to the system. Overall WER is 34.4%. A small but consistent gain is seen for the Module specific LMs but a larger drop is shown by the Investigation specific models. This result implies that there is enough data to train models specific to each module, but not specific to each investigation. The VB data have a considerably higher WER than the others, but its lexicon is not larger. This can be explained by the relative perplexity of the Language Models, which are ME= 62, MS= 62, VB= 81. Perplexity is a measure of the entropy, or predictability, of the model, and is a logarithm of a probability. The lower the perplexity, the more predictable the word sequences. We have observed that Variables is a more abstract topic that is more difficult for young students than the other two topics. This is reflected by less fluent and coherent and thus less predictable speech.

### Concept Accuracy

The behavior of the Virtual Tutor is more dependent on Concept Accuracy (CA) than on Word Error Rate. The only representation that the Dialog Manager has of what the student said is the extracted frame produced by the parser. If two different word strings have the same parser output, then they are equivalent to the Dialog Manager. One way to measure the effect of recognition errors on the system is to look at the accuracy of extraction of Frame Elements. Reference parses were created for each utterance by parsing the transcripts, which represent input with no recognition errors. The speech recognizer output for the utterances was also parsed and Recall and Precision of Frame Elements were calculated compared to the reference parses. The results for Concept Accuracy are shown in Table 3. The format of the cells is Recall/Precision. Recall is the percentage of the reference frame elements that were correctly extracted from the recognizer output. Precision is the percentage of the elements extracted from the recognizer output that were correct. Using a global LM resulted in an overall Word Error Rate of 34.4% with a Recall of 83% and Precision of 86% for Frame Elements. Increasing the specificity of the LM results in an increase in Recall at the expense of a decrease in Precision. This trend can be explained by realizing that more specific LMs tend to increase the likelihood that domain specific content words will be recognized, whether they were actually spoken or not.

<table>
<thead>
<tr>
<th></th>
<th>Global LM</th>
<th>Module Specific</th>
<th>Investigation Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>Lexicon</td>
<td>WER</td>
</tr>
<tr>
<td>ME</td>
<td>33.8</td>
<td>5134</td>
<td>33.5</td>
</tr>
<tr>
<td>MS</td>
<td>32.8</td>
<td>5134</td>
<td>32.5</td>
</tr>
<tr>
<td>VB</td>
<td>40.7</td>
<td>5134</td>
<td>40.1</td>
</tr>
<tr>
<td>Tot</td>
<td>34.4</td>
<td>34.1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2 – Results for Speech Recognition Output

<table>
<thead>
<tr>
<th></th>
<th>Global LM</th>
<th>Module Specific</th>
<th>Investigation Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>CA</td>
<td>WER</td>
</tr>
<tr>
<td>Tot</td>
<td>34.4</td>
<td>.83/.86</td>
<td>34.1</td>
</tr>
</tbody>
</table>

Table 3 – Results for Speech Recognition Output
We chose to use Module-specific Language Modules for the assessment, as these seemed to offer a good compromise where 85% of concepts in student responses are correctly extracted, and 85% of extracted concepts are correct.

Disfluencies
Approximately 2% of the words in the corpus were partial (truncated) words and 8% of the “words” are filled pauses. The system uses disfluency models to deal with this type of input. In addition to a Silence model, the system uses models to detect a number of other non-word speech events:
- BR – Breath Noise
- EM
- HMM
- HUH
- MMM
- UHM

We conducted an investigation to give some information about the performance of the disfluency models used in the system. Including the disfluency models improved overall WER from 37.3% (using only words and Silence model) to 34.4% (using the additional disfluency models). Normally the system strips out disfluencies from the recognition hypothesis. Disfluencies are annotated in the utterance transcripts and are also removed before the two are aligned for scoring. For this test, we generated recognition output for the test set, but did not strip out the disfluencies. Then all disfluency tags in the hypotheses and the transcripts were mapped to a single tag. For this analysis, we don’t really care what the disfluency was, just whether there was one. Strings of disfluency tags were also collapsed to a single occurrence. We then aligned the transcript and hypothesis strings and computed the Substitutions, Insertions and Deletions for disfluency tags. For the parameter settings used (which optimized WER without fillers) the results are:

Total words in reference: 57641
Total disfluencies in reference: 4397
Disfluency in Reference aligned with disfluency in hypothesis: 3999 (91%)
Disfluency in Reference Deleted (aligned with Silence): 221 (5%)
Disfluency in Reference aligned with word in hypothesis: 177 (4%)

Total disfluencies in hypotheses: 18123
Correctly aligned with disfluency in reference: 3999 (22%)
Insertions (not aligned with any word in the reference): 10267 (57%)
Disfluency in hypotheses aligned with word in reference: 3857 (21%)

