# Spoken Versus Typed Human and Computer Dialogue Tutoring

**Diane J. Litman,** *Learning Research and Development Center/Computer Science Department, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260*
*litman@cs.pitt.edu*

**Carolyn P. Rosé,** *Language Technologies Institute/Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15260*
*cprose@cs.cmu.edu*

**Kate Forbes-Riley,** *Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260*
*forbesk@cs.pitt.edu*

**Kurt VanLehn,** *Learning Research and Development Center/Computer Science Department, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260*
*vanlehn@cs.pitt.edu*

**Dumisizwe Bhembe,** *Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260*
*bhembe@pitt.edu*

**Scott Silliman,** *Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260*
*scotts@pitt.edu*

**Abstract.** While human tutors typically interact with students using spoken dialogue, most computer dialogue tutors are text-based. We have conducted two experiments comparing *typed* and *spoken* tutoring dialogues, one in a human-human scenario, and another in a human-computer scenario. In both experiments, we compared spoken versus typed tutoring for learning gains and time on task, and also measured the correlations of learning gains with dialogue features. Our main results are that changing the modality from text to speech caused changes in the learning gains, time and superficial dialogue characteristics of human tutoring, but for computer tutoring it made less difference.

**Keywords.** dialogue, evaluation of AIED systems, intelligent tutoring systems, natural language interfaces for instructional systems, spoken language interface

## INTRODUCTION

It is widely believed that the best human tutors are more effective than the best computer tutors, in part because Bloom (1984) found that human tutors could produce a larger difference in the learning gains, 2.0 standard deviations, than current computer tutors (e.g., (Anderson et al., 1995; VanLehn et al., 2005; Graesser et al., 1999)), which typically produce a 1.0 standard deviation gain. A major difference between human and computer tutors is that human tutors use face-to face spoken natural language dialogue, whereas computer tutors typically use menu-based interactions or typed natural language dialogue. This raises the question of whether making the interaction more natural, such as by changing the modality of the computer tutoring to spoken natural language dialogue, would decrease the advantage of human tutoring over computer tutoring.

In fact, as will be detailed below, several potential benefits of spoken tutorial dialogue with respect to increasing learning have already been hypothesized in the literature. One hypothesis is that spoken dialogue may be better at eliciting student behaviors that are believed to accelerate learning, such as student knowledge construction. A second hypothesis is that speech allows tutors to infer a more accurate student model, which similarly is believed to accelerate learning. A third hypothesis is that speech primes a more social interpretation of the tutorial environment, which again is hypothesized to accelerate learning.

It is thus important to test whether a move to spoken dialogues is likely to yield increased benefits with respect to learning and other performance measures. Furthermore, if the addition of speech can indeed increase learning gains, it is also important to understand why spoken dialogue accelerates learning. These are the overarching objectives of the work reported here.

It is particularly important given that natural language tutoring systems are becoming more common. Although a few use spoken dialogues (Schultz et al., 2003; Mostow & Aist, 2001), most still use typed dialogues (e.g. (Rosé et al., 2001; Heffernan & Koedinger, 2002; Ashley et al., 2002; Michael et al., 2003; Zinn et al., 2002; Aleven et al., 2002, 2001; Rosé & Freedman, 2000; Rosé & Aleven, 2002; VanLehn et al., 2002; Aleven & Rosé, 2003)). As shown by our work it is technically feasible to convert a tutor from typed dialogue tutor to spoken dialogue. Indeed that is just what we have done. While the details of this conversion are not covered in this paper, it took about nine person-months of effort. Thus, many developers may be wondering whether they should aim for a spoken or a typed dialogue tutoring system.

It is also important to study the difference between spoken and typed dialogue in two contexts: human tutoring and computer tutoring. As will be seen, our human and computer tutoring results do in fact differ somewhat. Given the current limitations of both speech and natural language processing technologies, computer tutors are far less flexible than human tutors, and also make more errors (e.g., in transcribing and interpreting student speech). The use of human tutors provides a benchmark for estimating the performance of an “ideal” computer system with respect to speech and natural language processing performance. That is, our analysis of human tutoring helps us to understand how the computer tutoring results might change as speech and natural language processing technologies continue to improve.

We thus conducted two experiments comparing *typed* and *spoken* tutoring dialogues. One experiment used an experienced *human* tutor who communicated with students either via speech or typing.

The other used the Why2-Atlas *computer* tutoring system (VanLehn et al., 2002) with either its original typed dialogue or a new spoken dialogue user interface. The new spoken dialogue system is called ITSPOKE (Litman & Silliman, 2004). Both experiments used qualitative physics as the task domain, similar pretests and posttests, and similar training sequences. The experiments were designed to test whether spoken interactions would yield better learning gains than typed dialogues, whether different dialogue characteristics would be predictive of learning in spoken versus typed dialogues, and whether our findings would generalize across human and computer tutoring.

This paper begins by reviewing the literature on both the potential benefits of spoken dialogue tutoring and previous studies of what aspects of dialogue accelerate learning. Next, we describe the common aspects of both our human and computer tutoring experiments: the task domain, the user interface, etc. Included in this discussion are brief descriptions of Why2-Atlas and ITSPOKE. Finally, the results of our two experiments are presented. Our results show that while in human tutoring, changing the modality from text to speech caused improvements in student learning and dialogue efficiency, in computer tutoring it made less difference. However, in both human and computer tutoring, we find that changing the modality caused differences in superficial dialogue characteristics, and differences in the type of dialogue characteristics that correlate with learning. In sum, while our results suggest that there are indeed potential payoffs for adding speech to text-based dialogue tutors, more research is still needed to fully achieve this potential. We conclude with a more general discussion, followed by our conclusions and current research directions.

## MOTIVATION

In this section we review the literature regarding the two major questions addressed in this paper: what are the potential benefits of using a spoken rather than a typed modality, and what aspects of dialogue accelerate learning. We first review the research on modality differences from the perspective of several communities: the dialogue tutoring community, the computer supported cooperative work community, and the spoken dialogue community. We then discuss previous approaches to investigating what aspects of dialogue accelerate learning.

### The Role of Speech in Tutoring Dialogues

Tutorial dialogue is a natural way to provide students with a learning environment that exhibits characteristics that have been shown to correlate with student learning gains, such as student activity. Thus, as natural language dialogue technology has improved over the years, the development of *computational tutorial dialogue systems* has also become an increasingly active research area (Rosé & Freedman, 2000; Rosé & Aleven, 2002; Aleven & Rosé, 2003). While most current systems are text-based (Evens et al., 2001; Zinn et al., 2002; Aleven et al., 2001; VanLehn et al., 2002), with recent advances in speech technology, several research groups have started developing speech-based natural language dialogue tutors. For example, there are now “talking head” tutors that use spoken language output (Graesser et al., 1999; Rickel & Johnson, 2000), as well as (typically non-animated) dialogue tutors that both accept spoken input and generate spoken output (Schultz et al., 2003; Mostow & Aist, 2001). However, while it has been hypothesized that such additions of speech technology will promote learning gains (due to reasons described below), little empirical work has actually investigated whether and how spoken language capabilities should be added to dialogue-based intelligent tutoring systems.

In particular, how important are *spoken* dialogue interactions in natural tutoring situations? As noted above, three main benefits of spoken tutorial dialogue with respect to increasing learning have been hypothesized. First, spoken dialogue may elicit more student engagement and knowledge construction. Chi et al. (1994, 2001) found that spontaneous and prompted self-explanation improves learning gains during human-human tutoring. However, when such studies (which involved spoken dialogue between human tutors and students) were repeated with the participants communicating via typed text, content-free prompting did not cause much increase in self-explanation or learning (Hausmann & Chi, 2002). Instead, the student typed in paraphrases of the text. Perhaps typing requires additional cognitive capacity and thus reduces the cognitive resources available for spontaneous self-explanation, or students preferred the "safety" of a paraphrase when using the typing modality. Regardless of the cause, the finding itself suggests that the benefits obtained from using prompting and open questions in a computer dialogue system might be easier to achieve in spoken rather than typed interactions. Self-explanation is just one form of student cognitive activity that is known to cause learning gains (Chi et al., 1994; Renkl, 1997; Chi et al., 2001). If it can be increased by using speech, perhaps other beneficial thinking can also be elicited as well.

