My Science Tutor: A Conversational Multimedia Virtual Tutor for Elementary School Science

WAYNE WARD, RONALD COLE, DANIEL BOLAÑOS, CINDY BUCHENROTH-MARTIN,
and EDWARD SVIRSKY, Boulder Language Technologies
SAREL VAN VUUREN, TIMOTHY WESTON, JING ZHENG,
and LEE BECKER, University of Colorado, Boulder

This article 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 engage 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. The focus of the interactions is to elicit self-expression from students. To this end, 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. 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 with the procedure, we present evaluations of the ASR and semantic parsing components. A formal evaluation of learning gains resulting from system use is currently being conducted. This paper presents survey results of teachers' and children's impressions of MyST.

Categories and Subject Descriptors: I 2.7 [**Artificial Intelligence**]: Natural language processing—*Speech recognition and synthesis*

General Terms: Design

Additional Key Words and Phrases: Semantic parsing, language model, dialog management, avatar

ACM Reference Format:
Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Van Vuuren, S., Weston, T., Zheng, J., and Becker, L. 2011. My Science Tutor: A conversational multimedia tutor for elementary school students. ACM Trans. Speech Lang. Process. 7, 4, Article 18 (August 2011), 29 pages.
DOI = 10.1145/1998384.1998392 http://doi.acm.org/10.1145/1998384.1998392

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 through Grants 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.

W. Ward is Chief Scientist and shareholder at Boulder Language Technologies and a research professor at the University of Colorado, Boulder
Author's address: email: wward@bltek.com

DOI 10.1145/1998384.1998392 http://doi.acm.org/10.1145/1998384.1998392

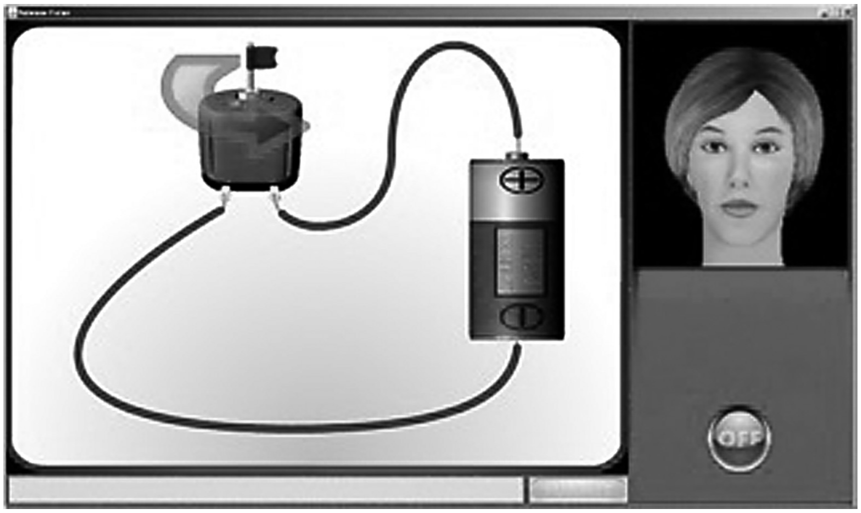


Fig. 1. Virtual Tutor Screen.

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 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.

Since 2007, our research team has been involved in an intensive effort to develop an intelligent tutoring system, My Science Tutor (MyST), intended to improve science learning by 3rd, 4th, and 5th grade children through natural spoken dialogs with Marni, a virtual science tutor. MyST requires the integration of automatic speech recognition, character animation, 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. Figure 1 displays a screenshot of the virtual tutor Marni asking questions about media displayed in tutorial dialog sessions.

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 [Mostow and Aist 2001; Rickel and Johnson 2000; Graesser et al. 2001; Aist and Mostow 2009; Mostow and 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 [Graesser et al. 2001; Littman and Stillman 2004]. Science tutors have been developed for children that incorporate embodied conversational agents (computer character that talk) in multimedia environments [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 dialogues with children have been used successfully to help children learn to read and comprehend text and to assess an individual's proficiency in a given language. For example, work in Project Listen integrated speech recognition and dialogue modeling to improve reading, vocabulary acquisition and text comprehension [Aist and Mostow 2009; Mostow and Chen 2009; Chen et al. 2010]. Bernstein and 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 dialogue systems have been developed for science tutoring for university-level students, and for children for reading and language assessment, we have no evidence 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.

The goal of the MyST project is to help struggling students learn the science concepts encountered in classroom science instruction. Each 15 to 20 minute MyST dialogue 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 dialogues with the virtual tutor Marni. The goal of these *multimedia dialogues* is to help students construct and generate explanations that express their ideas. The dialogues 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 dialogues are linked to the activities, observations and outcomes of classroom science investigations conducted by groups of three to five children in kit-based science investigations that are part of the FOSS (Full Option Science System) program used by over one million students in classrooms in all fifty states in the United States.[1] In addition to the science kits that support an average of 16 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, we developed 16 different tutorial dialogue sessions, lasting about 20 minutes each, for four different areas of science: Variables, Measurement, Water and Magnetism and Electricity. Thus, a total of 64 different tutorials, were developed across the four areas of science to help children think about and explain science concepts encountered during classroom activities.

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 dialogue, 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 dialogue 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?"

[1] www.fossweb.com.

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, and describe the development of a corpus of conversational tutorial sessions. Using the corpus, we present evaluations of the ASR and semantic parsing and dialogue components of the system. In addition to component level evaluations, the MyST project will also assess the system along the dimensions of Engagement (how satisfactory is the user experience?), feasibility (can the system be used in the way proposed in real world situations?), and efficacy (does the system produce learning gains?). A formal evaluation of these aspects of the system is currently being conducted. While these data are not yet available, this paper presents survey results of teachers' and children's impressions of MyST from the data collection done in the 2009–2010 academic year. These surveys give evidence for the Engagement and Feasibility of the system.