For fillers in the reference strings, 91% were aligned with fillers in the hypotheses, indicating good Recall accuracy for the filler models. When a filler (disfluency) in the reference string is deleted, that means that it does not align with any word in the hypothesis. This alone is harmless. It is still possible that it generated an error in the word before or after, but it did not get recognized as a separate word. This type of event only occurred for 5% of the reference fillers. When a filler in the reference string is aligned with a word in the hypothesis, it is an Insertion error, the filler was misrecognized...
as a word and generated an insertion in the hypothesis. Insertions can lead to a concept error (decreasing precision) if the word happens to be parsed into a frame element, but it is rare for word insertions to generate parse errors. Only 4% of the fillers in the reference strings generated Insertion errors.

A total of 18123 filler tokens appeared in the hypotheses, with 22% being correctly aligned to fillers in reference strings. Of the hypothesized fillers, 57% were insertions, i.e., they were not aligned with any word in the reference string. Since filler tokens are stripped out before parsing, the insertion of a filler has no harmful effect. It is possible that mapping frames of speech onto the filler model caused an error in the word before or after the filler, but the presence of the filler token causes no problem. When a filler model in the hypothesis is aligned with a word in the reference, it represents a Deletion error from the reference string, i.e., the word was misrecognized as a filler. This was the case for 21% of the fillers in the hypotheses, but compared to the 57641 words in the test set, this represents a deletion rate of about 7%. It is most often the case that words misrecognized as fillers are not content words necessary for correct understanding.

It was clearly beneficial to include filler models in the decoder, as it reduced the overall WER. While the filler alignment results do not show anything definitive, since we don’t see effects on surrounding words, they do give some insight into the effect of the models.

8. STUDENTS’ AND TEACHERS’ IMPRESSIONS OF MYST

A written survey was given to 167 students who used MyST in five elementary schools during the 2009-2010 school year. The schools in which students used MyST varied greatly in terms of the percentage of students who scored proficient or above on the state science test: from 21% proficient or above for the lowest scoring school, to 82% proficient or above in the highest scoring school. The survey had four main questions designed to assess students’ experiences with and impressions about the program: 1) How much did Marni help you with science? 2) How much did you enjoy working with Marni? 3) If you had your choice, when would you talk with Marni? 4) Now that you have worked with Marni, how do you feel about science? In addition, several other questions were included to assess usability issues, such as “Did you understand Marni’s voice?” Figures 10-13 display the distribution of students’ response choices to each question. The histograms are grouped by school, from lowest performing to highest performing. It can be seen that in general students had positive experiences and impressions about the program. Across schools, 50% to 75% of students said they would like to talk with Marni after every science investigation, 60% to 80% said they enjoyed working with Marni “a lot,” and 60% to 90% selected “I am more excited about science” after using the program. Perhaps most interesting, the majority of students in the lowest two performing schools felt that Marni “helped a lot” in learning science (75%, 55%), whereas the majority of students in the higher performing schools responded that Marni “helped some.” Since MyST dialogs are designed to help students learn the science concepts embedded in classroom investigations, MyST should provide the most benefit to students who are having difficulty understanding these concepts. The survey responses produce initial evidence that students who have most to gain from using MyST may have more positive impressions of the program.
Teachers for participating classrooms were also asked to fill out surveys. The results of the survey are shown in Figure 14. Even though students who used MyST left the classroom during tutoring sessions, teacher responses were in general very positive. They commented that students who used the system were more enthusiastic and engaged in classroom activities. Also, that their participation in science investigations and classroom discussions benefitted students who did not use the system. As Figure 14 reveals, teachers indicated that they would like to have all of their students use the system (not just struggling students) and that they would recommend it to other teachers.
9. SUMMARY AND CONCLUSIONS

The development and implementation of the multimedia dialogs described in this article resulted in an intelligent tutoring system that fully engaged children in natural spoken dialogs with a virtual tutor for substantial periods of time. Following a series of tutoring sessions with Marni, the great majority of students reported that they enjoyed spending time working with her, that they felt that Marni helped them learn science, and perhaps most interesting, that they felt more interested in science and more motivated to learn science than they had before using the system. Students in both high performing and low performing schools, the latter including significant populations of English language learners and students from families with low socioeconomic status, reported that Marni was “way cool.” One of the benefits of this shared perception that students whose parents did not sign the consent form allowing their child to work with Marni, often asked their parents to sign the form after learning how much other students enjoyed the experience; this resulted in a welcomed increase in the number of students who participated in the project.