A second hypothesis is that speech allows tutors to infer a more accurate student model, including long-term factors such as overall competence and motivation, and short-term factors such as whether the student really understood the tutor's utterance. Having a more accurate understanding of the students should allow the tutor to adapt the instruction to the student so as to accelerate the student's learning. For instance, while human tutors may not always choose to tailor their instruction to the individual characteristics of the knowledge state of their students, tutors who ignore signs of student confusion may run the risk of preventing learning (Wood et al., 1978; Chi, 1996). More recently, Siler (2004) has shown that human tutors with a better understanding of their tutees do indeed produce larger learning gains, and this occurs with both spoken and typed natural language dialogues. In particular, spoken dialogue tutors and typed dialogue tutors developed equally accurate assessments of their students' concept mastery, competence, motivation and confidence, perhaps because the spoken dialogue tutors exchanged more words but the typed dialogue tutors had longer to reflect on their communications. Both sets of tutors produced larger learning gains from tutees than tutors who had just met their tutees for the first time. However, this advantage disappeared after only a few minutes of tutoring, suggesting that tutors in both spoken and typed modalities rapidly acquire assessments of their students and increase their effectiveness. In addition, we have shown in other work that the prosodic and acoustic information of speech can improve the detection of student states such as confusion (Litman & Forbes-Riley, 2004), which may be useful for adapting tutoring to the student. There has been increasing interest in developing more affectively-aware systems throughout the tutorial dialogue systems community (Aist et al., 2002a; Craig & Graesser, 2003; Bhatt et al., 2004; Johnson et al., 2004; Moore et al., 2004; Craig et al., 2004).

Third, recent studies of artificial pedagogical agents have also suggested that both voice and dialogue are crucial components of effective interactions. Much of this research is based on the hypothesis that learning will be enhanced in computational environments that prime a more social interpretation of the teaching situation (as when an animated agent talks, and responds contingently to a learner). In a discovery environment for teaching plant design, when an interactive agent's words are conveyed using speech rather than text, student retention, transfer and interest increase; in contrast, the visual presence or absence of the agent image does not impact performance (Moreno et al., 2001). However, while Graesser et al. (2003) found that the use of a dialogue agent improved learning, there was no evidence that output media impacted learning. More recent work suggests that not only the presence or absence of an agent's spoken voice, but also the nature of the voice (e.g., whether the voice is machine-generated using a text-to-speech system, or a human voice that has been pre-recorded), can impact learning.[1] In experiments in both laboratory and school settings using a computer learning environment for teaching math, a human voice is preferable even when the agent is animated: students learn more deeply compared to when a machine-generated voice is used (Atkinson et al., 2005). As the authors note, however, future work should investigate how this finding might change as machine-generated voices improve, and/or if students are first given practice listening to machine-generated voices.

Although not directly in the area of dialogue tutoring, research in other computational learning environments has also investigated potential advantages of alternative modalities. For example, several experiments have shown that when monologue is combined with viewing graphical information, speech elicits larger learning gains than text (see Mayer (2002) for a summary). Note, however, that we are interested in dialogues rather than monologues, and we are not interested in having students split their attention between dialogues and graphics.

## The Role of Speech in Computer-Mediated Interactions

Within the Computer Supported Cooperative Work (CSCW) community, there is an extensive literature comparing the effects of differences in interaction modality. Modalities that have been commonly compared include face-to-face, video conferencing, synchronous or asynchronous email or newsgroup style interactions, and text-based chat with or without text-to-speech augmentation. Within the CSCW realm, what is most closely related to our work on tutorial dialogue is the large body of distance education literature, since it is concerned with both issues of communication effectiveness as well as learning, and the interaction between the two. However, while concerns of instructional effectiveness are at the forefront of our desiderata, in many published on-line learning studies learning gains are not formally evaluated. Other evaluation measures that are often used include coherence, volume, or depth of interaction, ability to form a consensus opinion or to coordinate on intended meaning, student motivation, identification with the learning community, feeling of copresence, satisfaction with the interaction, level of formality or likelihood of anti-social behavior, and longevity or frequency of voluntary participation.

With respect to the speech versus text question addressed in this article, in the CSCW community, neither of these modalities has consistently demonstrated clear advantages across the range of evaluation metrics noted above. Some evidence from the computer-mediated communication literature points to

[1] With respect to other measures besides learning, in the domain of instructional planning, students rate both visual and non-visual agents as more engaging and human-like when audio recordings of a human voice are used (Baylor et al., 2003). Student motivation also increases when the human voice is used with the non-visual version of the agent. However, with the visual agent, the machine-generated voice increases motivation.

advantages of text-based interaction over speech, due to the fact that the history of the dialogue is easy to access during the ongoing conversation. Herring (1999) has demonstrated that on-line communication leads to less coherent interactions than one typically finds in speech interactions, but that this reduction in coherence does not seem to lead to a decrease in satisfaction. The permanence of the conversational record compensates, and in fact, people seem to find new strategies of interaction that are not possible in other settings. Gergle et al. (2004) have similarly demonstrated an advantage for communication effectiveness resulting from displaying the discourse context during a dialogue interaction. Other evidence points to advantages of speech-based interaction. Jensen et al. (2000) have demonstrated that communication modality has a significant impact on cooperation and trust between interacting participants, with natural voice being significantly better than text-based interaction, but synthesized voice not being reliably better than text-based interaction.

Besides the difference in evaluation metrics, another difference from our research is that the CSCW work is primarily concerned with computer-mediated human-human interaction, while we are also interested in communication between humans and computer agents. It is thus unclear as to what extent the findings from computer-mediated communication can be directly applied to the design of dialogue computer tutors. For example, while asynchronous interaction has been shown to increase the volume and depth of participation in on-line communication (Albrektson, 1995; Newlands & McKean, 1996), these benefits may or may not carry into the context of interaction between humans and dialogue systems since humans interacting with a system may not feel obligated to respond to the agent in a timely fashion. As another example, Chester & Gwynne (1998) have studied how participants in computer-mediated communication are less hampered by social conventions, e.g., participants have the opportunity to explore identity issues that would not be possible to explore in a less anonymous environment. In conversations with computer agents, it is not clear that students would be motivated to engage in such explorations.

Nevertheless, despite such differences, we believe that observations from the CSCW community could potentially lend insights for providing a larger and richer context in which to evaluate experimental results from the dialogue tutoring community.

## Other Benefits of Spoken Dialogue Systems

Investigations of human-computer dialogue interactions outside the area of educational applications suggest that speech might yield other benefits in addition to those associated with learning. Spoken language is the most natural and easy to use form of human natural language interaction, and preliminary evidence suggests that spoken rather than typed dialogue might be a preferred modality in computer interactions as well. A study of keyboard versus spoken input in a task-oriented dialogue system evaluated the effect of input modality on task success and user preference (Allen et al., 1996). Subjects interacted with two versions of the same system, where just the input modality (speech or keyboard) differed. Even though keyboard input (with typos) is much less error prone than speech (with speech recognition errors), both input modalities yield the same level of task performance, with speech input being more efficient. In addition, subjects prefer using speech when given the ability to choose. An experiment using a multi-modal data-retrieval system also shows a user preference for speech over keyboard input, in this case despite the fact that speech is less efficient (Rudnicky, 1994).

A second advantage of speech is that the hands-free aspect of spoken versus typed dialogues will extend the applicability of computer tutoring dialogue systems to new domains, such as those where multi-modal (e.g., dialogue systems which involve parallel talking and pointing or clicking) and/or audio-only capabilities (e.g., training systems for use in space (Aist et al., 2002b)) are more crucial.

## Dialogue Features that Predict Learning

Even if it is shown that building a spoken dialogue tutoring system does have the potential to increase learning, when it comes time to actually implement a system, many design choices must be made that will likely influence the style of interaction, which in turn may influence a student's ability to learn from the system. Since it is not yet well understood how dialogue system design choices impact student learning, recent work has begun to try to determine what characteristics of tutoring dialogues positively correlate with learning gains, in order to put system building on a more empirical basis. In the computational community, there has been particular interest in using "shallow" features to characterize dialogue behaviors, as such features have the potential of being automatically computable by a tutoring system as it is operating.[2] The studies described below, for example, have used features such as turn length, percentages of words and turns, etc. In our work, we are particularly interested in learning whether a modality change from text to speech will cause the dialogue features that correlate with learning to also change. Understanding such differences is a prerequisite for constructing spoken language tutors that can engage students in the types of dialogues most likely to yield learning gains.

Rosé et al. (2003) hypothesize that if a tutor is responding directly to a student's revealed knowledge state, the effectiveness of the tutor's instruction should increase as average student turn length increases, as longer student answers to tutor questions reveal more of a student's reasoning. They indeed find a correlation between average length of student turns (in terms of number of words) and learning, in a corpus of typed human-human conceptual physics tutoring dialogues. This result complements findings in a corpus of typed human-human basic electricity and electronics dialogues (Core et al., 2003), which examines student language production relative to tutor language production. In particular, the percentage of words uttered by the student and the percentage of utterances produced by the student (as well as the percentage of tutor utterances that are questions) positively correlate with learning. Again, learning correlates with increased student language production.

