The Influence of Learner Characteristics on Task-Oriented Tutorial Dialogue

Kristy Elizabeth BOYER, Mladen A. VOUK, James C. LESTER
Department of Computer Science, North Carolina State University
Raleigh, NC 27695, USA
{keboyer, vouk, lester}@ncsu.edu

Abstract. Tutorial dialogue has been the subject of increasing attention in recent years, and it has become evident that empirical studies of human-human tutorial dialogue can contribute important insights to the design of computational models of dialogue. Students with particular characteristics may have specific dialogue profiles, and knowledge of such profiles could inform the design of tutorial dialogue systems whose strategies leverage the characteristics of the target population and address the communicative needs of those students. This paper reports on a study that was conducted to investigate the influence of learner characteristics (performance levels, self-efficacy, and gender) on the structure of task-oriented tutorial dialogue. A tutorial dialogue corpus was gathered from interactions transpiring in the course of problem-solving in a learning environment for introductory computer science. Analyses of the annotated dialogues suggest that the dialogue structure of (1) low-performing students differs significantly from that of high-performing students, (2) students with low self-efficacy differs significantly from that of students with high self-efficacy, and (3) males differs significantly from that of females.

1. Introduction

Providing intelligent tutoring systems with the ability to engage learners in rich natural language dialogue has been a goal of the AI & Education community since the inception of the field. With the investigation of tutorial dialogue in a number of systems devised to support a broad range of conversational phenomena (e.g., CIRCSIM [1], BEETLE [2], the GEOMETRY EXPLANATION TUTOR [3], WHY2/ATLAS [4], ITSPOKE [5], SCOT [6], ProPL [7] and AUTOTUTOR [8]), we have begun to the see the emergence of a core set of foundational requirements and functionalities for mixed-initiative natural language interaction. Moreover, recent years have witnessed the appearance of corpus studies empirically investigating speech acts in tutorial dialogue [9], dialogues' correlation with learning [10, 11, 12, 13], student uncertainty in dialogue [14, 15], and comparing text-based and spoken dialogue [5].

While all learners may engage in some common, "universal" form of tutorial dialogue, it may be the case that different populations of learners engage in qualitatively different forms of dialogue. It seems plausible that students with particular characteristics may have specific dialogue profiles, and knowledge of such profiles could inform the design of tutorial dialogue systems whose strategies leverage the characteristics of the target population and address the communicative needs of those students. This paper reports on a study investigating the influence of learners' achievement levels, self-efficacy, and gender on task-oriented tutorial dialogue.

Given that human-human tutorial dialogue offers a promising model for effective communication [16], an experiment was conducted to study naturally occurring tutorial dialogues in a task-oriented learning environment. A text-based dialogue interface was incorporated into a learning environment for introductory computer science. In the environment, students undertook a programming task and conversed with human tutors while designing, implementing, and testing Java programs. To ensure that only natural language was used for communication and to eliminate the possibility of non-verbal communication such as gesture and body language, tutors were physically separated from students in an adjoining lab. The tutors' interface included a real-time synchronized view of the students' problem-solving workspace. Dialogues were logged and the resulting corpus was then manually annotated with tutorial dialogue acts. Analyses of the annotated dialogues suggest that the dialogue structure of low-performing students differs significantly from that of high-performing students, that the dialogue structure of students with low self-efficacy differs significantly from that of students with high self-efficacy, and that the dialogue structure of males differs significantly from that of females.

2. Task-Oriented Tutorial Dialogue Corpus and Dialogue Acts

All tutorial dialogue is inherently task-oriented for it is undertaken in support of learning tasks. However, in contrast to some tutor-student conversations, one genre of tutorial dialogue is directly situated in the task at hand: these dialogues emerge as a result of the creation of learning artifacts such as designs, proofs, or computer programs. The domain investigated in this study, which has also been studied in the ProPL tutorial dialogue project [7], is that of computer programs. Here, students design, implement, and test programs (in the case of the study, Java programs) to meet a given specification. In the course of constructing the artifact, tutors and students pose questions to one another, tutors offer advice and feedback, and students make statements about the artifacts.

The Java Corpus was gathered by logging text-based dialogues between tutors and novice computer science students. (Specifics of the experimental design are described in Section 3.) The learning task was to complete a programming problem that required students to apply fundamental concepts such as iteration, modularization, and sequential-access data structures.