2. SCIENTIFIC RATIONALE

MyST is an example of a new generation of intelligent tutoring systems that facilitate learning through spoken dialogues 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 [Graesser et al. 2001; Wise et al. 2005; Lester et al. 1997; Mostow and Aist 2001]. Text, illustrations, and animations may be incorporated into the dialogues. 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 and Graesser 200l; 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" [Palincsar 1998]. Vygotsky [1978] stressed the critical role of social interaction within one's culture in acquiring the social and linguistic tools that are the 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]. 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]. Social constructivism provides the conceptual model for knowledge acquisition in MyST: to improve learning by scaffolding conversations using media to support hypothesis generation and coconstruction 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 and 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. 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 [Cohen et al. 1982], and positive outcomes obtained in large-scale tutoring programs [Topping and Whitley 1990; Madden and Slavin 1989]. 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 [Craig et al. 2000; Driscoll et al. 2003; King 1989] and school children [King 1994; King et al. 1998; Palincsar and 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 [King 1994; King et al. 1998; Palincsar and 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 and 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 coconstruction*. Students coconstruct 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 and Brown 1984; Pine and Messer 2000; Butcher 2006; Soter et al. 2008; Murphy et al. 2009]. Classroom conversations and tutorial dialogs increase the opportunity for occurrences of knowledge coconstruction 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 and Brave 2005]. Such programs foster social agency, enabling users to interact with them the way they interact with people. 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 [Moreno et al. 2001; Atkinson 2002; Baylor et al. 2005]. A number of researchers have observed that children become highly engaged with virtual tutors and appear to interact with a virtual tutor as if it were a real teacher and appear motivated to work hard to please the virtual tutor. Lester et al. [1997] termed this phenomenon the "Persona Effect." In our previous research using Marni as a virtual reading tutor [Cole et al. 2007], over 70% of over 250 students surveyed reported that they trusted Marni, that they felt Marni cared about them, that Marni was a good teacher, and that Marni helped them learn to read.

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; for instance, 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 dialogue 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 dialogues*, 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 automatically 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 dialogues 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 dialogue. 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 type of interaction is called a Wizard of Oz procedure and is described in the following. 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.

3.1. 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 and Beck 1999; McKeown et al. 1999].

QtA is a mature, scientifically-based and effective program used by hundred 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 and Edwards 2005; Murphy et al. 2009], QtA was identified as one of two approaches out of the nine examined that are 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 and 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 and Gamoran 1991; Soter and 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 and McKeown 2006], and has been demonstrated to be effective in improving comprehension of informational texts, we decided to incorporate principles of QtA into tutorial dialogues 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 dialogues, 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 iterative process of testing and refining these QtA-based multimedia dialogues.

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 datasets they enter into their science notebooks. During multimedia dialogues, 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.

3.2. 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 2 can be presented with questions like: "So what's going on here?" or "What's this all about?"

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 2 might be the following.

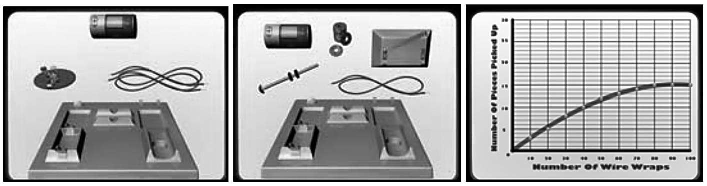
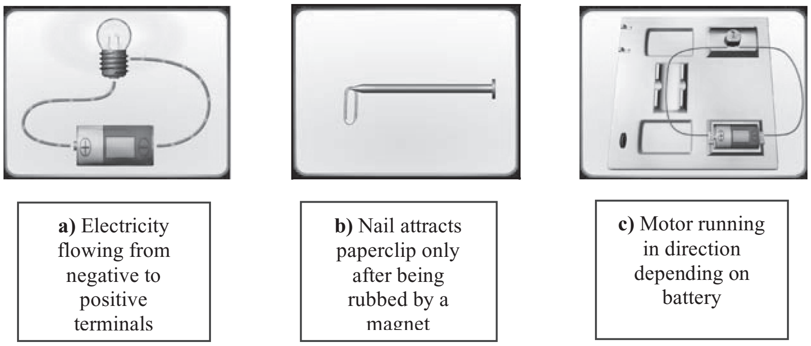


Fig. 2. Example static illustrations.



a) Electricity flowing from negative to positive terminals

b) Nail attracts paperclip only after being rubbed by a magnet

c) Motor running in direction depending on battery

Fig. 3. Example animations.

—*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 could be very helpful when working with a student who grasps what they are looking at, but not how to interpret it. A QtA inspired sequence about the graph might be the following.

—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 noninteractive Flash animations. Simple Animations can provide additional information and help students visualize concepts that can be difficult to capture in illustrations. Figure 3 describes several simple animations, such as the flow of electricity in a circuit and creation of a temporary magnet. In Figure 3(a), the direction of the flow of electricity is represented by blue dots moving

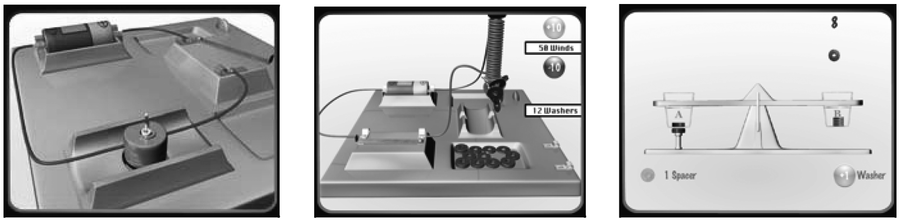


Fig. 4. Examples of interactive animations.

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; for instance, "You got it! The electricity is flowing through the circuit from the negative to the positive side of the D-Cell."

Interactive Animations. Interactive Animations 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 4). 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 DIALOGUES

Creating natural and effective interactions between Marni and the student is the overarching goal of the development process. It is necessary to design dialogues 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 dialogues 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 dialogue. 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 we have discussed.

A major challenge of the MyST project was how to design the spoken dialogues and media in a principled way to optimize engagement and learning. To meet this challenge, we developed an iterative approach to dialogue 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. These cycles involved initial human tutoring using a set of illustrations, to human tutoring with computer-based illustrations, animations, and interactive stimulations, to Wizard of Oz studies (described in the following), 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. In addition, we selected a specific approach to tutorial conversations, based on principles of QtA, and then developed, tested and refined dialogs administered first with human tutors, then to initial MyST dialogues monitored and sometime controlled by human tutors in Wizard of Oz sessions. At each step of the development process, sessions were recorded, transcribed and analyzed, leading to refinements and subsequent testing through a series of iterative design and test cycles, and to the final MyST dialogues now being evaluated in schools.

As noted, the concepts addressed in MyST tutorial sessions are aligned with the structure of FOSS content. 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).

4.1. 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 system (Marni), 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. Marni attempts to elicit the points by encouraging self-expression from the student. Questioning 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 dialogue content (referred to as marking it) followed by repeating (revoicing) the question back to the student using similar phrasing; for instance, You 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.