These results also complement results found in physics post-solution, reflective dialogues (Katz et al., 2003). With respect to turn length, the more students say in response to reflection questions after a solution has been reached, the more they learn. This result is statistically significant when response length is measured using average number of words, and shows a trend when average time is instead measured. However, these results are based on an analysis of student responses that are followed by canned rather than human natural language dialogue feedback. A related study, which does analyze human rather than canned post-solution reflective dialogues, finds that the following dialogue characteristics correlate with learning: the number of post-solution dialogues, the number of such dialogues that abstract from the particular problem, and the number of such abstraction dialogues that are initiated by the tutor. Note that

[2] The use of "deeper" features requiring manual coding (e.g, distinguishing between substantive contributions and groundings, between types of tutor moves such as scaffolding and explaining, etc. (Chi et al., 2001)) will be discussed below.

more student initiative does not lead to more learning, which is also a finding in Core et al. (2003). It remains to be tested whether the "post" or "reflection" part of such dialogues are more important (Katz et al., 2003), and thus whether the findings will generalize to problem solving (pre-solution) reflective dialogues.

## THE COMMON ASPECTS OF THE EXPERIMENTS

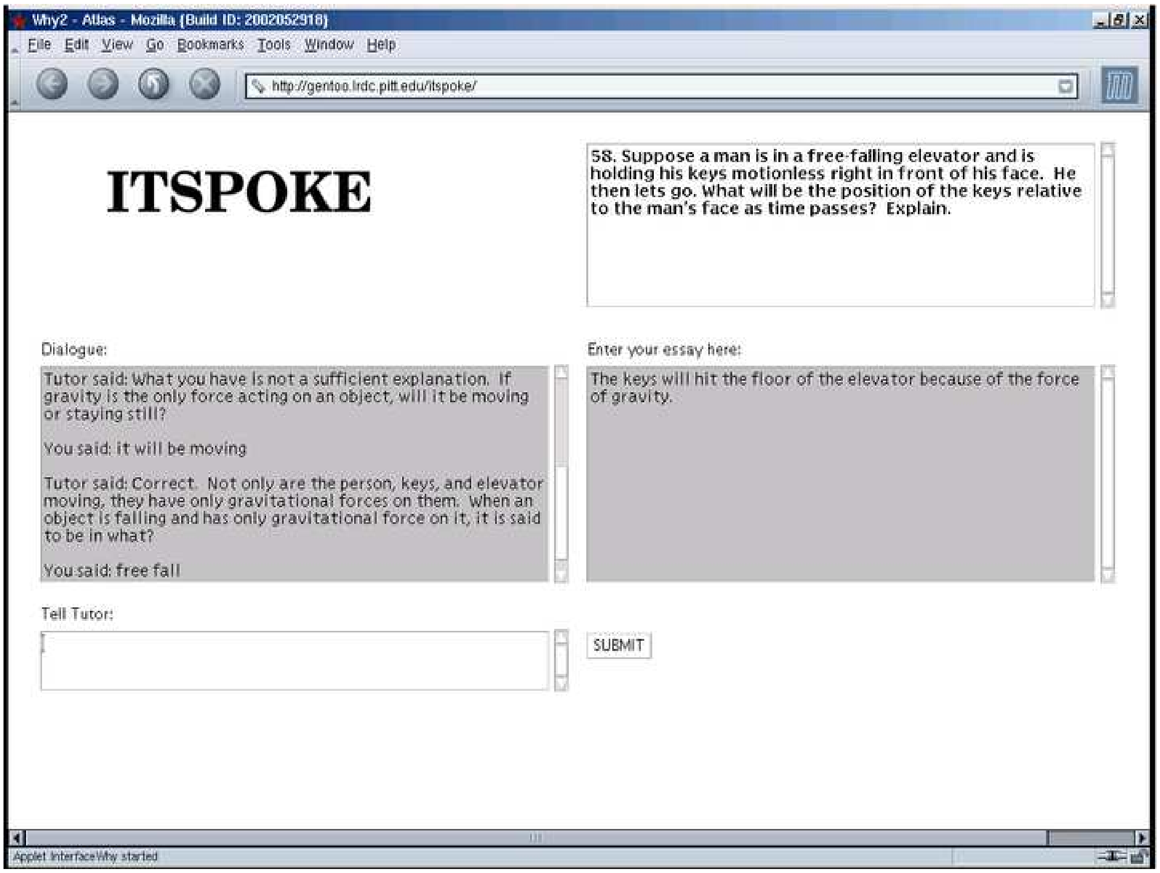
In both our human and computer tutoring experiments of typed versus spoken dialogue tutoring, the students learned how to solve qualitative physics problems, which are physics problems that can be answered without doing any mathematics. A typical problem is, "If a massive truck and a lightweight car have a head-on collision, and both were going the same speed initially, which one suffers the greater impact force and the greater change in motion? Explain your answer." The answer to such a problem is a short essay. A correct answer to this question should mention Newton's third law, which states that when one object, such as the truck, exerts a force on another object, such as the car, then the forces have the same magnitude. The answer should also mention Newton's second law, which implies that when two objects are acted on by forces of the same magnitude, their change in motion (acceleration) is inversely proportional to their mass, so a more massive object (e.g, the truck) will have a smaller acceleration than a lighter object (e.g., the car).

The experimental procedure was as follows. Students who have not taken any college physics were first given a pretest measuring their knowledge of physics. Next, students read a short textbook-like pamphlet, which described the major laws (e.g., Newton's first law) and the major concepts. Students then worked through a set of up to 10 training problems with the tutor. Finally, students were given a posttest that was similar to the pretest. The pre and post tests each included 40 multiple choice questions, and were isomorphic (that is, the problems on each test differed only in the identities of the objects (e.g., cars versus trucks) and other surface features that should not affect the reasoning required to solve them). The tests also included essay questions, but they did not turn out to be sensitive to learning, so they we will not be discussed further here (see VanLehn et al. (submitted) for details). The entire experiment took no more than 9 hours per student, and was usually performed in 1-3 sessions of no more than 4 hours each. Subjects were university students responding to ads, and were compensated with money or course credit.

The interface used for all experiments was a variant of that shown in Figure 1, which is a screenshot generated during an ITSPOKE interaction.

The student first typed an essay answering a qualitative physics problem, as in the middle and upper right of Figure 1. The tutor then engaged the student in a natural language dialogue to provide feedback, correct misconceptions, and to elicit more complete explanations (as shown in the dialogue window). At key points in the dialogue, the tutor asked the student to revise the essay. This cycle of instruction and revision continued until the tutor was satisfied with the student's essay, at which point the tutor presented the ideal essay answer to the student.

For the studies described below, we compare characteristics of student dialogues with both typed and spoken computer tutors (Why2-Atlas and ITSPOKE, respectively), as well as with a single human tutor performing the same task as the computer tutor for each system. Why2-Atlas is a *text-based* intelligent


Why2 - Atlas - Mozilla {Build ID: 2002052918}
File Edit View Go Bookmarks Tools Window Help
http://gentoo.lrdc.pitt.edu/itspoke/
ITSPOKE
Dialogue:
Enter your essay here:
Tell Tutor:
SUBMIT
Applet InterfaceWhy started

58. Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

Tutor said: What you have is not a sufficient explanation. If gravity is the only force acting on an object, will it be moving or staying still?

You said: it will be moving

Tutor said: Correct. Not only are the person, keys, and elevator moving, they have only gravitational forces on them. When an object is falling and has only gravitational force on it, it is said to be in what?

You said: free fall

The keys will hit the floor of the elevator because of the force of gravity.

Fig.1. Screenshot during ITSPOKE Human-Computer Spoken Dialogue

tutoring dialogue system (VanLehn et al., 2002), developed in part to test whether deep approaches to natural language processing elicit more learning than shallower approaches. A suite of natural language processing components are provided by a Why2-Atlas toolkit (e.g., sentence-level syntactic and semantic analysis (Rosé, 2000), discourse and domain level processing (Jordan et al., 2003), and a finite-state dialogue manager (Rosé et al., 2001)). ITSPOKE (Intelligent Tutoring SPOKEn dialogue system) (Litman & Silliman, 2004) is a *speech-enabled* version of Why2-ATLAS. Student speech is digitized from microphone input and sent to the Sphinx2 recognizer (Huang et al., 1993). ITSPOKE uses 56 dialogue-state dependent stochastic language models for speech recognition (initially trained using 4551 student utterances from a 2002 evaluation of Why2-Atlas, then later enhanced using utterances from ITSPOKE's pilot testing). The most probable "transcription" output by Sphinx2 is sent to the Why2-Atlas natural language processing "back-end". Finally, the text response produced by Why2-Atlas is sent to the Cepstral text-to-speech system (a commercial outgrowth of the Festival system (Black & Taylor, 1997)), then played to the student through a headphone.

