A Comparison of Tutor and Student Behavior in Speech Versus Text Based Tutoring

Carolyn P. Rosé, Diane Litman, Dumisizwe Bhembe, Kate Forbes, Scott Silliman, Ramesh Srivastava, Kurt VanLehn
Learning Research and Development Center, University of Pittsburgh,
3939 O'Hara St., Pittsburgh, PA 15260
rosecp,bhembe,forbesk,scotts,rcsriva@pitt.edu
litman,vanlehn@cs.pitt.edu

Abstract

This paper describes preliminary work in exploring the relative effectiveness of speech versus text based tutoring. Most current tutorial dialogue systems are text based (Evens et al., 2001; Rose and Aleven, 2002; Zinn et al., 2002; Aleven et al., 2001; VanLehn et al., 2002). However, prior studies have shown considerable benefits of tutoring through spoken interactions (Lemke, 1990; Chi et al., 1994; Hausmann and Chi, 2002). Thus, we are currently developing a *speech based* dialogue system that uses a text based system for tutoring conceptual physics (VanLehn et al., 2002) as its "back-end". In order to explore the relative effectiveness between these two input modalities in our task domain, we have started by collecting parallel human-human tutoring corpora both for text based and speech based tutoring. In both cases, students interact with the tutor through a web interface. We present here a comparison between the two on a number of features of dialogue that have been demonstrated to correlate reliably with learning gains with students interacting with the tutor using the text based interface (Rosé et al., submitted).

1 Introduction

This paper describes preliminary work in exploring the relative effectiveness of speech versus text based tutorial dialogue systems. 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. For example, it has been demonstrated that generating words rather than simply reading them promotes subsequent recall of those words (Slamecka and Graf, 1978). (Chi et al., 1994) notes that there is a general momentum in the science education literature toward the importance of talking, reflecting and explaining as ways to learn (Lemke, 1990). Moreover, encouraging student self-explanation, which includes both generating inferences from material they have read and relating new material to old material, has been shown to correlate with learning (Chi et al., 1981; Chi et al., 1994; Renkl, 1997; Pressley et al., 1992). In a further study, prompting students with zero content prompts to encourage them to self-explain was also associated with student learning (Chi et al., 2001). A second important advantage to dialogue is that it affords the tutor the opportunity to tailor instruction to the needs of the student. 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 (Chi, 1996). (Rosé et al., submitted) explore the benefits of tutor adaptation by comparing learning gains for naive learners and review learners in a human tutoring condition and a non-adaptive reading condition.

In recent years tutorial dialogue systems have become more and more prevalent, most of which are text based (Evens et al., 2001; Rose and Aleven, 2002; Zinn et al., 2002; Aleven et al., 2001; VanLehn et al., 2002). Many of these systems have yielded successful evaluations with students (Rosé et al., 2001; Heffernan and Koedinger, 2002; Ashley et al., 2002; Graesser et al., 2001a). However, while the majority of current tutorial dialogue systems are text based, there is reason to believe that speech based tutorial dialogue systems could be more effective.

Prior studies have shown considerable benefits of human-human tutoring through spoken interactions (Lemke, 1990; Chi et al., 1994). (Hausmann and Chi, 2002) has shown that spontaneous self-explanation occurs much more frequently in spoken tutoring then in text based tutoring, suggesting that typing requires additional cognitive capacity and thus reduces the cognitive resources available for spontaneous self-explanation. Other research projects (Mostow and Aist, 2001; Fry et al., 2001) have shown that basic spoken natural language capabilities can be implemented quite effectively in computer tutoring systems. Moreover, speech contains prosodic and acoustic information which has been shown to improve the accuracy of predicting emotional states (Ang et al., 2002; Batliner et al., 2000) and user responses to system errors (Litman et al., 2001) that are useful for triggering system adaptation. We are thus currently developing a *speech based* dialogue system that uses a text based system (VanLehn et al., 2002) as its "back-end". These systems and their goals will be discussed in Section 2.