4.2. Implementing Tutorial Sessions

Marni's behavior in a dialogue with a student, including the presentation of media within dialogues, 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; the extracted information;
—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: 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 Marni speak it. In order to elicit speech to fill a frame element, the system 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

```
Frame: FlowDirection
        [Flow]
        [DirFlow]
                Action: flash(Flow); synth(Tell me about what's going on here.)
                Action: synth(What do you notice about the flow?)
        [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?)
```

Fig. 5. Example task frame.

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 5 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*."

Rules are also 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.

The 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.

As noted above, development of dialogues, as represented in the *task* files, proceeds through an 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.

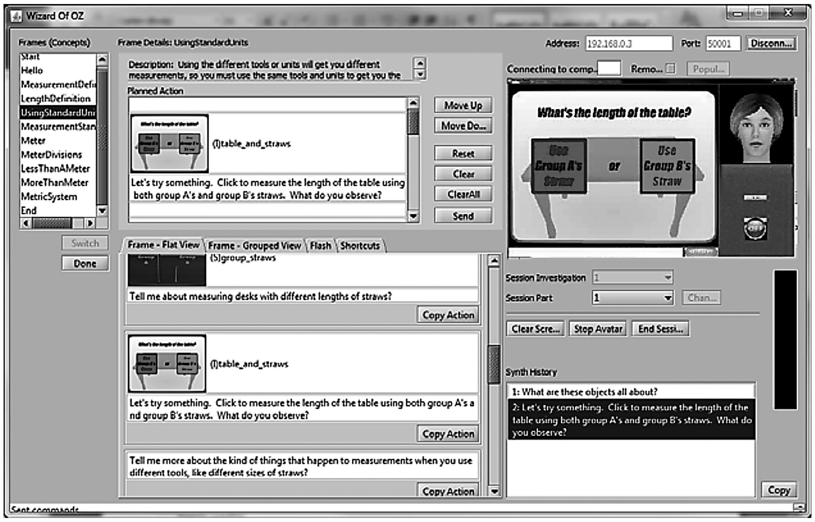
5. WIZARD-OF-OZ INTERFACE

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 Marni, while the human tutor can monitor the student's interaction with the system and alter system behavior when desired. This type of data collection system is often referred to as a Wizard-of-Oz system (WOZ). 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 (either at some other location at the school or at Boulder Language Technologies offices) have monitored the tutoring sessions remotely. One project tutor goes 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. For new tutorials, wizards intervene on an average of about 33% of the turns. This number reduces quickly to about 20%. Fewer than 1% of the wizard interventions involve changing the focus frame. The correct concept was being discussed, but the wizard wanted to say something different.

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 6 shows the layout of the Wizard display, which contains the following.

- A screenshot of the screen that the student sees
- The action Marni is about to take
- The frame in focus, including all action sequences associated with elements of the frame


Description: Using the different tools or units will get you different measurements, so you must use the same tools and units to get you the

Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Tell me about measuring desks with different lengths of straws?

Let's try something. Click to measure the length of the table using both group A's a nd group B's straws. What do you observe?

Tell me more about the kind of things that happen to measurements when you use different tools, like different sizes of straws?

1: What are these objects all about?

2: Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Fig. 6. Wizard screen.

- 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 Marni 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 system keeps a log of time-stamped events occurring during the session, including any wizard generated actions. 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 system to behave more like the wizards.

Student Interface. An example of the student's screen is shown in Figure 1. 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 systems work well with children 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.

Dialogue 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 dialogue 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 dialogue 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 classroom 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 dialogue 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 QtA 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 dialogue, 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.

6. MYST SYSTEM ARCHITECTURE

MyST was developed using Boulder Language Technologies Virtual Human Toolkit (VHT). The BLT VHT is a resource for designing and experimenting with multimedia programs that support real time conversational interaction with virtual humans. The VHT 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

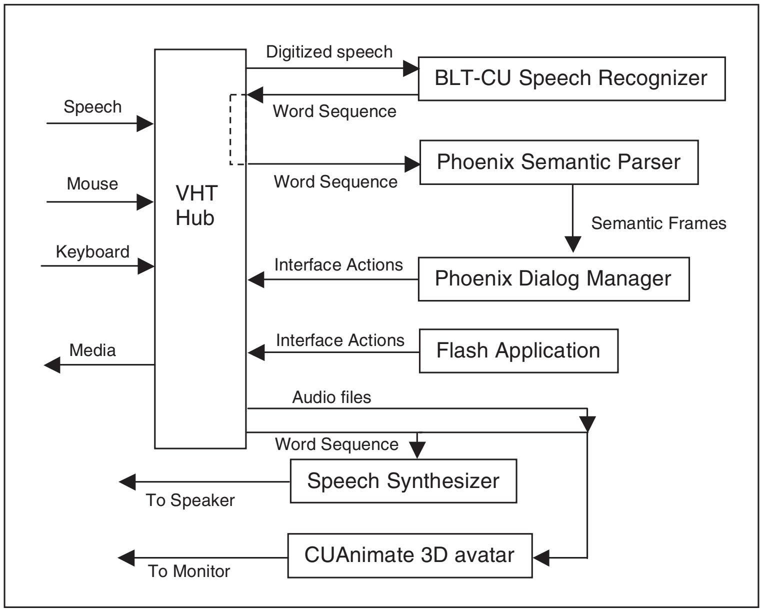


Fig. 7. Virtual human toolkit architecture.

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.

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 7.

6.1. 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 the following.

—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 prerecorded 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 7, 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.

6.2. Speech Recognizer

The speech recognizer used in the VHT is a large vocabulary continuous speech recognition (LVCSR) system written by Daniel Bolaños [Bolaños et al. 2011], 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].

Feature extraction from the audio was carried out using Mel Frequency Cepstral Coefficients (MFCC) plus the logarithm of the signal energy. Cepstral coefficients were extracted using a 20ms window size and a 10ms shift, which produces a feature vector each 10ms that is composed of 12 MFCCs plus the log energy and the first and second order derivatives. Acoustic Models are clustered triphones based on Hidden Markov Models using Gaussian Mixtures to estimate the probabilities of the acoustic observation vectors. The system uses filler 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. The recognizer is configured to run in approximately real time so the delay after the student quits speaking and before the system is ready to respond is kept short. This is necessary to promote a fluent and engaging dialog.

6.3. 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.