# HUMAN-HUMAN TUTORING: EXPERIMENT 1

## Experimental Procedure

Experiment 1 compared typed and spoken tutoring, using the same *human* tutor in both conditions. In the typed condition, the interaction was in the form of a *typed* dialogue between the student and tutor, where the human tutor performed the same task that Why2-Atlas was designed to perform. A text-based chat web interface (as in Figure 1) was used, with student and tutor in separate rooms; students knew that the tutor was human. A sample typed dialogue excerpt is displayed in Figure 2.

**PROBLEM:** Suppose that you released 3 identical balls of clay in a vacuum at exactly the same instant. They would all hit the ground at the same instant. Now you stick two of the balls together, forming one ball that is twice as heavy as the remaining, untouched clay ball. Both balls are released in a vacuum at exactly the same instant. Which ball hits the ground first?

**ESSAY:** Both balls will hit the ground at the same time. The balls are in free fall (only gravitational forces). The ratio of the masses and weight are equal.

**. . . excerpt from 2.0 minutes into** *typed* **dialogue . . .**

**Tutor:** You have correctly said the balls are in free fall. What do you conclude from this fact?

**Student:** There is no air resistance and the balls' accelerations are constant in regards to one another

**Tutor:** Right, so the conclusion is that they both have the same acceleration, the free fall acceleration. Now you have to show that time taken by them to reach the ground will be the same. How do you do that?

**Student:** F (net forces) / mass = f/m because F = m*a therefore ratio should equal one another

**Tutor:** But that does not tell you anything about the time taken. What physical quantities will determine the time taken. Will the initial velocity of the balls matter?

**Student:** yes, assuming starting at rest? sorry, kind of at a loss at the moment

**Tutor:** The initial velocity will matter, and here they both have the same initial velocity, zero. What else, will the height through wich (sic) they fall matter?

**Student:** not so much as long as they are both dropped from the same height

**Tutor:** Height of fall will determine the time of fall, and here the height of fall is the same for both the balls. Now on the basis of these facts can you conclude about the time taken by them to reach the ground?

**Student:** since initial velocities are equal and the height of falls are equal, then the time taken is equal

**Tutor:** How about acceleration, if they had different accelerations, even if they fall through the same height and have same initial velocity , will they reach the ground at the same time?

Fig.2. Excerpt from Human-Human Typed Dialogue

**PROBLEM:** Suppose that you released 3 identical balls of clay in a vacuum at exactly the same instant. They would all hit the ground at the same instant. Now you stick two of the balls together, forming one ball that is twice as heavy as the remaining, untouched clay ball. Both balls are released in a vacuum at exactly the same instant. Which ball hits the ground first?

**ESSAY:** The balls will hit the ground at exactly the same instant.

**. . . excerpt from 8.3 minutes into *spoken* dialogue . . .**

**Tutor:** So the accel- the- both the balls will have same acceleration G is that right?

**Student:** Yeah.

**Tutor:** Ok. So now, can you uh argue about the time?

**Student:** Yeah. Uh the reason the two balls hit the ground at exactly the same instant is because the net forces acting on it, which is acceleration due to gravity, are equal.

**Tutor:** Force is gravity and acceleration is the same, ok, and what else?

**Student:** Um -(sniff)

**Tutor:** What other factors determine the time taken are the same?

**Student:** Uh, the height at which they're dropped -

**Tutor:** mm-hm

**Student:** Uh, the the acceleration due to gravity. Uh -

**Tutor:** What other factor did we say will determine time taken for a body to fall to the ground? Height is one factor acceleration is another.

**Student:** And -

**Tutor:** Initial velocity, isn't it?

**Student:** Ok, yeah, starting from rest it's initial velocity. That's right.

**Tutor:** Right. Now are all these factors same for these two balls?

**Student:** Are the factors all the same?

**Tutor:** Uh-huh.

**Student:** Yeah.

**Tutor:** Ok so that is the reason why they will take the same time to hit the ground, and they were dropped at the same instant so they will hit the ground also at the same instant.

Fig.3. Excerpt from Human-Human Spoken Dialogue

In the spoken condition, the interaction was in the form of a *spoken* dialogue, where the human tutor performed the same task that ITSPOKE was designed to perform. (While the dialogue was changed to speech, students still typed the essay.) The tutor and student spoke through head-mounted microphones, allowing all speech to be digitally recorded to the computer. The student and tutor were in the same room (due to constraints of speech recording), but separated by a partition. The same web interface was used as in the typed condition, except that no dialogue history was displayed (this would have required manual transcription of utterances). In contrast to the typed condition, where strict turn-taking was enforced, interruptions and overlapping speech were permitted in the spoken condition. This was because we plan to add "bargein" to ITSPOKE, which will enable students to interrupt ITSPOKE. A sample dialogue excerpt from the spoken human tutoring condition is displayed in Figure 3. Note that turns ending in "-" indicate speech overlapping with the following turn.

The same human tutor was used in both conditions. The tutor was a retired physics professor who had logged hundreds of hours tutoring students in a set of experiments preceding this study (VanLehn et al., submitted). He was thus quite familiar with the topics, students, and experimental set up used in our studies. The tutor was instructed to cover the expectations for each problem, to watch for the specific set of expectations and misconceptions associated with the problem, and to end the discussion of each problem by showing the ideal essay to the student. He was encouraged to avoid lecturing the student and to attempt to draw out the student's own reasoning. He knew that transcripts of his tutoring would be analyzed. Nevertheless, he was not required to follow any prescribed tutoring strategies.

Pre and posttest items were scored as right or wrong, with no partial credit. Students who were not able to complete all 10 problems due to lack of time took the posttest after only working through a subset of the training problems.

Experiment 1 resulted in two human tutoring corpora. The *typed* dialogue corpus consists of 171 physics problems with 20 students, while the *spoken* dialogue corpus consists of 128 physics problems with 14 students. In subsequent analyses, a "dialogue" refers to the transcript of one student's discussion of one problem with the tutor.

## Results

Table 1 presents the means and standard deviations for two types of analyses, learning and training time, across conditions. Based on the literature discussed earlier, we hypothesized that both of these analyses would show higher levels of performance in our spoken as compared to our typed dialogue tutoring conditions.

With respect to learning, the pretest scores were not reliably different across the two conditions, F(1,32) = 1.574, p = 0.219, MSe = 0.009. In an ANOVA with condition by test phase factorial design, there was a robust main effect for test phase, F(1,66) = 90.589, p = 0.000, MSe = 0.012, indicating that students in both conditions learned a significant amount during tutoring. However, the main effect for condition was not reliable, F(1,32) = 1.823, p = 0.186, MSe = 0.014, and there was no reliable interaction. In an ANCOVA, the adjusted posttest scores (where pretest score was factored out) showed a strong trend of being reliably different across conditions, F(1,31)=4.044, p=0.053, MSe = 0.01173. Our results thus suggest that the human speech tutored students learned more than the human text tutored students; the effect size was 0.74.

Table 1
Learning and Time: Human Tutoring Spoken (14) and Typed (20) Conditions

| Dependent Measure | Human Spoken | Human Typed |
|---|---|---|
| Pretest Mean (standard deviation) | .42 (.10) | .46 (.09) |
| Posttest Mean (standard deviation) | .72 (.11) | .67 (.13) |
| Adjusted Posttest Mean (standard deviation) | .74 (.11) | .66 (.11) |
| Dialogue Time (standard deviation) | 166.58 (45.06) | 430.05 (159.65) |

With respect to training time, students in the spoken condition completed their dialogue tutoring in less than half the time than in the typed condition, where dialogue time was measured as the sum over the training problems of the number of minutes between the time that the student was shown the problem text and the time that the student was shown the ideal essay. The extra time needed for both the tutor and the student to type (rather than speak) each dialogue turn in the typed condition was a major contributor to this difference. An ANOVA shows that the difference in means across the two conditions was reliably different, with F(1,32) = 35.821, p = 0.00, MSe = 15958.787.

In sum, for human tutoring, our results thus support our hypothesis that spoken tutoring is indeed more effective than typed tutoring, for both learning and training time.

It is important to understand why the change in modality (and interruption policy) increased learning. Table 2 presents the means for a variety of measures characterizing different aspects of dialogue, to determine which aspects differ across conditions, and to examine whether different dialogue characteristics correlate with learning across conditions (although the utility of correlation analysis might be limited by our small subject pool). For each dependent measure (explained below), the second through fourth columns present the means (across students) for the spoken and typed conditions, along with the statistical significance of their differences. The fifth through eighth columns present a Pearson's correlation between each dialogue measure and raw posttest score. However, in the spoken condition, the pre and posttest scores are highly correlated (R=.72, p =.008); in the typed condition they are not (R=.29, p=.21). Because of the spoken correlation, the last four columns show the correlation between posttest and the dependent measure, after the correlation with pretest is regressed out.