We expect that the different modalities used by these systems (e.g. text based vs speech based) will display interesting differences with respect to the characteristics of dialogue interaction that may determine their relative merits with respect to increasing student performance. Although human-computer data from the speech based system is not yet available for comparison, we have collected parallel human-human corpora both for text based and speech based tutoring, as discussed in Sections 3-4, and these corpora already display similarities and differences with respect to features of their dialogue interactions, as discussed in Section 5, that are wholly modality based and that will likely be displayed to an even greater extent in the comparable human-computer data.

2 Why2-Atlas and ITSPOKE Dialogue Systems

Why2-Atlas is a *text based* intelligent tutoring dialogue system (Rosé et al., 2002a; VanLehn et al., 2002). The goal of Why2-Atlas is to provide a platform for testing whether deep approaches to natural language processing elicit more learning than shallower approaches, for the task domain of qualitative physics explanation generation. Using Why2-Atlas, the activity in which students engage is answering deep reasoning questions involving topics in conceptual physics. One such question that we used is, "A lightweight car and a massive truck have a head-on collision. On which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Explain." This is an appropriate task domain for pursuing questions about the benefits of tutorial dialogue for learning because questions like this one are known to elicit robust, persistent misconceptions from students, such as "heavier objects exert more force." (Hake, 1998; Halloun and Hestenes, 1985). We designed a set of 10 essay questions to use as training problems. Two physics professors and a computer science professor worked together to select a set of expectations (i.e., correct propositions that the tutors expected students to include in their essays) and potential misconceptions associated with each question. Additionally, they agreed on an ideal essay answer for each problem. In Why2-Atlas, a student first types an essay answering a qualitative physics problem. A computer tutor then engages the student in a natural language dialogue to provide feedback, correct misconceptions, and to elicit more complete explanations. The first version of Why2-Atlas was deployed and evaluated with undergraduate students in the spring of 2002; the system is continuing to be actively developed (Graesser et al., 2002).

We are currently developing a *speech-enabled* version of Why2-ATLAS, called ITSPOKE (Intelligent Tutoring SPOKEn dialogue system), that uses the Why2-Atlas system as its "back-end". To date we have interfaced the Sphinx2 speech recognizer (Huang et al., 1993) with stochastic language models trained from example user utterances, and the Festival speech synthesizer (Black and Taylor, 1997) for text-to-speech, to the Why2-Atlas back-end. The rest of the needed natural language processing components, e.g. the sentence-level syntactic and semantic analysis modules (Rosé, 2000), discourse and domain level processors (Makatchev et al., 2002), and a finite-state dialogue manager (Rosé et al., 2001), are provided by a toolkit that is part of the Why2-Atlas back-end. The student speech is digitized from microphone input, while the tutor's synthesized speech is played to the student using a speaker and/or headphone. We are now in the process of adapting the knowledge sources needed by the spoken language components to our application domain. For example, we have developed a set of dialogue dependent language models using the experimental human-computer typed corpus (4551 student utterances) obtained during the Why2-Atlas 2002 evaluation. Our language models will soon be enhanced using student utterances from our parallel human-human spoken language corpus.

One goal of the ITSPOKE system is simply replacing text based dialogue interaction with spoken dialogue interaction and leaving the rest of the Why2-Atlas back-end unchanged, in order to test the hypothesis that student self-explanation (which leads to greater learning (Hausmann and Chi, 2002)) might be easier to achieve in spoken dialogues. This hypothesis is discussed further in Section 5. Although not the focus of this paper, another goal of the ITSPOKE system is to take full advantage of the speech modality. For example, speech contains rich acoustic and prosodic information about the speaker's current emotional state that isn't present in typed dialogue. Connections between learning and emotion have been well documented (Coles, 1999), so it seems likely that the success of computer-based tutoring systems could be greatly increased if they were capable of predicting and adapting to student emotional states, e.g. reinforcing positive states, while rectifying negative states (Evens, 2002). Preliminary machine learning experiments involving emotion annotation and automatic feature extraction from our corpus suggest that ITSPOKE can indeed be enhanced to automatically predict and adapt to student emotional states (Litman et al., 2003).