6.4. Dialog Manager

The Dialog Manager controls the system's dialogue 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 was discussed in Section 4.

6.5. Character Animation

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 [Ma et al. 2002, 2004] 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). In MyST, the character for Marni shown in Figure 1 was used in all applications.

6.6. 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.

7. USE OF SPOKEN RESPONSES

In the tradition of other systems using children's speech [Mostow and 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. When these occur, the system asks a follow-on question, which may be accompanied by a new illustration, animation or interactive investigation, that is designed to scaffold learning and elicit an appropriate response. Thus, the interaction style used in Questioning the Author is especially well suited to ASR errors that can occur during spoken dialogues.

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 sessions where the system is able to accurately recognize and parse student responses, it is able to adapt the tutorial dialogue to the individual student. It may move on as soon a student expresses an understanding of a point, or 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 dialogue, the system will conclude the session with a default tutorial presentation as specified in the *task* file for 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 generate a follow-on question. These dialogue techniques are designed to assure the student that Marni is listening to and understands what the student is saying. Marni 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, and communicates this understanding back to the student. 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.

8. 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 are also using the data for modeling tutorial dialogs and determining features that are associated with learning gains. At the completion of the project, the corpus, which will contain over 150 hours of children's speech during tutorial dialogs, will be made available to the research community.

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, which vary from low to high performing on state science tests. 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.

Data are being collected in three basic conditions.

(1) *Human Tutor.* 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.
(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. This is the procedure being used to assess the effectiveness of the MyST system in schools. Data collection in this condition recently started and is not included in Table I, or in any of the data sets used in this paper.

Table I shows the amount of data (number of speakers and hours of speech) collected for each module. The Water module was developed last and collection is just beginning.

Speech Files. The speech data are stored in files by student turns, that is, whatever is said from the time the student pressed the space bar to talk until the bar is released.

Table I. Data Collected by Module

Module	All		Training		Development		Evaluation	
	speakers	hours	speakers	hours	speakers	hours	speakers	hours
Magnetism and Electricity	176	35	149	31	14	2	13	2
Measurement	222	48	185	38	20	5	17	5
Variables	60	20	44	18	6	1	10	1
Water	25	8	22	6	1	1	2	1
Total	483	111	400	93	41	9	42	9

The speech is sampled at 16 KHz, 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, 111 hours of speech have been transcribed.

Log files. Each MyST 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. The speech recognition output is logged. Manual transcription of the speech files is performed offline and is introduced into the log file later. Some additional pieces of information stored in the log file are: extracted frame elements, current context, frame name, and frame element or rule that is generating the system response, the number of times this frame element or rule has been used, and the 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.

9. COMPONENT EVALUATIONS

Since only a small amount of data has been collected for the Water (WA) module, and transcripts for those are not completed, experiments were conducted using data from only 3 modules; Magnetism & Electricity (ME), Measurement (MS), and Variables (VB). As shown in Table I, the data were partitioned by speaker into training, development and evaluation 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.

9.1. Automatic Speech Recognition Performance

The recognizer was trained and parameterized using the training and development data and run on the evaluation set using a language model, trained on all training data, that has a perplexity of 63 for the evaluation set. The vocabulary size was 6235 words. The Word Error Rate (WER) for the recognizer on the Evaluation set is shown in Table II in the *Baseline* column. The Out of Vocabulary word rate was very low for all modules, ranging from 0.6% for Magnetism and Electricity to 0.7% for Variables. There were a total of 65,496 words in the evaluation set.

Table II. Results for Speech Recognition

	Baseline		+VTLN		+VTLN +MLLR	
	WER(%)	CA	WER(%)	CA	WER(%)	CA
ME	29.8	.85/.89	28.1	.87/.91	26.1	.87/.91
MS	29.6	.83/.87	28.6	.84/.87	26.7	.86/.89
VB	36.1	.82/.89	34.3	.80/.87	31.9	.82/.90
Tot	30.9	.84/.89	29.5	.85/.89	27.4	.86/.90

The WER for the pooled data (Tot) was 30.9%. For the individual modules, the WER for ME and MS were very similar, while the WER for VB was substantially higher. Using a global LM, the perplexity of each module was: 56 for ME, 63 for MS and 74 for VB. Even though the ME data had a lower perplexity than the MS, the WERs are similar. VB had a substantially higher perplexity and WER. The higher perplexity of the VB data can be attributed both to less training data and to the topic of the module. The ME and MS modules are about concrete topics with which students are generally familiar. Variables introduces more abstract ideas like dependent and independent variables and graphing data. Students generally have a more difficult time with this topic, even with human tutors.

The baseline results reported above were obtained using speaker-independent acoustic models, but not adapted to the current user. A number of speaker adaptation techniques are commonly used in ASR systems. Two of the most effective are Maximum Likelihood Linear Regression [Leggetter and Woodland 1995] and Vocal Track Length Normalization [Lee and Rose 1998]. Vocal Tract Length Normalization (VTLN) is motivated by the fact that different speakers have vocal tracts of different length, which results in a variation of the format frequencies. VTLN compensates for this variability by applying a warping factor to the speech spectrum in the frequency domain. For each speaker, a first pass of the decoder was run to generate a hypothesis word string. A warping factor was then computed for the speaker to maximize the likelihood of the features extracted from the speech given the hypothesis. This warping factor is then used to produce a final hypothesis in a second decoding pass. The application of VTLN reduced the WER from 30.9% to 29.5%. MLLR works in the acoustic model space, rather than feature space like VTLN, and consists of applying a set of transforms to the Gaussian means and covariances of the speaker independent acoustic models to better match the speech characteristics of the target speaker. Transforms are estimated so that, when applied to the parameters of the acoustic models, the likelihood of the speaker data is maximized with respect to the hypothesized sequence of words. Speaker data are then re-decoded after applying the transforms. The number of transforms is determined dynamically based on the adaptation data available. A regression class tree is used to cluster the Gaussian components in the system. The number of base classes in the tree was set to 50 and the tree was built using EM clustering. Full transformation matrices for the means and diagonal transformation matrices for the variances were used. The minimum class occupancy count was set to 3500. Adding MLLR adaptation reduced the error rate further to 27.4%.