The measures in Table 2 were motivated by previous work suggesting that learning correlates with increased student language production, as discussed above. In pilot studies of the typed corpus, average student turn length was found to correlate with learning. We thus computed the average length of student turns in words (Ave. Stud. Wds/Turn), as well as the total number of words and turns per student, summed across all training dialogues (Tot. Stud. Words, Tot. Stud. Turns).[3] We also computed these figures for the tutor's contributions (Ave. Tut. Wds/Turn, Tot. Tut. Words, Tot. Tut. Turns). The slope and intercept measures will be explained below. Similarly, the studies of Core et al. (2003) examined

[3] In the spoken data, turn boundaries were manually annotated by a paid transcriber. The transcriber added a turn boundary when: 1) the speaker stopped speaking and the other party in the dialogue began to speak, 2) the speaker asked a question and stopped speaking to wait for an answer, 3) the other party in the dialogue interrupted the speaker and the speaker paused to allow the other party to speak.

Table 2
Dialogue Aspects & Learning: Human Spoken (14) & Typed (20) Conditions

| Dependent Measure | Spoken mean | Typed mean | p | Zero Order Correlations | | | | Controlled for Pre-Test Correlations | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Spoken | | Typed | | Spoken | | Typed | |
| | | | | R | p | R | p | R | p | R | p |
| Tot. Stud. Words | 2322.43 | 1569.30 | .03 | -.473 | .09 | .065 | .78 | -.261 | .39 | .013 | .96 |
| Tot. Stud. Turns | 424.86 | 109.30 | .00 | -.340 | .24 | -.148 | .53 | -.016 | .96 | -.213 | .38 |
| Ave. Stud. Wds/Turn | 5.21 | 14.45 | .00 | -.167 | .57 | .491 | .03 | -.209 | .49 | .515 | .03 |
| Slope: Stud. Wds/Trn | -.01 | -.05 | .04 | -.275 | .34 | -.375 | .10 | .379 | .20 | -.291 | .23 |
| Int: Stud. Wds/Trn | 6.51 | 16.39 | .00 | -.176 | .55 | .625 | .00 | -.441 | .13 | .593 | .01 |
| Tot. Tut. Words | 8648.29 | 3366.30 | .00 | -.482 | .08 | .027 | .91 | -.164 | .59 | -.034 | .89 |
| Tot. Tut. Turns | 393.21 | 122.90 | .00 | -.436 | .12 | -.171 | .47 | -.110 | .72 | -.239 | .32 |
| Ave. Tut. Wds/Turn | 23.04 | 28.23 | .01 | -.139 | .64 | .496 | .03 | -.086 | .78 | .536 | .02 |
| S-T Tot. Wds Ratio | .27 | .45 | .00 | .067 | .82 | .275 | .24 | -.202 | .51 | .268 | .27 |
| S-T Wd/Trn Ratio | .25 | .51 | .00 | .026 | .93 | .283 | .23 | -.237 | .44 | .277 | .25 |

student language production relative to tutor language production, and found that the percentage of words and utterances produced by the student positively correlated with learning. This led us to compute the number of students words divided by the number of tutor words (S-T Tot. Wds Ratio), and a similar ratio of student words per turn to tutor words per turn (S-T Wd/Trn Ratio).

Table 2 shows interesting differences between the spoken and typed corpora of human-human dialogues. For every measure examined, the means across conditions are significantly different, verifying that the style of interactions is indeed quite different. In spoken tutoring, both student and tutor take more turns on average than in typed tutoring, but these spoken turns are on average shorter. Moreover, in spoken tutoring both student and tutor on average use more words to communicate than in typed tutoring. However, in typed tutoring, the ratio of student to tutor language production is higher than in speech.

The remaining columns attempt to uncover which aspects of tutorial dialogue in each condition were responsible for its effectiveness. Although the zero order correlations are presented for completeness, our discussion will focus only on the last four columns, which we feel present the more valid analysis.

In the typed condition, as in its earlier pilot study, there is a positive correlation between average length of student turns in words and learning (R=.515, p = .03). We hypothesize that longer student answers to tutor questions reveal more of a student's reasoning, and that if the tutor is adapting his interaction to the student's revealed knowledge state, the effectiveness of the tutor's instruction might increase as average student turn length increases. Note that there is no correlation between total student words and learning; we hypothesize that how much a student explains (as estimated by turn length) is more important than how many questions a student answers (as estimated by total word production). There is also a positive correlation between average length of tutor turn and learning (R=.536, p=.02). Perhaps more tutor words per turn means that the tutor is explaining more or giving more useful feedback. A deeper coding of our data would be needed to test all of these hypotheses. Finally, as in the typed pilot study (Rosé et al., 2003), student words per turn usually decreased gradually during the sessions. In speech, turn length decreased from an average of 6.0 words/turn for the first problem to 4.5 words/turn by the last problem. In text, turn length decreased from an average of 14.6 words for the first problem to 10.7 words by the last problem. This led us to fit regression lines to each subject and compare the intercepts and slopes to learning. These measures indicate roughly how verbose a student was initially and how quickly the student became taciturn. Table 2 indicates a reliable correlation between intercept and learning (R=.593; p=.01) for the typed condition, suggesting that inherently verbose students (or at least those who initially typed more) learned more in typed human dialogue tutoring.

Since there were no significant correlations in the spoken condition, we have begun to examine other measures that might be more relevant in speech. For example, the mean number of total syntactic questions per student is 35.29, with a trend for a negative correlation with learning (R=-.500, p=.08). This result suggests, that as with our text-based correlations, our current surface level analyses (which had the advantage of being automatically computable from the transcriptions) will need to be enhanced with deeper codings before we can fully interpret our results (e.g., by manually coding non-interrogative form questions, and by distinguishing question types).

In sum, our results suggest that a change in modality influences the dialogue features found in human tutoring. However, it is still an open question as to how such differences might explain our finding that spoken dialogue is superior to text-based dialogue with respect to learning. That is, while our results demonstrate differences across our typed and spoken conditions, further experimentation is still needed to understand why these differences cause higher learning gains in our spoken condition.

## HUMAN-COMPUTER TUTORING: EXPERIMENT 2

### Experimental Procedure

Experiment 2 compared typed and spoken tutoring using the Why2-Atlas and ITSPOKE *computer* tutors, respectively. The experimental procedure was the same as for Experiment 1, except that students worked through only 5 physics problems, and the pretest was taken after the background reading (allowing us to measure gains caused by the experimental manipulation, without confusing them with gains caused by background reading). Strict turn-taking was now enforced in both conditions as bargein had not yet been implemented in ITSPOKE.

While Why2-Atlas and ITSPOKE used the same web interface, during the dialogue, Why2-Atlas students typed while ITSPOKE students spoke through a head-mounted microphone. In addition, the Why2-Atlas dialogue history contained what the student actually typed, while the ITSPOKE history contained the potentially noisy output of ITSPOKE's speech recognizer. The speech recognizer's hypothesis for each student utterance, and the tutor utterances, were not displayed until after the student or ITSPOKE had finished speaking.

A sample Why2-Atlas dialogue excerpt is shown in Figure 4, while a sample ITSPOKE dialogue excerpt is shown in Figure 5. Note that for ITSPOKE, the output of the automatic speech recognizer (the ASR annotations) sometimes differed from what the student actually said. Thus, ITSPOKE dialogues contained rejection prompts (when ITSPOKE was not confident of what it thought the student said, it asked the student to repeat, as in the third ITSPOKE turn). On average, ITSPOKE produced 1.4 rejection prompts per dialogue. ITSPOKE also misrecognized utterances; when ITSPOKE heard something different than what the student said but was confident in its hypothesis, it proceeded as if it heard correctly. While the ITSPOKE word error rate was 31.2%, semantic analysis based on speech recognition versus perfect transcription differed only 7.6% of the time. Semantic accuracy is more relevant for dialogue evaluation, as it does not penalize for unimportant word errors.

Experiment 2 resulted in two computer tutoring corpora. The *typed* Why2-Atlas dialogue corpus consists of 115 problems (dialogues) with 23 students, while the ITSPOKE *spoken* corpus consists of 100 problems (dialogues) with 20 students.

**PROBLEM:** Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

**ESSAY:** The keys will be in front of his face, because he and his keys will be moving in the same direction.

**. . . excerpt from 3.0 minutes into *typed* dialogue . . .**

**Why2-Atlas:** You have answered the problem correctly except that what you have in your essay is not a sufficient explanation. Before the release, what is the relationship between the man's velocity and the keys' velocity?

**Student:** They are the same

**Why2-Atlas:** Uh huh. If gravity is the only force acting on an object, will it be moving or staying still?

**Student:** Moving

**Why2-Atlas:** Good. Not only are the person, keys, and elevator moving, they have only gravitational forces acting on them. When an object is falling and has only gravitational force on it, it is said to be in what?