3 Typed Human-Human Tutoring Corpus

The Why2-Atlas Human-Human Typed Tutoring Corpus is a collection of typed tutoring dialogues between (human) tutor and student collected via typed interface, which the tutor plays the same role that Why2-Atlas is designed to perform. The experimental procedure is as follows: 1) students are given a pretest measuring their knowledge of physics, 2) students are asked to read through a small document of background material, 3) students work through a set of up to 10 Why2-Atlas training problems with the human tutor, and 4) students are given a post-test that is similar to the pretest. The entire experiment takes no more than 15 hours per student, and is usually performed in 1-3 sessions of no more than 4 hours each. Data collection began in the Fall 2002 semester and is continuing in the Spring 2003 semester. The subjects are all University of Pittsburgh students who have never taken any college physics courses. One tutor currently participates in the study.

As in the Why2-Atlas system, when the tutoring session then begins, the student first types an essay answering a qualitative physics problem. Once the student submits his/her essay, the tutor then engages the student in a typed natural language dialogue to provide feedback and correct misconceptions, and to elicit more complete explanations. This instruction is in the form of a dialogue between the student and the tutor through a text based chat interface with student and tutor in separate rooms. At key points in the dialogue, the tutor asks the student to revise the essay. This cycle of instruction and revision continues until the tutor is satisfied with the student's essay. A sample tutoring dialogue from the Why2-Atlas typed human-human tutoring corpus is displayed in Figure 1.

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. So his tutoring style was much more naturalistic than in previous studies such as the BEE study (Rosé et al., 2001) in which two specific tutoring styles, namely Socratic and Didactic, were contrasted. The results of that study revealed a trend for students in the Socratic condition to learn more than those in the Didactic condition. A further analysis of the corpus collected during the BEE study (Core et al., 2002) verified that the Socratic dialogues from the BEE study were more interactive than the Didactic ones. The biggest reliable difference between the two sets of tutoring dialogues was the percentage of words spoken by the student, i.e, number of student words divided by total number of words. The Didactic dialogues contained on average 26% student words, whereas the Socratic dialogues contained 33% student words. On average with respect to percentage of student words, the dialogues in our text based human tutoring corpus were more like the Didactic dialogues from the BEE study, with average percentage of student text being 27%. Nevertheless, because the tutor was not constrained to follow a prescribed tutoring style, the level of interactivity varied widely throughout the transcripts, at times being highly Socratic, and at other times being highly Didactic.

Pre and post tests were used to measure learning gains to be used for evaluating the effectiveness of various features of tutorial dialogue found in our corpora. Thus, we developed two tests: versions A and B, which were isomorphic to one another. That is, the problems on test A and B differed only in the identities of the objects (e.g., cars vs. trucks) and other surface features that should not affect the reasoning required to solve them. Each version of the test (A and B) consisted of 40 multiple choice questions. Each multiple choice question was written to address a single expectation covered in the training problems. Some students were not able to complete all 10 problems before they reached the end of their participation time. Thus, they took the post-test after only working through a subset of the training problems.

4 Spoken Human-Human Tutoring Corpus

The ITSPOKE Human-Human Spoken Tutoring Corpus is a parallel collection of spoken tutoring dialogues collected via a web interface supplemented with a high quality audio link, where a human tutor performs the same task that our ITSPOKE system is being designed to perform. The experimental procedure used to collect the corpus is exactly the same as the procedure used to gather the Why2-Atlas Human-Human Corpus: the same tutor is used, the same subject pool[1] is used, the same pre-test and post-test are used, and the same set of physics prob-

[1] We assigned a greater percentage of students to the text based condition as part of a separate experiment. Thus, the text based corpus is larger than the speech based corpus.

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 minutes into a typed dialogue . . .

Tutor1: You have correctly said the balls are in free fall. What do you conclude from this fact?

Student1: There is no air resistance and the balls' accelerations are constant in regards to one another

Tutor2: 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?