For the numbers we have listed, the adaptation techniques were applied in a batch unsupervised mode using all of the data for the particular speaker. In a live application, for new users, warping factors and transforms would need to be computed incrementally as more data come in, or after a certain minimum amount of speech data were available. The benefits of adaptation would initially be small and should improve as more speech data become available. In this intervention (MyST), it is anticipated that an individual student will use the system repeatedly over a period of time. A single FOSS Module will have 16 tutorial sessions associated with it, each lasting about 20 min. The cumulative data from each user will be used to precompute warp factors

Table III. Speech Recognition Performance by LM

	Overall			ME			MS			VB		
	WER (%)	PP	CA	WER (%)	PP	CA	WER (%)	PP	CA	WER (%)	PP	CA
Global	27.4	63	.86 .90	26.1	56	.87 .91	26.7	63	.86 .89	31.9	74	.82 .90
ModA	27.8	62	.87 .90	26.3	55	.88 .91	27.1	63	.87 .89	32.0	73	.83 .89
ModS	27.9	60	.88 .89	26.3	53	.89 .90	27.1	59	.87 .88	32.0	73	.86 .87
InvA	27.2	61	.87 .89	26.0	54	.89 .91	26.3	62	.87 .89	31.7	72	.85 .88
InvS	29.2	64	.89 .87	27.8	56	.90 .89	28.3	64	.88 .87	34.0	75	.86 .85

and transforms that are stored and loaded when the user logs in. On average, first time users will initially experience system performance similar to that in the Baseline column in Table II, WER of around 31%. The system will incrementally adapt as more data from the user are available over sessions. Since the batch unsupervised adaptation described above not only adapts to the speaker, but also to the test data, performance in live use would not be expected to fully reach the same level of performance.

Effect of Language Model Specificity. The VHT speech decoder uses standard trigram language models that were trained using the CMU-Cambridge Language Model Toolkit [Clarkson and Rosenfeld 1997]. In creating language models for structured data such as this, the developer has the opportunity to tune the model to the specific topic of the investigations. In this case, a general language model is trained and adaptation data is used to tune the model for a specific topic. One effective method for language model adaptation is to use MAP (maximum a posteriori) adaptation, which combines weighted word counts from the general data and adaptation data [Federico 1996]. We used a simple approximation to this procedure by mixing adaptation data with general data with a weighting factor. The weighting factor was determined using a development set. Performance of the recognizer was determined using five sets of Language Models (LMs).

(1) *Global*. A single LM is trained by pooling all training data;
(2) *ModA*. A separate LM is generated for each module by adapting the Global model with the training data for the module;
(3) *ModS*. A separate LM is trained for each module using just the training data for the module;
(4) *InvA*. A separate LM is generated for each investigation by adapting the Global model with the training data for the investigation;
(5) *InvS*. A separate LM is trained for each investigation using just the training data for the investigation.

The WER in %, Perplexity (PP) and Concept Accuracy (CA) for modules in the Evaluation set are shown in Table III. For CA, the top number is Recall and the bottom number is Precision. The WER, PP and CA numbers for investigation specific models are an average across the investigations of each module. There is a small difference in WER as the LMs become more specific. Results with the Module Adapted (ModA) and Module Specific (ModS) LMs are substantially equivalent and are slightly worse than the WER achieved with the Global LM. The Investigation Adapted (InvA) LMs had, for each module and overall, lower WER than the Investigation Specific (InvS) LMs which had the highest WER. The data were not clearly sufficient to train investigation specific LMs, but LM adaptation helped a little bit in this case, although not enough to ensure a significant improvement with respect to the WER achieved with the Global LM. Variations in perplexity across LMs are also small.

Disfluencies. Conversational speech contains many events that are nonwords, such as breath and filled pauses. A common technique to deal with these events is to use acoustic filler models to match the input. In addition to a Silence model, the system uses acoustic models to match nonword speech events (br, EM, HMM, HUH, MMM, UHM). The decoder that produced the results in Tables II and III used filler models. Fillers are allowed to occur between any events (words or other fillers) with an insertion penalty that is set to minimize WER using the development set. We conducted an investigation to give some information about the performance of the filler models used in the system. Using a global language model, the overall WERs of the baseline system and the adapted system were 30.9% and 27.4%, respectively. Approximately 6.7% of the annotated tokens in the evaluation set transcriptions were fillers. Filler tokens are stripped out of the recognition hypotheses before computing WER and before parsing, so insertions of filler tokens do not in themselves cause a problem. A problem can occur when recognizing the filler causes a word deletion or substitution error. Without using filler models the WERs increased to 35.1% and 29.3%. It was clearly beneficial to overall WER to include filler models in the decoder. Even using filler models, disfluencies continue to be a significant problem in ASR for children's conversational speech.

9.2. Concept Accuracy

The behavior of the virtual tutor is more dependent on Concept Accuracy than on Word Error Rate. The only representation that the Dialog Manager has of what the student said are the extracted frames 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. Grammars are created for each investigation (there are 4 investigations for each module) using the training data. The investigations have an average of 8 frames with an average of 5 frame elements per frame, thus there are about 40 frame element classes on average in an investigation. Reference parses were created for each hand transcribed utterance by parsing the transcripts, which represent word input with no ASR 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. Recall is the percentage of the reference elements that were correctly extracted from the recognizer output. Precision is the percentage of the elements extracted from the recognizer output that were correct. The results for Concept Accuracy are shown in the columns labeled CA in Tables II and III. The first (or top) number in the accuracy is Recall and the second (or bottom) number is Precision. As seen in Table II, using a global LM the baseline system had a WER of 30.9% with an overall Recall of .84 and Precision of .89. With batch unsupervised speaker adaptation, a WER of 27.4% with a Recall of .86 and a Precision of .90 were achieved. This generally would be the expected effect of recognizing more content words correctly. As seen in Table III, 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. This expectation is consistent with the CA results.