**Student:** Free fall

**Why2-Atlas:** Yes. Thus, after the release the only force on the keys and man is the downward force of earth's gravity, so they are in freefall. What force acts on all objects within close proximity to earth?

**Student:** The earths gravity

**Why2-Atlas:** Good. Besides the Earth's gravitational force, are there any other forces acting on the person, keys, and elevator after the release?

**Student:** No

**Why2-Atlas:** Fine. Thus, keep in mind that after the release, the only force on the person, keys, and elevator is the force of gravity. We know that after the release, the keys, the person, and the elevator are in freefall. Is freefall acceleration near earth a constant? Or does it vary depending upon which object is in freefall?

Fig.4. Excerpt from Why2-Atlas Typed Dialogue

**PROBLEM:** Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

**ESSAY:** As time passes, the keys will have gotten higher than the man's face. The reason the keys would have moved, is due to force and direction acting on the keys. The man will have stayed where he was, but the keys, which are smaller in mass, would have been forced upward. Since the direction of the elevator is south, the opposite direction will be forced on the keys, which is north. Therefroe, the keys will have traveled north and have passed the man's face.

**. . . excerpt from 11.1 minutes into** *spoken* **dialogue . . .**

**ITSPOKE:** If you push on the same box with three times the force, what happens to the acceleration of the box?

**Student:** it increases (ASR: it increases)

**ITSPOKE:** Terrific. OK, what is the box's acceleration when you are pushing three times relative to its acceleration when you were only pushing with one-third the force?

**Student:** it decreases (ASR: it decreases since)

**ITSPOKE:** Could you please repeat that?

**Student:** it increases (ASR: it increases)

**ITSPOKE:** Well... The acceleration will be three times the acceleration from before. Given several different forces acting on the box, what must you do to the forces to find the acceleration of the box?

**Student:** have a constant (ASR: out the instant)

**ITSPOKE:** Could you please repeat that?

**Student:** have mass stay the constant (ASR: i have the mass stay added constants)

**ITSPOKE:** That's not right. Are forces vector or scalar quantities?

**Student:** vector (ASR: vector)

Fig.5. Excerpt from ITSPOKE Spoken Dialogue

## Results

Table 3 presents the means and standard deviations for the learning and training time measures previously examined in Experiment 1. The pre-test scores were not reliably different across the two conditions, F(1,41) = 0.037, p= 0.848, MSe = 0.036. In an ANOVA with condition by test phase factorial design, there was a robust main effect for test phase, F(1,84) = 29.57, p = 0.000, MSe = 0.032, indicating that students learned during their tutoring. The main effect for condition was not reliable, F(1,41)=0.029, p=0.866, MSe=0.029, and there was no reliable interaction. In an ANCOVA of the multiple-choice test data, the adjusted post-test scores were not reliably different, F(1,40)=0.004, p=0.950, MSe=0.01806. Thus, the Why-Atlas tutored students did not learn reliably more than the ITSPOKE tutored students.

Table 3
Learning and Time: Computer Tutoring Spoken (20) and Typed (23)

| Dependent Measure | Computer Spoken | Computer Typed |
|---|---|---|
| Pretest Mean (standard deviation) | .48 (.17) | .49 (.20) |
| Posttest Mean (standard deviation) | .69 (.18) | .70 (.16) |
| Adjusted Posttest Mean (standard deviation) | .69 (.13) | .69 (.13) |
| Dialogue Time (standard deviation) | 97.85 (32.8) | 68.93 (29.0) |

With respect to training time, students in the spoken condition took more time to complete their dialogue tutoring than in the typed condition. In the spoken condition, extra utterances were needed to recover from speech recognition errors; also, listening to tutor prompts often took more time than reading them, and students sometimes needed to both listen to, then read, the prompts. An ANOVA shows that this difference was reliable, with F(1,41)=9.411, p=0.004, MSe=950.792.

In sum, while adding speech to Why2-Atlas (in the form of ITSPOKE) did not yield the hoped for improvements in learning, the degradation in tutor understanding due to speech recognition (and potentially in student understanding due to text-to-speech) also did not decrease student learning. In fact, although many ITSPOKE students experienced problems with speech recognition, in other research we have found no correlation between learning and numerous quantitative measures of speech recognition error (including number of rejection prompts, word error rate, and semantic error rate) (Litman & Forbes-Riley, 2005).

Table 4 presents the means for the measures used in Experiment 1 to characterize dialogue, as well as for a new "Tot. Subdialogues per KCD" measure for our computer tutors. A Knowledge Construction Dialogue (KCD) is a line of questioning targeting a specific concept (such as Newton's Third Law). When students answer questions incorrectly, the KCDs correct them through a "subdialogue", which may involve more interactive questioning or simply a remedial statement. Thus, subdialogues per KCD is the number of student responses treated as wrong. We hypothesized that this measure would be higher in speech, due to the previously noted degradation in semantic accuracy.

Compared to Experiment 1, Table 4 shows that there are less differences between spoken and typed *computer* tutoring dialogues. The total words produced by students, the average length of turns and initial verbosity, and the ratios of student to tutor language production are no longer reliably different across

Table 4

Dialogue & Learning: Computer Spoken (20) & Typed (23) Conditions

| Dependent Measure | Spoken mean | Typed mean | p | Zero Order Correlations | | | | Controlled for Pre-Test Correlations | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Spoken | | Typed | | Spoken | | Typed | |
| | | | | R | p | R | p | R | p | R | p |
| Tot. Student Words | 296.85 | 238.17 | .12 | .043 | .86 | -.354 | .10 | .394 | .10 | .050 | .82 |
| Tot. Student Turns | 116.75 | 87.96 | .02 | -.093 | .70 | -.549 | .01 | .210 | .39 | -.168 | .46 |
| Ave. Student Words/Turn | 2.42 | 2.77 | .29 | .061 | .80 | .167 | .45 | .119 | .63 | .202 | .37 |
| Slope: Student Wds/Trn | -.02 | .00 | .02 | -.179 | .45 | -.084 | .70 | -.287 | .23 | -.102 | .65 |
| Intercept: Stud. Wds/Trn | 3.21 | 2.88 | .40 | .246 | .30 | .250 | .25 | .321 | .18 | .281 | .21 |
| Tot. Tutor Words | 6314.90 | 4972.61 | .03 | -.100 | .68 | -.576 | .00 | .283 | .24 | -.159 | .48 |
| Tot. Tutor Turns | 148.20 | 110.22 | .01 | -.061 | .80 | -.529 | .01 | .252 | .30 | -.133 | .56 |
| Ave. Tutor Words/Turn | 42.11 | 44.33 | .06 | -.261 | .27 | -.565 | .01 | .-062 | .80 | -.164 | .47 |
| Stud-Tut Tot. Word Ratio | .05 | .05 | .57 | .219 | .35 | .238 | .27 | .281 | .25 | .201 | .37 |
| Stud-Tut Wds/Trn Ratio | .06 | .06 | .64 | .089 | .71 | .278 | .20 | .094 | .70 | .212 | .35 |
| Tot. Subdial/KCD | 3.29 | 1.98 | .01 | -.304 | .19 | -.732 | .00 | -.018 | .94 | -.457 | .03 |

conditions. As hypothesized, Tot. Subdialogues per KCD is reliably different (p=.01). Finally, the last four columns show a significant negative correlation between Tot. Subdialogues per KCD and posttest score (after regressing out pretest) in the typed condition. We hypothesize that as the argumentation in the dialogue becomes increasingly embedded and thus more complex (due to the tutor's addition of subdialogues when student responses are incorrect), it becomes more difficult for students to learn. There is also a trend for a positive correlation with total student words in the spoken condition, consistent with previous results on learning and increased student language production. Note that although the same measure was examined in Experiment 1, we did not see a similar correlation. We hypothesize that this is due to the fact that in the human tutoring condition, a higher percentage of a student's words correspond to disfluencies (e.g., *um*) and other lexical phenomena not related to physics. As will be discussed below, in future work we plan to determine the impact of coding for such phenomena (for example, by removing the associated words from the total word count).

## DISCUSSION

Our two experiments provide first results in generating an empirically-based understanding of the implications of adding spoken language capabilities to text-based dialogue tutors, and how these implications might differ across human-human and human-computer dialogues. With respect to performance evaluation, our main result was that while changing the modality from text to speech caused improvements in the learning gains and time in human tutoring, for computer tutoring it made less difference. In contrast, with respect to dialogue correlations with learning, our main result was that in both human and computer tutoring, changing the modality from text to speech did cause differences in both the nature of the dialogues (at least as quantified by our shallow measures), and in the set of dialogue characteristics that correlated with learning.