Table 1 presents two sample annotated dialogue excerpts from the Java Corpus. In Dialogue Excerpt A, the tutor interacts with a low performing student, Student A, whose pre-test score was well below the median. The structure of Dialogue A illustrates many features commonly seen with low performing students. Student A responds to the tutor's first question with an unsure answer. After receiving a hint from the tutor, Student A types a proposed problem-solving step into the dialogue interface before implementing it in the problem-solving environment. This pattern of receiving a hint and then requesting feedback repeats. It appears that Student A, who also happens to be in the low self-efficacy group in addition to being in the low performing group, seeks to establish confirmation of his proposed plan before proceeding to implementation. In contrast to Dialogue A, Dialogue B illustrates some common characteristics of dialogues with high performing students. Student B asks a specific question and after receiving tutorial advice begins problem-solving work. Student B does not type his proposed problem-solving step into the dialogue interface to obtain feedback from the tutor; rather, he proceeds directly to implementation.

Table 1: Sample Dialogue Excerpts

Dialogue Excerpt A	Dialogue Excerpt B
Tutor: Do you know how to do that? [TQ] **Student A:** Not really. [A] **Tutor:** Well we first need a new String that will hold zipCode's string value. [HA] **Student A:** So String z = zipCode? [RF] **Tutor:** Close. [PLF] **Tutor:** Then you can set that string equal to ""+zipcode. [HA] **Student A:** Ok so Strong z = ""+zipCode [RF] **Tutor:** Yeah. [PPF] **Student A:** Then what? [TQ] **Tutor:** Ok, so now we need somewhere to keep the individual digits. [A]	**Student B:** So I need an if for each digit? [TQ] **Tutor:** One if should suffice, since it will be called in each iteration. [A] **Tutor:** You just need to know which element to reference. [A] **Tutor:** This would be done in the inner loop. [HA] **Student B:** Ok. [ACK] (*Student works for 2.5 minutes.*) **Tutor:** You've got the right idea. [UPF] **Student B:** Yeah, I had programmers block. [EX] (*Student works for 3 minutes.*) **Tutor:** Perfect. [UPF]

The Java Corpus consists of 5034 dialogue acts: 3075 tutor turns and 1959 student turns. The average number of student turns per dialogue was 56 (SD=24, min=20, max=125), and the average number of tutor turns per dialogue was 86 (SD=27, min=16, max=126). The corpus was manually annotated with a set of tutorial dialogue acts designed to capture the salient characteristics of task-oriented tutorial dialogues. The coding scheme (Table 2) draws on a scheme devised for tutorial dialogue on qualitative physics problems [10]. While most of the acts in this scheme are present in the Java corpus as well, the particular dialogues in the Java corpus made it difficult to make judgements about short answer questions versus deep answer questions and to make fine-grained distinctions between hinting levels. The four-category scheme [9] and a more expansive non-tutorial dialogue act catalogue [17] also contributed commonly occurring acts.

Table 2: Dialogue Acts

	Act	Description	Examples	Counts Student	Counts Tutor	Frequency
Student/Tutor Dialogue Acts	Task Question (TQ)	Questions about goals to pursue, ordering of goals, and the specific problem being solved.	“Where should we start?” “Should we use an array?”	594	313	18.0%
	Concept Question (CQ)	Questions about domain elements, concepts, or facts that would apply over many different problems.	“How do I declare an array?” “I don’t know how to write a loop.”	39	56	1.9%
	Answer (A)	Answers to a task or conceptual question.	“No.” or “Yes.” “We need to give it an index.”	312	695	20.0%
	Acknowledgement (ACK)	Positive acknowledgement of a previous statement.	“Okay.” or “Yeah.” “Alright.”	440	91	10.6%
	Extra Domain (EX)	A statement not related to the computer science discussion.	“Hello.” or “Sorry.” “Nice working with you.”	357	415	15.3%
Student Dialogue Acts	Request Feedback (RF)	A request for evaluative feedback on completed problem solving steps. May also be a request for feedback on a specific proposed problem solving step.	“So should I do array[0] = 1?” “Does that look good?”	98	-	2.0%
	Signal Non-Understanding (SNU)	An indication that a previous statement by the tutor is not clear.	“Kind of makes sense.” “Not really.” or “I’m confused.”	9	-	0.2%
	Statement (S)	Assertion of fact.	“I am going to use a for loop.” “We need to initialize that variable.”	110	-	2.2%
Tutor Dialogue Acts	(Un)Prompted Positive Feedback (UPF/PPF)	Unmitigated positive feedback regarding problem solving action. Prompted is in direct response to a student request for feedback (RF); other feedback is unprompted.	“Good job.” or “Looks great.” “Yep.”	-	270	5.4%
	(Un)Prompted Lukewarm Feedback (ULF/PLF)	Partly positive, partly negative feedback regarding student problem solving action. Prompted is in direct response to a student request for feedback (RF); all other feedback is unprompted.	“The first part is right, but…” “You’re close.” or “Well, almost.”	-	57	1.1%
	(Un)Prompted Negative Feedback (UNF/PNF)	Negative feedback regarding student problem solving action. Prompted is in direct response to a student request for feedback (RF); all other feedback is unprompted.	“No.” “Actually, that won’t work.”	-	36	0.9%
	Hint/Advice (HA)	Problem solving or conceptual hint or advice not in answer to a direct question.	“Each digit is represented by 5 bars.” “Let’s move on.”	-	1115	22.2%
	Request to Confirm Understanding (RCU)	A request for student to confirm or disconfirm understanding.	“Does that make sense?” “Are you with me?”	-	27	0.5%