10. 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. All of these students used MyST in WOZ mode. Measures were taken to avoid bias wherein students give overly positive answers to questionnaires including: 1) written (versus oral) surveys for students were administered, 2) students were verbally assured of anonymity, 3) questionnaires were anonymous in

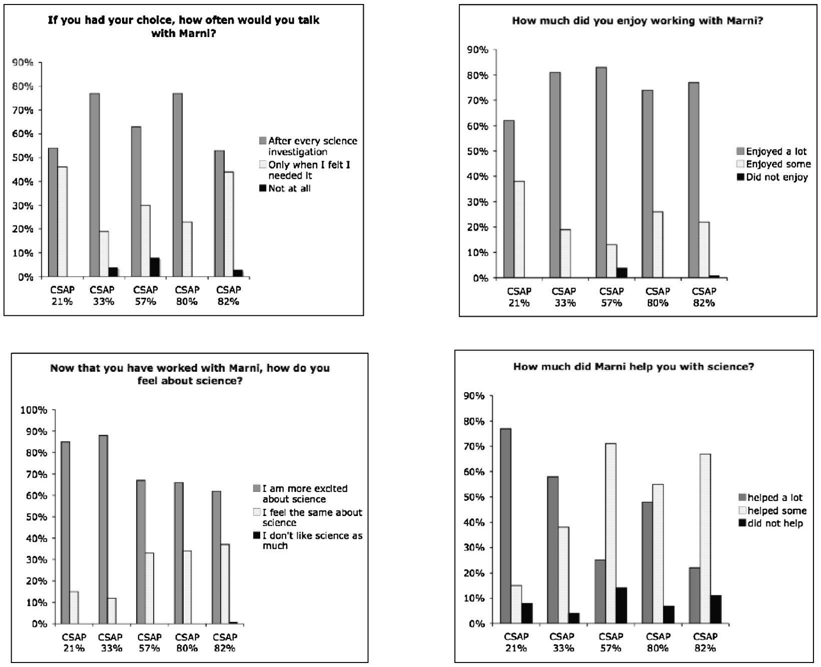


Fig. 8. Student survey responses by school CSAP score.

that students did not write their names on the survey, and 4) adults from the program did not directly observe or interfere with students while they completed the survey. The survey included ten questions that asked for ratings of student experience and impressions of the program and its usability. Three point rating scales for survey items were keyed to each question. A typical question, such as "How much did Marni help with science?" had responses such as: "Did not help, helped some, helped a lot." Items were written to reflect the reading level of the students. Four main questions assessed student experiences with Marni: 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?"

The schools in which students used MyST varied greatly in terms of the percentage of students who scored proficient or above in science on the state Colorado Student Assessment Program (CSAP) test: from 21% proficient or above for the lowest scoring school, to 82% proficient or above in the highest scoring school. Figure 8 displays the distribution of students' response choices to each question. The histograms are grouped by school, using the percentage of students at the school who scored as proficient or above. 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

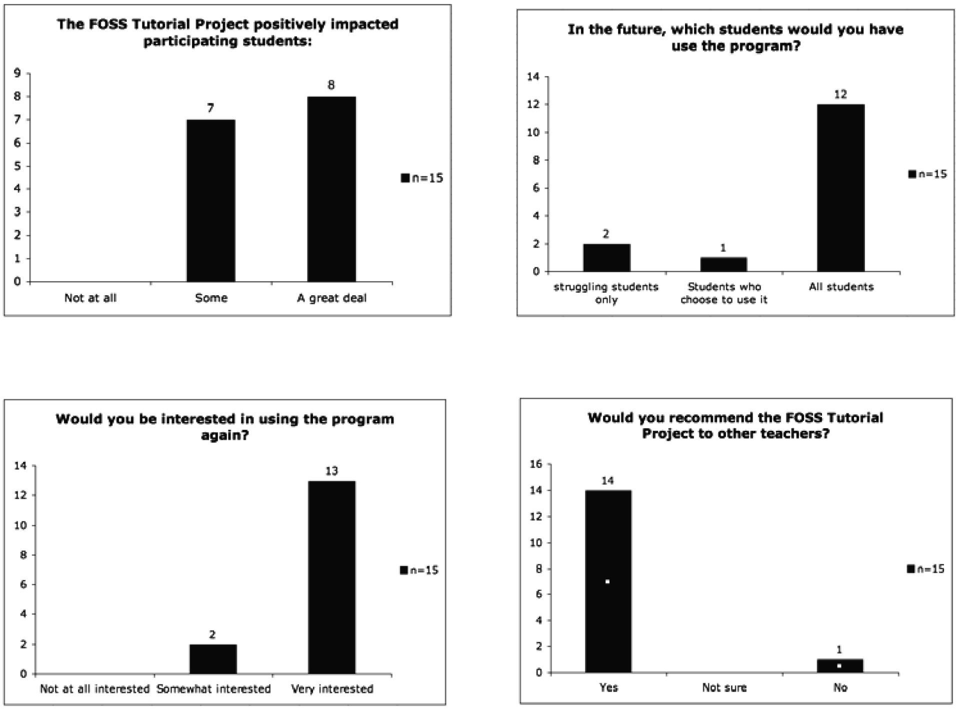


Fig. 9. Teacher survey results.

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 have more positive impressions of the program.

Teachers were asked for feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. A teacher survey was administered to all participating teachers directly after their students completed tutoring. Teachers were assured anonymity in their responses both verbally and in written form. The questionnaire contained 22 rating items as well as 9 open-ended questions. The survey asked teachers about the perceived impact of using Marni for student learning and engagement, impacts on instruction and scheduling, willingness to potentially adopt Marni as part of classroom instruction, and overall favorability toward participating in the research project. Additionally, teachers answered items related to potential barriers in implementing new technology in the classroom. The results of the survey are shown in Figure 9. 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 enthused about and engaged in classroom activities, and that their participation in science investigations and classroom discussions benefited students who did not use the system. 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.

11. CONCLUSIONS AND FUTURE WORK

This article has presented the design of a conversational multimedia virtual tutor for elementary school science. Speech and language technologies play a central role because the focus of the system is on engagement and self-expression by the students. It was argued that current speech and natural language technology is a good match to this task because it takes advantage of speech understanding capabilities to improve the interaction while minimizing the effects of errors in recognition and understanding.

A corpus is being developed which will be used to evaluate the MyST system as well as enable research by others on tutorial dialog systems. Evaluation results were presented for the Automatic Speech Recognition and Spoken Language Understanding components of the system. Using a global LM, the baseline system had a WER of 30.9% with an overall Recall of 84% and Precision of 89% for extraction of frame elements. With batch unsupervised speaker adaptation, a WER of 27.4% with a Recall of 86% and a Precision of 90% were achieved. The accuracy of extraction of frame elements measures how well the system is understanding the student. Performance of live systems would average somewhere between these performance numbers.

During data collection using a WOZ paradigm, surveys were collected from students and teachers that bear on the engagement and feasibility of the proposed tutoring system. 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 unanticipated benefits of this shared perception to our project was 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.

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.