One hypothesis for the lack of improvement in our spoken computer tutoring condition is that simply adding a spoken “front-end”, without also modifying the tutorial dialogue system “back-end”, is either not enough to change how students interact with a computer tutor, or doesn’t exploit that fact that different types of dialogues might be required to accelerate learning in each modality (as suggested by our findings). For example, the same natural language processing components (e.g. sentence level semantic analysis, knowledge construction dialogues) were used in both Why2-Atlas and ITSPOKE, even though these components were originally authored with only the Why2-Atlas text-based system in mind. Another hypothesis is that the limitations of the particular natural language technologies used in Why2-Atlas (or the expectations that the students had regarding such limitations) are inhibiting the modality differences. Finally, if there were differences between conditions, perhaps the shallow measures used in our experiments and/or our small number of subjects prevented us from discovering them. In sum, while the results of human tutoring suggest that spoken tutoring is a promising approach for enhancing learning, more exploration is required to determine how to productively incorporate speech into computer tutoring systems.

By design, the modality change left the content of the computer dialogues completely unchanged – the tutors said nearly the same words and asked nearly the same questions, and the students gave their usual short responses. On the other hand, the content of the human tutoring dialogues probably changed considerably when the modality changed. This suggests that modality change makes a difference in learning only if it also facilitates content change. We will investigate this hypothesis in future work by coding for content and other deep features. For example, active (self-) construction has been shown to enhance student learning in human tutorial contexts, particularly in response to certain tutor moves (Chi et al., 2001); in peer learning contexts, collaboration and interaction have also been shown to enhance student learning. We plan to apply some of the codings used in those studies (e.g., substantive versus non-substantive contributions, type of tutor move, conversation act) to better measure both student self-construction and dialogue interactivity.

In addition, we had hypothesized that the spoken modality would encourage students to become more engaged and to self-construct more knowledge. Although a deeper coding of the dialogues would be necessary to test this hypothesis, we can get a preliminary sense of its veracity by examining the total number of words uttered. Student verbosity (and perhaps engagement and self-construction) did not increase significantly in the spoken computer tutoring experiment. In the human tutoring experiment, the number of student words did significantly increase, which is consistent with the hypothesis and may explain why spoken human tutoring was probably more effective than typed human tutoring. However, the number of tutor words also significantly increased, which suggests that the human tutor may have “lectured" more in the spoken modality. Perhaps these longer explanations contributed to the benefits of speaking compared to the text, but it is equally conceivable that they reduced the amount of engagement and knowledge construction, and thus limited the gains. This suggests that although we considered how the modality might effect the student, we neglected to consider how it might effect the tutor, and how that might impact the students’ learning. Clearly, these issues deserve more research. Our goal is to use such investigations to guide the development of future versions of Why2-Atlas and ITSPOKE, by modifying the dialogue behaviors in each system to best enhance the possibilities for increasing learning.

Finally, note that for each of the spoken and typed modalities, the specific dialogue characteristics that correlated with learning differed in Experiment 1 versus Experiment 2. This suggests that it is unclear to what extent findings regarding effective student and tutor behaviors in human tutoring are directly applicable when designing computational tutorial dialogue systems.

## CONCLUSION AND CURRENT DIRECTIONS

In this paper we presented the results of both human and computer dialogue tutoring experiments, investigating whether adding spoken language capabilities to text-based tutoring yields performance gains and along what metrics, and whether the dialogue characteristics previously shown to correlate with learning gains in text also correlate with increased learning in speech.

The results of Experiment 1 on human tutoring suggest that spoken dialogue (allowing interruptions) is more effective than typed dialogue (prohibiting interruptions), with mean adjusted posttest score increasing and training time decreasing. We also find that typed and spoken dialogues are very different for the surface measures examined, and for the typed condition we see a benefit for longer turns (evidenced by correlations between learning and average and initial student turn length and average tutor turn length). While we do not see these results in speech, spoken utterances are typically shorter than written sentences (Jurafsky & Martin, 2000) (and in our experiment, turn length was also impacted by interruption policy), suggesting that other measures might be more relevant. We are in fact starting to explore the use of more sophisticated measures, as will be described below.

While the results of Experiment 1 offer the hope that a shift to a spoken dialogue modality can yield an increase in the effectiveness of tutorial dialogue technology, the results of Experiment 2 on computer tutoring are less conclusive than expected. On the negative side, we do not see any evidence that replacing typed dialogue in Why2-Atlas with spoken dialogue in ITSPOKE improves student learning. However, on the positive side, we also do not see any evidence that the degradation in understanding caused by speech recognition decreases learning (see also (Litman & Forbes-Riley, 2005)). A similar result showing that speech recognition errors did not decrease learning was found in an evaluation of the SCoT spoken dialogue tutoring system (Pon-Barry et al., 2004). Furthermore, compared to human tutoring, we see fewer stylistic difference between spoken and typed computer dialogue interactions, at least for the dialogue aspects measured in our experiments.

We are currently continuing our studies in several ways. As discussed above, one hypothesis as to why we failed to find some of our predicted differences was that the set of shallow measures used in our experiments just prevented us from discovering them. Recall that in our spoken corpora, our “shallow” ways of characterizing dialogue particularly failed to correlate with learning. We have thus started to investigate the use of measures derived from “deeper” types of dialogue codings (as discussed above), and are indeed starting to find more significant correlations with learning (Forbes-Riley et al., 2005). Similar types of analyses are also starting to be productively used in the Autotutor system (Jackson et al., 2004). Based on the promising results from both projects, we plan to annotate our text-based corpora with such deeper measures, to enable a more sophisticated comparison of correlations across modalities. In our spoken corpora we also plan to investigate whether spoken phenomena such as disfluencies and grounding[4] might also explain the lack of correlations with shallow measures. Such phenomena increase

[4] Example disfluencies in Figure 3 include filled pauses (*uh*, *um*), word repetition (*the the*), and a false start (*So the accel-*

the word count in the spoken condition and generally do not occur in the typed condition. Further analysis is needed to determine the impact of coding for such phenomena, and removing the associated words from the total word count.

We also hypothesized that we might not have seen learning differences across modalities in computer tutoring, because either the limitations of our particular language technologies - or the expectations that students had regarding these limitations - might have inhibited the potential modality differences. In Experiment 2, for example, there was a paucity of student explanation behavior, perhaps creating a situation in which the gains observed in human tutoring did not have the opportunity to be demonstrated. Thus, exploring and ameliorating both perceived and actual system limitations is an important step in demonstrating whether speech interaction yields benefits for tutorial dialogue systems.

To this end, we have started to investigate whether the expectations that students bring with them to a computer interaction contributes to our observed lack of student explanation. In a recent study, we had students interact with a computer using the same KCD technology used in Experiment 2, and found that students who believed they were interacting with a human offered more explanation behavior than students who believed they were interacting with a computer (Rosé & Torrey, 2005). Issues of how student expectations affect behavior, and how patterns of behavior do or do not take advantage of affordances of the technology, are very complex and require much further investigation.

In addition, although we found that the errors introduced by speech recognition did not decrease student learning in computer tutoring, perhaps these errors nonetheless inhibited any potential increases in learning. We have recently completed the implementation of ITSPOKE Version 2, to improve both its input and output spoken language components. We have used the data from Experiment 2 to enhance our language models, to increase the accuracy of our speech recognizer. We have also replaced ITSPOKE's original machine-generated voice (using a text-to-speech system) with a human voice (using pre-recorded audio), and plan to evaluate if improving ITSPOKE's voice increases student learning.

We hope that the results of studies such as those reported here will impact the development of future dialogue tutoring systems incorporating speech, by highlighting the performance gains that can be expected, and the requirements for their achievement. New tutorial strategies optimized for speech can then be incorporated into future systems for experimentation.

## ACKNOWLEDGMENTS

This research is supported by ONR (N00014-00-1-0600, N00014-04-1-0108) and NSF (0325054).

## REFERENCES

Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002a). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems* (pp. 16–28), online.

*the- both*), while the acknowledgment *mm-hm* is an example grounding.

Aist, G., Dowding, J., Hockey, B. A., & Hieronymus, J. (2002b). A demonstration of a spoken dialogue interface to an intelligent procedure assistant for astronaut training and support. In *Demo Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 83–84).

Albrektson, J. R. (1995). Mentored online seminar: A model for graduate-level distance learning. *T. H. E. Journal*, 23(3), 102–105.

Aleven, V. & Rosé, C. P., (Eds.) (2003). *Proceedings of the AIED 2003 Workshop on Tutorial Dialogue Systems: With a View toward the Classroom*, online.

Aleven, V., Popescu, O., & Koedinger, K. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.) *Artificial Intelligence in Education* (pp. 246–255). Amsterdam:IOS Press.