The third, fourth and fifth grade teachers whose students were tutored by Marni were also excited about the program. The teachers noticed that most of their students who used the program increased their participation and contributions during science investigations and
classroom discussions, and this benefited all students, including those who were not being tutored. Teachers reported that they would like to use MyST in the future to tutor all of their students, and that they would recommend the program to other teachers. This is an important outcome for two reasons. First, it suggests that teachers perceive MyST as a tool that can be used to improve student motivation and learning, rather than as a distraction or threat to their position. Second, since MyST is linked to science instruction and aligned with the learning objectives of a widely used and highly respected science program (FOSS) used by millions of children, it appears likely that teachers and school administrators who use FOSS would be willing to adopt the program if further research demonstrates that it improves student learning and achievement as well as improving their attitudes and motivation.

At the time of this writing, three months into the 2010-2011 school year, MyST is being evaluated for its potential to improve student achievement during independent use by children in each of the four areas of science. In the evaluation phase of the project, children in classrooms (whose parents consent to their child being tutored) are randomly assigned to one of two groups: being tutored by Marni, or being tutored in small groups by one of the project tutors trained in QtA who tutored children and served as wizard in the development phase of the project. Gains in science learning will be compared for students in these two groups based on their performance on the ASK assessment administered to each student before after each science module. In addition, the performance of these students on the ASK assessments will be compared to the performance of students in matched classrooms who do not receive either human or computer-based tutoring. Our hypothesis is that students who engage in multimedia dialogs with Marni will produce benefits similar to students who interact with human tutors, since students in both of these treatment conditions will achieve a thorough understanding of the science through the proven benefits of constructing self explanations and engaging in dialogs that result in co- construction of knowledge with an expert tutor (whether virtual or human) that lead to accurate explanations. In addition, consistent with a substantial body of research reviewed above, we expect that students in both treatment groups will significantly outperform students in classrooms who did not receive tutoring.

In terms of the focus of this special issue on child-computer interaction using automatic speech and language processing, our research demonstrates that state of the art spoken dialog and character animation technologies can be integrated with media (illustrations, animations and interactive simulations) to fully engage children as young as seven years of age in natural spoken dialogs for up to 20 minutes, resulting in positive learning experiences and increased motivation to learn science. One of the most important outcomes of the study, in our opinion, is the demonstrated feasibility and potential of speech and language processing technologies to support sustained conversations between a child and a virtual tutor. By combining open-ended questions with media during tutoring sessions the VT provides students with the opportunity to visualize, think and reason about science, to put their thoughts into words through self-explanations, and to reflect on their explanations during the ensuing dialog to revise and refine their ideas and explanations.
In our view, these conversations are greatly aided by the presence of Marni because students perceive her to be a knowledgeable, sensitive and effective tutor. It is clear from observing students interact with Marni, from the way they talk about her, and from their responses to our questionnaire, that students form a strong bond with Marni, and are motivated to perform well and impress her with their answers, just as they would with a sensitive and effective human tutor. The fact that students are able to engage in productive and sustained dialogs with Marni and enjoy the conversation indicates that speech and language processing technologies are capable of supporting natural spoken dialogs with children that can produce great benefits in the future.

Over the past decade, advances in speech and language processing have enabled research and development of a growing number of intelligent tutoring systems that use spoken dialogs to tutor children and adults (e.g., Mostow & Aist, 2001; Rickel & Johnson, 2000; Graesser et al., 2001; Aist & Mostow, 2009; Mostow & Chen, 2009; Chen et al., 2010). These systems have focused mainly on science, reading and language learning. Our literature review indicated that science tutors that incorporate spoken dialogs have been designed for use by university-level students (e.g., Graesser et al., 2001; Littman & Stillman, 2004). Science tutors have been developed for children that incorporate embodied conversational agents (computer character that talk) in multimedia environments (e.g., Lester et al., 1997; 1999; Dede et al., 2010), but these systems do not support natural spoken dialogs between a child and the agent. Spoken dialogs with children have been used successfully to help children learn to read, comprehend text, and to assess an individual’s proficiency in a given language. For example, in Project Listen, Mostow and his colleagues have integrated speech recognition and dialog modeling to improve reading, vocabulary acquisition and text comprehension (Aist & Mostow, 2009; Mostow & Chen, 2009; Chen et al., 2010). Bernstein & Cheng (2007) demonstrated the validity of scores from fully automated tests that use ASR to assess a child’s ability to understand and communicate in English. While spoken dialog systems have been developed for university-level students for science tutoring and for children for reading and language assessment, we have found no examples in the literature of intelligent tutoring systems that support spoken conversational interaction between children and a virtual science tutor. To our knowledge, MyST is unique in this regard.

We believe the ability of MyST to support sustained and productive dialogs with children can be attributed to two major factors: steady advances in speech and language processing technologies over the past twenty five years which are incorporated into the MyST system architecture, and the way that spoken responses are used by the system.

REFERENCES


My Science Tutor

ACM Trans..