The survey responses reported in this article are based on experiences with a WOZ system. Students interacted with a virtual tutor, but a human tutor was moderating the interaction. Survey responses and anecdotal evidence in observing interactions indicate that both students and teachers are accepting of the virtual tutor. What remains to be shown is how well the virtual tutor is able to maintain engagement without the assistance of a wizard. The efficacy of the system in the form of learning gains also needs to be determined. At the time of this writing, during the 2010–2011 school year, MyST is being evaluated in stand-alone mode. In addition to student and teacher surveys, the system 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 Wizards in the development phase of the project. ASK assessments are given to students before and after each science module. 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 and after each science module. In addition, the performance of these students will be compared to the performance of students who are administered ASK assessments in classrooms that did not receive tutoring. Our hypothesis is that students who engage in multimedia dialogs with Marni will produce benefits similar to students who interact with human tutors. One of the most important outcomes of the assessment procedure will be determining the feasibility and potential of using speech and language processing technologies in multimedia tutoring dialogs with children.

REFERENCES

Aist, G. and Mostow, J. 2009. Designing spoken tutorial dialogue with children to elicit predictable but educationally valuable responses. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*.

Atkinson, R. K. 2002. Optimizing learning from examples using animated pedagogical agents. *J. Educ. Psych.* 94, 416–427.

Baylor, A. L. and Ryu, J. 2003. Does the presence of image and animation enhance pedagogical agent persona? *J. Edu. Comput. Resear., 28*, 4, 373–395.

Baylor, A. L. and Kim, Y. 2005. Simulating instructional roles through pedagogical agents. *Int. J. Artific. Intell. Edu. 15*, 1.

Beck, I. L., McKeown, M. G., Worthy, J., Sandora, C. A., and Kucan, L. 1996. Questioning the author: A year-long classroom implementation to engage students with text. *Elem. School J. 96*, 4, 387–416.

Beck, I. and McKeown, M. 2006. *Improving Comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach*. Scholastic.

Bernstein, J. and Cheng, J. 2007. Logic and validation of fully automatic spoken English test. In *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice,* M. Holland and F. P. Fisher, Eds., Routledge, 174–194. http://www.ordinate.com/samples/Versant-English/Sample-TEST-PAPER-Versant-English-Test-watermark.pdf

Bloom, B. S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educ. Resear. 13*, 4–16.

Bolanos, D., Cole, R., Ward, W., Borts, E., and Svirsky, E. 2011. FLORA: Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process.*

Bruner, J. S. 1966. *Toward a Theory of Instruction*. Harvard University Press, Cambridge, MA.

Bruner, J. S. 1990. *Acts of Meaning*. Harvard University Press, Cambridge, MA.

Butcher, K. R. 2006. Learning from text with diagrams: Promoting mental model development and inference generation. *J. Edu. Psych. 98*, 1, 182–197.

Chapin, S. H., O'Connor, C., and Anderson, N. C. 2003. *Classroom Discussions Using Math Talk to Help Students Learn*. Math Solution Publications, Sausalito, CA.

Chen, W., Mostow, J., and Aist, G. 2010. Exploiting predictable response training to improve automatic recognition of children's spoken questions. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS2010)*, Springer-Verlag, 55–64.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci. 13*, 145–182.

Chi, M. T. H., De Leeuw, N., Chiu, M., and Lavancher, C. 1994. Eliciting self-explanations improves understanding. *Cogn. Sci. 18*, 439–477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., and Hausmann, R. G. 2001. Learning from human tutoring. *Cogn. Sci. 25,* 471–533.

Clarkson, P. R. and Rosenfeld, R. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech*.

Cohen, P. A., Kulik, J. A., and Kulik, C. L. C. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *Am. Edu. Resear. J. 19*, 237–248.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., and Yan, J. 2003. Perceptive animated interfaces: First steps toward a new paradigm for human–computer interaction. In*Proc. IEEE 91*, 9, 1391–1405.

Cole, R., Wise, B., and Van Vuuren, S. 2007. How Marni teaches children to read. *Educ. Techn.*

Craig, S. D., Gholson, B., Ventura, M., Graesser, A. S., and Tutoring Research Group. 2000. Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *Int. J. Artif. Intell. Edu. 11,* 242–253.

Dede, C., Salzman, M., Loftin, B., and Ash, K. (in press). *Using virtual reality technology to convey abstract scientific concepts*. In *Learning the Sciences of the 21st Century: Research, Design, and Implementing Advanced Technology Learning Environments*. M. J. Jacobson and R. B. Kozma, Eds. Lawrence Erlbaum, Hillsdale, NJ.

Driscoll, D., Craig, S. D., Gholson, B., Ventura, M., Hu, X., and Graesser, A. 2003. Vicarious learning: Effects of overhearing dialog and monolog-like discourse in a virtual tutoring session. *J. Educ. Comput. Resear. 29,* 431–450.

Federico, M. 1996. Bayesian Estimation Methods for n-gram language model adaptation. In *Proceedings of ICSLP'96*, 240–243.

Graesser, A. C., Hu, X., Susarla, S., Harter, D., Person, N. K., Louwerse, M., Olde, B. and the Tutoring Research Group. 2001. AutoTutor: An intelligent tutor and conversational tutoring scaffold. In *Proceedings of the 10th International Conference of Artificial Intelligence in Education,* 47–49.

Graesser, A., N., Person, N., and Harter D. 2001. Teaching tactics and dialog in Autotutor. *Int. J. Artific. Intell. Edu.*

Hausmann, R. G. M. and Vanlehn, K. 2007a. Explaining self-explaining: A contrast between content and generation. *Artificial Intelligence in Education*, R. Luckin, K. R. Koedinger, and J. Greer, Eds. IOS Press, Amsterdam, Netherlands, 417–424.

Hausmann, R. G. M. and Vanlehn, K. 2007b. Self-explaining in the classroom: Learning curve evidence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* D. McNamara and G. Trafton Eds., Erlbaum, Mahwah, NJ, 1067–1072.

King, A. 1989. Effects of self-questioning training on college students' comprehension of lectures. *Contemp. Educ. Psy. 14*, 366–381.

King, A. 1991. Effects of training in strategic questioning on children's problem-solving performance. *J. Educ. Psych. 83*, 307–317.

King, A. 1994. Guiding knowledge construction in the classroom: Effect of teaching children how to question and explain. *Am. Educ. Resear. J. 31,* 338–368.

King, A., Staffieri, A., and Adelgais, A. 1998. Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *J. Educ. Psych. 90,* 134–15.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psych. Rev. 95,* 163–182.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, England.