Student2: F (net forces) / mass = f/m because F = m*a therefore ratio should equal one another

Tutor3: 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?

Student3: yes, assuming starting at rest? sorry, kind of at a loss at the moment

Tutor4: The initial velocity will matter, and here they both have the same initial velocity, zero. What else, will the height through wich they fall matter?

Student4: not so much as long as they are both dropped from the same height

Tutor5: 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?

Student5: since initial velocities are equal and the height of falls are equal, then the time taken is equal

Tutor6: 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?

Student6: no...

Tutor7: Right...

Figure 1: **Excerpt from Human-Human Typed Dialogue Corpus**

lems are used. Only the modality differs. In particular, once the tutoring session begins and the student submits his/her typed essay, the tutor and student then discuss the student's essay using spoken English. In contrast to the text condition, where strict turn-taking is enforced, in the spoken condition, interruptions and overlapping speech are common. An example excerpt from the corpus is shown in Figure 2. Note that turns ending in "-" indicate speech overlapping with the following turn. Eventually, the student will edit his/her typed explanation. As in the text condition, the tutor then either expresses satisfaction and ends the tutoring for the current problem, or continues with another round of spoken dialogue interaction and typed essay revision. As in the text condition, students are presented with the ideal essay answer for a problem upon completing that problem.

5 Differences between Typed and Spoken Human-Tutoring

(Rosé et al., submitted) presents an analysis to uncover which aspects of the tutorial dialogue were responsible for its effectiveness in the *text based* condition. Longer student answers to tutor questions reveal more of a student's reasoning. Very short answers, i.e., 10 words or less, are normally composed of a single clause at most. Longer, multi-clausal answers have the potential to communicate many more inter-connections between ideas. Thus, if a tutor is attending to and responding directly to the student's revealed knowledge state, it would be expected that the effectiveness of the tutor's instruction would increase as average student turn length increases. To test this prediction, we computed a linear regression of the sequence of student turn lengths over time for each student in the text based condition in order to obtain an intercept and a slope, since student turn lengths have been observed to decline on average over the course of their

PROBLEM: If a car is able to accelerate at 2 m/s2, what acceleration can it attain if it is towing another car of equal mass?

ESSAY: If the car is towing another car of equal mass, the maximum acceleration would be the same because the car would be towed behind and the friction caused would only be by the front of the first car.

. . . excerpt from 6.5 minutes into spoken dialogue . . .

Tutor1: So twice the mass multiplied by the acceleration should be equal to the force which you have already determined as the mass of the first car times the acceleration. So essentially you are dividing it by two and that gives you the uh acceleration by twice the because mass has become twice. Now this law that force is equal to mass times acceleration, what's this law called? This is uh since this it is a very important basic uh fact uh it is it is a law of physics. Um you have you have read it in the background material. Can you recall it?

Student1: Um no it was one of Newton's laws-

Tutor2: Right, right-

Student2: but I don't remember which one. (laugh)

Tutor3: That-

Student3: he I-

Tutor4: is Newton's second law of motion.

Student4: Ok, because I remember one, two, and three, but I didn't know if there was a different name

Tutor5: Yeah that's right you know Newton was a genius and-

Student5: (laugh)-

Tutor6: uh he looked at a large number of experiments and experimental data that was available and from that he could come to this general law and it is known as Newton's second law of motion. Um many many other scientists before him had seen all this data which was collected by scientists but had not concluded this. Now it looks very simple but to come to a conclusion from a mass of data was something which required the genius of Newton.

Student6: mm hm

Tutor7: So now you will give Newton full credit isn't it? (laugh)

Student7: (laugh)

Figure 2: **Excerpt from Human-Human Spoken Dialogue Corpus.**

interaction with the turn. We then computed a multiple regression with pre-test score, intercept, and gradient as independent variables and post test score as the dependent variable. We found a reliable correlation between intercept and learning, with pre-test scores and gradients regressed out (R=.836; p<.05). This result is consistent with (Core et al., 2002) where percentage of student talk is strongly correlated with learning. Consistent with this, we found a strong and reliable correlation between ratio of student words to tutor words and learning[2]. We computed a correlation between ratio of student words to tutor words and post-test score after pre-test scores were regressed out (R=.866, p<.05).