Aleven, V., Popescu, O., & Koedinger, K. R. (2002). Pilot-testing a tutorial dialogue system that supports self-explanation. In S. A. Cerri, G. Gouardères, & F. Paraguaų (Eds.) *Intelligent Tutoring Systems* (pp. 344–354). Berlin:Springer.

Allen, J. F., Miller, B. W., Ringger, E. K., & Sikorski, T. (1996). A robust system for natural spoken dialogue. In A. Joshi & M. Palmer (Eds.) *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 62–70). San Francisco:Morgan Kaufmann Publishers.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.

Ashley, K. D., Desai, R., & Levine, J. M. (2002). Teaching case-based argumentation concepts using didactic arguments vs. didactic explanations. In S. A. Cerri, G. Gouardères, & F. Paraguaų (Eds.) *Intelligent Tutoring Systems* (pp. 585–595). Berlin:Springer.

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, 30, 117–139.

Baylor, A. L., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. In P. Kommers & G. Richards (Eds.) *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 452–458). Chesapeake, VA: AACE.

Bhatt, K., Evens, M., & Argamon, S. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Black, A. & Taylor, P. (1997). Festival speech synthesis system: system documentation (1.1.1). Human Communication Research Centre Technical Report 83, University of Edinburgh.

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4–16.

Chester, A. & Gwynne, G. (1998). Online teaching: Encouraging collaboration through anonymity. *Journal of Computer Mediated Communication*, 4(2), online.

Chi, M., Leeuw, N. D., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.

Chi, M. T. H. (1996). Learning processes in tutoring. *Applied Cognitive Psychology*, 10, S33–S49.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 67–74).

Craig, S. D. & Graesser, A. (2003). Why am I confused: An exploratory look into the role of affect in learning. In A. Mendez-Vilas & J. Gonzalez (Eds.) *Advances in Technology-based Education: Towards a Knowledge-based Society Volume 3* (pp. 1903–1906).

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3), 241–250.

Evens, M., Brandle, S., Chang, R., Freedman, R., Glass, M., Lee, Y. H., Shim, L. S., Woo, C. W., Zhang, Y., Zhou, Y., Michael, J. A., & Rovick, A. A. (2001). Circsim-tutor: An intelligent tutoring system using natural language dialogue. In *Proceedings of the Twelfth Midwest AI and Cognitive Science Conference (MAICS)* (pp. 16–23).

Forbes-Riley, K., Litman, D., Huettner, A., & Ward, A. (2005). Dialogue-learning correlations in spoken dialogue tutoring. In C. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) *Artificial Intelligence in Education* (pp. 225–232). Amsterdam:IOS Press.

Gergle, D., Millen, D., Kraut, R. E., & Fussell, S. (2004). Persistence matters: Making the most of chat in tightly-coupled work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 431–438). New York:ACM Press.

Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group (1999). Autotutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1(1), 35–51.

Graesser, A. C., Moreno, K. N., Marineau, J. C., Adcock, A. B., Olney, A. M., & Person, N. K. (2003). Autotutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Artificial Intelligence in Education* (pp. 47–54). Amsterdam:IOS Press.

Hausmann, R. & Chi, M. (2002). Can a computer interface support self-explaining? *The International Journal of Cognitive Technology*, 7(1), 4–14.

Heffernan, N. T. & Koedinger, K. R. (2002). An intelligent tutoring system incorporating a model of an experienced tutor. In S. A. Cerri, G. Gouardères, & F. Paraguaȩ (Eds.) *Intelligent Tutoring Systems* (pp. 596–608). Berlin:Springer.

Herring, S. (1999). Interactional coherence in CMC. *Journal of Computer Mediated Communication*, 4(4), online.

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., & Rosenfeld, R. (1993). The SphinxII speech recognition system: An overview. *Computer, Speech and Language*, 7(2), 137–148.

Jackson, G., Person, N., & Graesser, A. (2004). Adaptive tutorial dialogue in AutoTutor. In *Proceedings of the ITS 2004 Workshop on Dialog-based Intelligent Tutoring Systems*, online.

Jensen, C., Farnham, S., Drucker, S., & Kollock, P. (2000). The effect of communication modality on cooperation in online environments. *CHI Letters*, 2(1), 1–6.

Johnson, W. L., Rizzo, P., Bosma, W., Kole, S., Ghijsen, M., & van Welbergen, H. (2004). Generating socially appropriate tutorial dialog. In E. André et al. (Eds.) *Proceedings of the ISCA Workshop on Affective Dialogue Systems (ADS)* (pp. 254–264). Berlin:Springer-Verlag.

Jordan, P., Makatchev, M., & VanLehn, K. (2003). Abductive theorem proving for analyzing student explanations. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Artificial Intelligence in Education*. Amsterdam:IOS Press.

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence and Education*, 13, 641–650.

Litman, D. & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* (pp. 233-236).

Litman, D. J. & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 352-359).

Litman, D. & Forbes-Riley, K. (2005). Speech recognition performance and learning in spoken dialogue tutoring. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech/Eurospeech)* (pp. 161–164).

Mayer, R. E. (2002). *Multimedia Learning*. Cambridge University Press.

Michael, J., Rovick, A., Glass, M. S., Zhou, Y., & Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments*, 11(3), 233–262.

Moore, J. D., Porayska-Pomsta, K., Varges, S., & Zinn, C. (2004). Generating tutorial feedback with affect. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*.

Moreno, R., Mayer, R. E., Spires, H. A., & 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. *Cognition and Instruction*, 19(2), 177–213.

Mostow, J. & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.) *Smart Machines in Education: The coming revolution in educational technology* (pp. 169–234). MIT/AAAI Press.

Newlands, D. & McKean, A. (1996). The potential of live teacher supported distance learning: A case-study of the use of audio conferencing at the university of aberdeen. *Studies in Higher Education*, 21(3), 285–297.

Pon-Barry, H., Clark, B., Bratt, E., Schultz, K., & Peters, S. (2004). Evaluating the effectiveness of SCoT: A spoken conversational tutor. In *Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions*, online.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21(1), 1–29.

Rickel, J. & Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.) *Embodied Conversational Agents* (pp. 95-122). Cambridge:MIT Press.

Rosé, C. P. (2000). A framework for robust sentence level interpretation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1129–1135).

Rosé, C. P. & Aleven, V. (2002). C. P. Rosé & V. Aleven (Eds.) *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, online.

Rosé, C. P. & Freedman, R. (2000). Building dialogue systems for tutorial applications. Technical Report FS-00-01 (Working Notes of the Fall Symposium), AAAI.

Rosé, C. P. & Torrey, C. (2005). Interactivity and expectation: Eliciting learning oriented behavior with tutorial dialogue systems. In *Proceedings of INTERACT 2005* (pp. 323–336).

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., and Weinstein, A. (2001). Interactive conceptual tutoring in atlas-andes. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.) *Artificial Intelligence in Education* (pp. 256–266). Amsterdam:IOS Press.

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). The role of why questions in effective human tutoring. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Artificial Intelligence in Education*. Amsterdam:IOS Press.

Rudnicky, A. I. (1994). Mode preference in a simple data-retrieval task. In *Proceedings of the ARPA Human Language Technology Workshop '93* (pp. 364–369).

Schultz, K., Bratt, E. O., Clark, B., Peters, S., Pon-Barry, H., and Treeratpituk, P. (2003). A scalable, reusable spoken conversational tutor: Scot. In *Proceedings of the AIED 2003 Workshop on Tutorial Dialogue Systems: With a View toward the Classroom*, (pp. 367–377).

Siler, S. (2004). *Does tutors' use of their knowledge of their students enhance one-to-one-tutoring?* Ph.D. thesis, University of Pittsburgh.

VanLehn, K., Jordan, P., Rosé, C., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., Srivastava, R., & Wilson, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouardères, & F. Paraguaų (Eds.) *Intelligent Tutoring Systems* (pp. 158–167). Berlin:Springer.

VanLehn, K., Lynch, C., Schultz, K., Shapiro, J. A., Shelby, R. H., Taylor, L., Treacy, D. J., Weinstein, A., & Wintersgill, M. C. (2005). The Andes physics tutoring system: Five Years of Evaluations. In C. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) *Artificial Intelligence in Education* (pp. 678–685). Amsterdam:IOS Press.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (submitted). Natural language tutoring: A comparison of human tutors, computer tutors and text.

Wood, D. J., Wood, H., & Middleton, D. (1978). An experimental evaluation of four face-to-face teaching strategies. *International journal of Behavioral Development*, 1, 131–147.

Zinn, C., Moore, J. D., & Core, M. G. (2002). A 3-tier planning architecture for managing tutorial dialogue. In S. A. Cerri, G. Gouardères, & F. Paraguaų (Eds.) *Intelligent Tutoring Systems* (pp. 574–584). Berlin:Springer.