Spoken Versus Typed Human and Computer Dialogue Tutoring

Diane J. Litman[1], Carolyn P. Rosé[2], Kate Forbes-Riley[1], Kurt VanLehn[1], Dumisizwe Bhembe[1], and Scott Silliman[1]

[1] Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260
{litman,vanlehn}@cs.pitt.edu

[2] Language Technologies Institute/Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15260
{rosecp,forbesk,bhembe,scotts}@pitt.edu

Abstract. While human tutors typically interact with students using spoken dialogue, most computer dialogue tutors are text-based. We have conducted 2 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 large differences in the learning gains, time and superficial dialogue characteristics of human tutoring, but for computer tutoring it made less difference.

1 Introduction

It is widely believed that the best human tutors are more effective than the best computer tutors, in part because [1] found that human tutors could produce a larger difference in the learning gains than current computer tutors (e.g., [2,3,4]). 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 tutoring to spoken natural language dialogue, would decrease the advantage of human tutoring over computer tutoring.

Three main benefits of spoken tutorial dialogue with respect to increasing learning have been hypothesized. One is that spoken dialogue may elicit more student engagement and knowledge construction. [5] found that students who were prompted for self-explanations produced more when the self-explanations were spoken rather than typed. Self-explanation is just one form of student cognitive activity that is known to cause learning gains [6,7,8]. 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

J.C. Lester et al. (Eds.): ITS 2004, LNCS 3220, pp. 368–379, 2004.

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. In other work we have shown that the prosodic and acoustic information of speech can improve the detection of speaker states such as confusion [9], which may be useful for adapting tutoring to the student.

A third hypothesis is 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 (as in dialogue) to a learner. While [10] found that the use of a dialogue agent improved learning, there was no evidence that output media impacted learning. In [11], an interactive pedagogical agent using speech rather than text output improved student learning, while the visual presence or absence of the agent did not impact performance.

It is thus important to test whether a move to spoken dialogues is likely to cause higher learning gains, and if so, to understand why it accelerates learning. It is particularly important given that natural language tutoring systems are becoming more common. Although a few use spoken dialogues [12], most still use typed dialogues (e.g. [13,14,15]), although as shown by our work it is technically feasible to convert a tutor from typed dialogue tutor to spoken dialogue. While the details of this conversion are not covered in this paper, it took about 9 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. 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. 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. We thus conducted two experiments. Both used qualitative physics as the task domain, similar pretests and posttests, and similar training sequences. However, one experiment used an experienced human tutor who communicated with students either via speech or typing. The other used the Why2-Atlas tutoring system [16] with either its original typed dialogue or a new spoken dialogue user interface. The new system is called ITSPOKE [9].

2 The Common Aspects of the Experiments

In both experiments, 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.

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 (eg., 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 isomorphic to the pretest; both consisted of 40 multiple choice questions. The entire experiment took no more than 9 hours per student, and was usually performed in 1-3 sessions. Subjects were University students responding to ads, and were compensated with money or course credit.

The interface used for all experiments was basically the same. The student first typed an essay answering a qualitative physics problem. The tutor then engaged the student in a natural language dialogue to provide feedback, correct misconceptions, and to elicit more complete explanations. 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 tutoring dialogue system [16], developed in part to test whether deep approaches to natural language processing (e.g., sentence-level syntactic and semantic analysis, discourse and domain level processing, and finite-state dialogue management) elicit more learning than shallower approaches. ITSPOKE (Intelligent Tutoring SPOKEn dialogue system) [9] is a *speech-enabled* version of Why2-ATLAS. Student speech is digitized from microphone input and sent to the Sphinx2 recognizer. 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.

3 Human-Human Tutoring: Experiment 1

3.1 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 was used, with student and tutor in separate rooms; students knew that the tutor was human. 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 the typed condition

Excerpt from 2 minutes into a *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

Excerpt from 8.3 minutes into a *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 -

Fig. 1. Excerpts from Human-Human Dialogues

strict turn-taking was enforced, while in the spoken condition interruptions and overlapping speech were permitted. This was because we plan to add "bargein" to ITSPOKE, which will enable students to interrupt ITSPOKE. Sample dialogue excerpts from both conditions are displayed in Figure 1.

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.

3.2 Results

Table 1 presents the means and standard deviations for two types of analyses, learning and training time, across conditions. The pretest scores were not reliably different across the two conditions, $F(33) = 1.574$, $p = 0.219$, $MSe = 0.009$. In

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)

an ANOVA with condition by test phase factorial design, there was a robust main effect for test phase, $F(67) = 90.589$, $p = 0.000$, $MSe = 0.012$, indicating that students in both conditions learned a significant amount during tutoring. However, the main effect for condition was not reliable, $F(33) = 1.823$, $p = 0.186$, $MSe = 0.014$, and there was no reliable interaction. In an ANCOVA, the adjusted posttest scores show a strong trend of being reliably different, $F(1,33)=4.044$, $p=0.053$, $MSe = 0.01173$. Our results thus suggest that the human speech tutored students learned more than the human text tutored students; the effect size is 0.74. 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 $F(33) = 35.821$, $p = 0.00$, $MSe = 15958.787$. 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. 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

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

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). 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 [17] examined 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 [18], 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 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 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).