One of our current research objectives is to compare the relative effectiveness of speech based and text based tutoring. Thus, when we have enough speech data, we would like to compare learning gains between the speech and text based conditions to test whether or not speech based tutoring is more effective than text based tutoring. We also plan to test whether the same features that correlate with learning in the text based condition also correlate with learning in the speech based condition. Since both average student turn length and overall ratio of student words to tutor words correlated strongly with learning gains in the text based condition, in this paper we compare these two measures between the text based tutoring condition and the speech based tutoring condition, but not yet in connection with learning gains in the speech-based corpus.

Since strict turn taking was not enforced in the speech condition, turn boundaries were manually annotated (based on consensus labellings from two coders) when ei-

[2] Note that ratio of student words to tutor words is number of student words divided by number of tutor words, whereas percentage of student words is number of student words divided by total number of words

ther (1) the speaker stopped speaking and the other party in the dialogue began to speak, (2) when the speaker asked a question and stopped speaking to wait for an answer, or (3) when the other party in the dialogue interrupted the speaker and the speaker paused to allow the other party to speak.

Currently, 13 students have started the typed human-human tutoring experiment, 7 of whom have finished. We have so far collected 78 typed dialogues from the text based condition, 69 of which were used in our analysis. 9 students have started the spoken human-human tutoring experiment, 6 of whom have finished. Thus, we have collected 63 speech based dialogues (1290 minutes of speech from 4 female and 4 male subjects), and have transcribed 25 of them. We hope to have an analysis covering all of our data in both conditions by the time of the workshop.

As shown in Table 1, analysis of the data that has been collected and transcribed to date is already showing interesting differences between the ITSPOKE (spoken) and WHY2-ATLAS (text) corpora of human-human dialogues. The #trns columns show mean and standard deviation for the total number of turns taken by the students or tutor in each problem dialogue, while the next pair of columns show the mean and standard deviation for the total number of words spoken or typed by the students or tutor (#wds) in each problem dialogue. The last pair of columns show mean and standard deviation for the average number of student or tutor words per turn in each problem dialogue.

Due to the fact that data is still being collected for both corpora (and the fact that the speech corpus also requires manual transcription), the sizes of the two data sets represented in the table differ somewhat. However, even at this early stage in the development of both corpora, these figures already show that the style of the interactions are very different in each modality. In particular, in spoken tutoring, both student and tutor take more turns on average than in text based 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 text based tutoring. Another interesting difference is that although in the speech condition both student and tutor take more turns, students finish the speech condition in less time. In particular, on average, students in the text based tutoring condition require 370.58 minutes to finish the training problems, with a standard deviation of 134.29 minutes, students in the speech condition require only 159.9 minutes on average, with a standard deviation of 58.6 minutes. We measured the statistical reliability of the difference between the two measures that correlated reliably with learning in the text-based condition. A 2-tailed unpaired t-test indicates that this difference is significant (t(30)=8.99, p<.01). There are also similarities across the two conditions. In particular, a 2-tailed unpaired t-test shows that the relative proportion of student and tutor word or turns do not differ significantly on average across the two modalities (t(13)=1.225, p=.242). As an illustration, Table 2 shows mean and standard deviation for the ratios of the total number of student and tutor words (#Swds/#Twds) and turns (#Strns/#Ttrns) in each problem dialogue[3].

Average student turn length is significantly lower in the speech based condition (t(13) = 4.5, p< .001). This might predict that speech based tutoring may be less effective than text based tutoring. However, since ratio of student words to tutor words does not differ significantly, this would predict that learning will also not differ significantly between conditions. Total number of words uttered in the speech condition is larger than in the text based condition as are total number of turns. This difference between the two conditions will likely be even more pronounced in the human-computer comparison, due to noisy student input that results from use of automatic speech recognition. For example, clarifications and corrections made necessary by this will likely lead to an increase in dialogue length. More careful analysis is required to determine whether this means that more self-explanation took place overall in the speech based condition. If so, this would predict that the speech based condition would be more effective for learning than the text based condition. Thus, much interesting exploration is left to be done after we have collected enough speech data to compute a reliable comparison between the two conditions.