Lee, L. and Rose, R. C. 1998. A frequency warping approach to speaker normalization. *IEEE Trans. Speech Audio Process. 6*, 1, 49–60.

Leggetter, C. J. And Woodland, P. C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. *Comput. Speech Langu. 9*, 171–185.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., and Boghal, R. 1997. The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI'97, ACM,* New York, 359–366.

Lester, J., Stone, B., and Stelling. G. 1999. Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Model. User-Adap. Interact. 9*, 1–2, 1–44.

Littman, D. and Silliman, X. 2004. ITSPOKE: An intelligent tutoring spoken dialog system. In *Proceedings of HLT-NAACL*, 5–8.

Ma, J., Cole, R. A., Pellom, B., Ward, W., and Wise, B. 2004. Accurate automatic visible speech synthesis of arbitrary 3d models based on concatenation of di-viseme motion capture data. *J. Comput. Anim. Virt. Worlds 15*, 485–500.

Ma, J. Yan, J., and Cole, R. 2002. CU Animate: Tools for enabling conversations with animated characters. In *Proceedings of the International Conference on Spoken Language Processing*.

Madden, N. A. and Slavin, R. E. 1989. Effective pullout programs for students at risk. in *Effective Programs for Students At Risk*, R. E. Slavin, N. L. Karweit, and N. A. Madden, Eds., Allyn and Bacon, Boston, MA.

Mayer, R. 2001. *Multimedia Learning*. Cambridge University Press, Cambridge, UK.

McKeown, M. G. and Beck, I. L. 1999. Getting the discussion started. *Educ. Leader. 57*, 3, 25–28.

McKeown, M. G., Beck, I. L., Hamilton, R., and Kucan, L. 1999. *Accessibles—Questioning the Author (Easy-Access Resources for Classroom Challenges)*. Wright Group, Bothell, WA.

Moreno, R., Mayer, R. E., Spires, H. A., and Lester, J. C. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cogn. Inst. 19*, 2, 177–213.

MOSTOW, J. AND AIST, G. 1999. Giving help and praise in a reading tutor with imperfect listening—Because automated speech recognition means never being able to say you're certain. *CALICO J. 16*, 3, 407–424.

MOSTOW, J. AND AIST, G. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In *Smart Machines in Education*, K. Forbus and P. Feltovich, Eds.

MOSTOW, J., AIST, G., BURKHEAD, P., CORBETT, A., CUNEO, A., EITELMAN, S., HUANG, C., JUNKER, B., SKLAR, M. B., AND TOBIN, B. 2003. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *J. Educa. Comput. Resear. 29*, 1, 61–117.

MOSTOW, J. AND CHEN, W. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED'09)*. 465–472.

MURPHY, P. K. AND EDWARDS. M. N. 2005. What the studies tell us: A meta-analysis of discussion approaches. In *Making Sense of Group Discussions Designed to Promote High-Level Comprehension of Texts. Symposium Presented at the Annual Meeting of the American Educational Research Association*.

MURPHY, P. K., WILKINSON, I. A. G., SOTER, A. O., HENNESSEY, M. N., AND ALEXANDER, J. F. 2009. Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *J. Educ. Psych. 101*, 740–764.

NAEP. 2002. http://nces.ed.gov/nationsreportcard

NASS C. AND BRAVE S. 2005. Wired for Speech: How Voice Activates and Advances The Human-Computer Relationship. MIT Press, Cambridge, MA.

NYSTRAND, M. AND GAMORAN, A. 1991. Instructional discourse, student engagement, and literature achievement. *Resear. Teach. English 25,* 261–290.

PALINCSAR, A. S. 1998. Social constructivist perspectives on teaching and learning. *Annual Revi. Psych. 49*, 345–375.

PALINCSAR, A. S. AND BROWN, A. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cogn. Instr. 1,* 117–175.

PINE, K. J. AND MESSER, D. J. 2000. The effect of explaining another's actions on children's implicit theories of balance. *Cogn. Instr. 18*, 1, 35–51.

REEVES, B. AND NASS, C. 1996. *The Media Equation,* Cambridge University Press, Cambridge, UK.

RICKEL, J. AND JOHNSON, W. L. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds.

SOTER, A. O. AND RUDGE, L. 2005. What the discourse tells us: Talk and indicators of high-level comprehension. In *Proceedings of the Annual Meeting of the American Educational Research Association*. 11–15.

SOTER, A. O., WILKINSON, I. A. G., MURPHY, P. K., RUDGE, L., RENINGER, K., AND EDWARDS, M. 2008. What the discourse tells us: Talk and indicators of high-level comprehension. *Int. J. Educ. Resear. 47*, 372–391.

TAYLOR, P., BLACK, A. W., AND CALEY, R. 1998. The architecture of the festival speech synthesis. In *Proceedings of the 3rd ESCA Workshop in Speech Synthesis*. 147–151.

TOPPING, K. AND WHITLEY, M. 1990. Participant evaluation of parent-tutored and peer-tutored projects in reading, In *Educa. Resear. 32*, 1, 14–32.

VAN LEHN, K. AND GRAESSER, A. C. 2002. Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Unpublished report prepared by the University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.

VAN LEHN, K., LYNCH, C., TAYLOR, L., WEINSTEIN, A., SHELBY, R., SCHULZE, K., TREACY, D., AND WINTERSGILL, M. 2003. In *Intelligent Tutoring Systems,* S. A. Cerri, G. Gouarderes, and F. Paraguacu, Eds. Springer, Berlin, Germany, 367–376.

VANLEHN, K., LYNCH, C., SCHULZE, K. SHAPIRO, J. A., SHELBY, R., TAYLOR, L., TREACY, D., WEINSTEIN, A., AND WINTERSGILL, M. 2005. The Andes physics tutoring system: Five years of evaluations. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*. G. McCalla and C. K. Looi, Eds. IOR Press, Amsterdam.

VYGOTSKY, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, MA.

WARD, W. 1994. Extracting information from spontaneous speech, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*

WARD, W. AND PELLOM, B. 1999. The CU Communicator system. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.

Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J., and Pellom, B. 2005. Learning to read with a virtual tutor: foundations to literacy. In *Interactive Literacy Education,* C. Kinzer and L. Verhoeven, Eds., Environments through Technology. Lawrence Erlbaum, Mahwah, NJ.

Wood, D. and Middleton, D. 1975. A study of assisted problem solving. *Brit. J. Psych. 66,* 181–191.

Received June 2010; revised October 2010 and April 2011; accepted April 2011