4 Human-Computer Tutoring: Experiment 2

4.1 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.

Figure 2 contains excerpts from both Why2-Atlas and ITSPOKE dialogues. Note that for ITSPOKE, the output of the automatic speech recognizer (the

Excerpt from 3.0 minutes into a *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

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)

Fig. 2. Excerpts from Why2-Atlas and ITSPOKE Dialogues

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.

4.2 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, $F(42) = 0.037$, $p = 0.848$, $MSe = 0.036$. In an ANOVA with condition by test phase factorial design, there was a

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)

robust main effect for test phase, $F(85) = 29.57$, $p = 0.000$, $MSe = 0.032$, indicating that students learned during their tutoring. The main effect for condition was not reliable, F(42)=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(1,42)=0.004, p=0.950, MSe=0.01806. Thus, the Why-Atlas tutored students did not learn reliably more than the ITSPOKE tutored students. 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(42)=9.411, p=0.004, MSe=950.792. In sum, while adding speech to Why2-Atlas 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. A separate analysis showed no correlation between word error or semantic degradation (discussed in Section 4.1) with learning or training time.

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

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 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 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. 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.

5 Discussion and Current Directions

The main results of our study are that changing the modality from text to speech caused large differences in the learning gains, time and superficial dialogue characteristics of human tutoring, but for computer tutoring it made less difference. Experiment 1 on human tutoring suggests 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 (and in our experiment, turn length was also impacted by interruption policy), suggesting that other measures might be more relevant. However, we plan to investigate whether spoken phenomena such as disfluencies and grounding might also explain the lack of correlation.

The results of Experiment 2 on computer tutoring are less conclusive. 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. Furthermore, compared to human tutoring, we see less difference between spoken and typed computer dialogue interactions, at least for the dialogue aspects measured in our experiments. One hypothesis is that simply adding a spoken "front-end", with-

out also modifying the tutorial dialogue system "back-end", is not enough to change how students interact with a computer tutor. 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.

Finally, 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.

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

References

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

2. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: Lessons learned. The Journal of the Learning Sciences **4** (1995) 167–207
3. VanLehn, K., Lynch, C., Taylor, L., Weinstein, A., Shelby, R.H., Schulze, K., Treacy, D.J., Wintersgill, M.C.: Minimally invasive tutoring of complex physics problem solving. In: Proc. Intelligent Tutoring Systems (ITS), 6th International Conference. (2002) 367–376
4. Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R.: Autotutor: A simulation of a human tutor. Journal of Cognitive Systems Research **1** (1999)
5. Hausmann, R., Chi, M.: Can a computer interface support self-explaining? The International Journal of Cognitive Technology **7** (2002)
6. Chi, M., Leeuw, N.D., Chiu, M., Lavancher, C.: Eliciting self-explanations improves understanding. Cognitive Science **18** (1994) 439–477
7. Renkl, A.: Learning from worked-out examples: A study on individual differences. Cognitive Science **21** (1997) 1–29
8. Chi, M.T.H., Siller, S.A., Jeong, H., Yamauchi, T., Hausmann, R.G.: Learning from human tutoring. Cognitive Science (2001) 471–477
9. Litman, D.J., Forbes-Riley, K.: Predicting student emotions in computer-human tutoring dialogues. In: Proc. Association Computational Linguistics (ACL). (2004)
10. Graesser, A.C., Moreno, K.N., Marineau, J.C., Adcock, A.B., Olney, A.M., Person, N.K.: Autotutor improves deep learning of computer literacy: Is it the dialog or the talking head? In: Proc. AI in Education. (2003)
11. Moreno, R., Mayer, R.E., Spires, H.A., Lester, J.C.: 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** (2001) 177–213
12. Schultz, K., Bratt, E.O., Clark, B., Peters, S., Pon-Barry, H., Treeratpituk, P.: A scalable, reusable spoken conversational tutor: Scot. In: AIED Supplementary Proceedings. (2003) 367–377
13. Michael, J., Rovick, A., Glass, M.S., Zhou, Y., Evens, M.: Learning from a computer tutor with natural language capabilities. Interactive Learning Environments (2003) 233–262
14. Zinn, C., Moore, J.D., Core, M.G.: A 3-tier planning architecture for managing tutorial dialogue. In: Proceedings Intelligent Tutoring Systems, Sixth International Conference (ITS 2002), Biarritz, France (2002) 574–584
15. Aleven, V., Popescu, O., Koedinger, K.R.: Pilot-testing a tutorial dialoque system that supports self-explanation. In: Proc. Intelligent Tutoring Systems (ITS): 6th International Conference. (2002) 344–354
16. 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.: The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In: Proc. Intelligent Tutoring Systems (ITS), 6th International Conference. (2002)
17. Core, M.G., Moore, J.D., Zinn, C.: The role of initiative in tutorial dialogue. In: Proc. 11th Conf. of European Chapter of the Association for Computational Linguistics (EACL). (2003) 67–74
18. Rose, C.P., Bhembe, D., Siler, S., Srivastava, R., VanLehn, K.: The role of why questions in effective human tutoring. In: Proc. AI in Education. (2003)