6 Current Directions

Currently we are continuing to collect data both in the speech and text based human tutoring conditions. Since human tutors differ with each other with respect to both their tutoring styles and their conversational styles, we plan to collect data using several different human tutors in order to test the robustness of our comparisons between speech and text based human tutoring. Another possible direction for further inquiry would be to contrast naturalistic speech (where strict turn taking is not enforced, as in this data collection effort), with a speech condition in which strict turn taking is enforced, in order to separate the effects of speech on learning from the effects of alternative turn taking policies.

As discussed in Section 2, we are currently developing both text based and speech based human-computer tutorial systems. Our ultimate goal is to test the relative effectiveness of speech versus text based computer tutors. We expect differences both between text and speech con-

[3] Note that strict-turn taking is enforced in the text condition, but not in the speech condition.

Table 1: Student and Tutor Characteristics in Human-Human Speech and Text Conditions

		#trns		#wds		Avg#wds/trn	
Condition	Participant	Mean	STD	Mean	STD	Mean	STD
Speech	Student	47.49	25.95	264.18	125.47	5.72	1.35
Text	Student	9.71	6.79	146.72	57.96	13.39	5.55
Speech	Tutor	46.94	20.90	1199.14	605.87	26.78	14.20
Text	Tutor	11.03	7.04	391.85	136.89	39.04	6.23

Table 2: Student-Tutor Word and Turn Ratios in Speech and Text Conditions

Speech Condition				Text Condition			
#Swds/#Twds		#Strns/#Ttrns		#Swds/#Twds		#Strns/#Ttrns	
Mean	STD	Mean	STD	Mean	STD	Mean	STD
0.29	0.15	0.99	0.15	0.37	0.08	0.81	0.15

ditions in the human-computer data and between human-human and human-computer data. One of our first tasks will thus be to use the baseline version of ITSPOKE described in Section 2 to generate a corpus of *human-computer* spoken dialogues, using a process comparable to the human-human corpus collection described here. This will allow us to 1) compare the ITSPOKE human-human and human-computer corpora 2) compare the ITSPOKE human-computer spoken corpus with a comparable Why2-Atlas text corpus, e.g. by expanding on the just described pilot study of the two human-human corpora, and 3) use the ITSPOKE human-computer corpus to guide the development of a new version of ITSPOKE that will attempt to increase its performance, by taking advantage of information that is only available in speech, and modifying its behavior in other ways to respect the interaction differences in item 2.

7 Acknowledgments

This research was supported by the Office of Naval Research, Cognitive Science Division under grant number N00014-0-1-0600 and by NSF grant number 9720359 to CIRCLE.

References

V. Aleven, O. Popescu, and K. Koedinger. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In J. D. Moore, C. L. Redfield, and W. L. Johnson, editors, *Proceedings of Artificial Intelligence in Education*, pages 246–255.

J. Ang, R. Dhillon, A. Krupski, E.Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. ICSLP*.

K. D. Ashley, R. Desai, and J. M. Levine. 2002. Teaching case-based argumentation concepts using didactic arguments vs. didactic explanations. In *Proceedings of the Intelligent Tutoring Systems Conference*, pages 585–595.

A. Batliner, R. Huber, H. R. Niemann, E. Nöth, J. Spilker, and K. Fischer. 2000. The recognition of emotion. In *Proc. of the ISCA Workshop on Speech and Emotion*.

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

M. Chi, N. de Leeuw, M. Chiu, and C. LaVancher. 1981. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3).

Michelene Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18:439–477.

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

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

G. Coles. 1999. Literacy, emotions, and the brain. Reading Online, March 1999.

M. Core, J. D. Moore, and C. Zinn. 2002. Initiative in tutorial dialogue. In *Proceedings of the ITS Workshop on Empirical Methods for Tutorial Dialogue Systems*, pages 46–55.

M. Evens, S. Brandle, R. Chang, R. Freedman, M. Glass, Y. H. Lee, L. S. Shim, C. W. Woo, Y. Zhang, Y. Zhou,

J. A. Michaeland, and A. A. Rovick. 2001. Circsim-tutor: An intelligent tutoring system using natural language dialogue. In *Proceedings of the Twelfth Midwest AI and Cognitive Science Conference, MAICS 2001*, pages 16–23, Oxford, OH.

M. Evens. 2002. New questions for Circsim-Tutor. Presentation at the 2002 Symposium on Natural Language Tutoring, University of Pittsburgh.

J. Fry, M. Ginzton, S. Peters, B. Clark, and H. Pon-Barry. 2001. Automated tutoring dialogues for training in shipboard damage control. In *Proc. 2nd SigDial Workshop on Discourse and Dialogue*.

A. Graesser, N. Person, and D. Harter et al. 2001a. Teaching tactics and dialog in Autotutor. *International Journal of Artificial Intelligence in Education*.

A. Graesser, K. Vanlehn, TRG, and NLT Group. 2002. Why2 report: Evaluation of why/atlas, why/autotutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Technical report, LRDC Tech Report, University of Pittsburgh.

R. R. Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics students. *American Journal of Physics*, 66(64).

I. A. Halloun and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*, 53(11):1043–1055.

Robert Hausmann and Michelene Chi. 2002. Can a computer interface support self-explaining? *The International Journal of Cognitive Technology*, 7(1).

N. T. Heffernan and K. R. Koedinger. 2002. An intelligent tutoring system incorporating a model of an experienced tutor. In *Proceedings of the Intelligent Tutoring Systems Conference*, pages 596–608.

X. D. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, K. F. Lee, and R. Rosenfeld. 1993. The SphinxII speech recognition system: An overview. *Computer, Speech and Language*.

J. L. Lemke. 1990. *Talking Science: Language, Learning and Values*. Ablex, Norwood, NJ.

D. Litman, J. Hirschberg, and M. Swerts. 2001. Predicting user reactions to system error. In *Proc.of ACL*.

Diane Litman, Kate Forbes, and Scott Silliman. 2003. Towards emotion prediction in spoken tutoring dialogues. Submitted.

M. Makatchev, P. Jordan, and K. VanLehn. 2002. Discourse processing for explanatory essays in tutorial applications. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*.

J. Mostow and G. Aist. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus and P. Feltovich, editors, *Smart Machines in Education*.

M. Pressley, E. Wood, V. E. Woloshyn, V. Martin, A. King, and D. Menke. 1992. Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27:91–109.

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

C. P. Rose and V. Aleven. 2002. Proc. of the ITS 2002 workshop on empirical methods for tutorial dialogue systems. Technical report, San Sebastian, Spain, June.

C. P. Rosé, J. D. Moore, K. VanLehn, and D. Allbritton. 2001. A comparative evaluation of socratic versus didactic tutoring. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 869–874.

C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. 2001. Interactive conceptual tutoring in atlas-andes. In *Proceedings of Artificial Intelligence in Education*, pages 256–266.

C. P. Rosé, D. Bhembe, A. Roque, S. Siler, R. Srivastava, and K. Vanlehn. 2002a. A hybrid language understanding approach for robust selection of tutoring goals. In *Proceedings of the Intelligent Tutoring Systems Conference*, pages 552–561.

C. P. Rosé, K. VanLehn, and The Natural Language Tutoring Group. Submitted. Is human tutoring always more effective than reading. In *Annual Meeting of the Cognitive Science Society*.

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

N. J. Slamecka and P. Graf. 1978. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, (4):592–604.

K. VanLehn, P. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. of ITS*.

Claus Zinn, Johanna D. Moore, and Mark G. Core. 2002. A 3-tier planning architecture for managing tutorial dialogue. In *Proceedings Intelligent Tutoring Systems, Sixth International Conference (ITS 2002)*, Biarritz, France, June.