The corpus was annotated by a single annotator. In an agreement study to evaluate the consistency of the coding scheme and its application to the corpus, the services of a second annotator were enlisted. The second annotator labelled a subset of 969 dialogue acts (of the total 5034 acts in the corpus). This yielded a 0.75 Kappa between the two annotators, indicating a reasonable inter-rater reliability.

3. Experimental Design

Subjects were students enrolled in an introductory computer science course. The course section from which they were drawn was primarily comprised of students who were not computer science majors; the majority of students were engineering majors in disciplines such as mechanical, electrical, and computer engineering. Most subjects were freshmen or sophomores of traditional college age and included 7 women.

The corpus was gathered from tutor-student interactions between 35 students and 6 tutors during a one-week study. Tutors were blind to student characteristics including gender, age, race, self-efficacy, and pre-test performance. Reciprocally, students were not aware of any tutor characteristics. Students and tutors worked together remotely from separate labs. While tutors could observe student problem-solving actions (e.g., programming, scrolling, running programs) in real time, tutors could not themselves take programming actions. The tutors all had some level of tutoring experience. Four were graduate students in computer science, while two were nearing completion of their undergraduate studies. Five were male and one was female. Tutors were given only general guidelines. They were instructed to ask students to explain themselves as often as felt natural, to avoid off-topic discussions, and to try to help with concepts and not merely problem-solving tasks. Tutors were not instructed about specific tutorial strategies.

To begin each session, subjects completed a pre-survey which included items about self-efficacy, attitude toward computer science, and attitude toward collaboration. Subjects then completed a ten item pre-test over specific topic content. All instructions were provided via a three page printed tutorial on the software, accompanied by a three minute instructional video explaining the environment and the learning task. The problem itself was presented on a two page document describing the problem-solving objectives.

The subjects were instructed to greet their tutor through the dialogue interface upon completing the instructional video. This signalled the start of the session, which was controlled at 50 minutes for all subjects. After 50 minutes, the subject was stopped and given a post-survey and post-test containing variants of the items on the pre- versions. Any subject whose session was interrupted due to technical difficulties or external factors was omitted from the data set for analysis. One subject who finished well ahead of time was also discarded from the data set to ensure time on task remained consistent among subjects ($n_{omitted}$=6).

Students worked on a problem-solving task that was part of their regularly scheduled weekly laboratory exercise. The problem required students to take a five-digit "zip code," as used in United States mailing addresses, and translate it into a bar code digit-by-digit. The programming problem was scaffolded with much of its required functionality such as a runtime graphical interface for input and output already provided as source code. Students were asked to complete three modules which involved: (1) extracting each individual digit from a five-digit integer code and placing the digits into a random-access data structure of the student's choosing, (2) writing a single loop to traverse this random-access data structure in order to apply an algorithm which calculated a sixth "correction digit" constructed from the previous five, and (3) writing a nested loop to traverse a two-dimensional array which contained guidelines for representing each individual digit as a sequence of long and short bars in the final bar code.

4. Results and Discussion

To compare dialogue structure based on learner characteristics, three partitioning criteria were applied to the student population: incoming performance level (as indicated by performance on the pre-test), level of self-efficacy (as indicated by rating of self-efficacy in the pre-survey), and gender. After briefly noting overall learning effectiveness, this section reports on dialogue structure characteristics for each student sub-population based on each of the three partitioning criteria.

For each student, learning gain was gauged by the difference between pre and post-test scores. On average, students scored 13% higher on the post-test than the pre-test. A pair-wise difference t-test indicates that the difference is statistically significant with $p < 0.05$. This learning gain was statistically significant among both high performing and low performing groups as well as among high self-efficacy and low self-efficacy groups. Both gender groups exhibited learning gains, but the female population was too small to yield a statistically significant learning gain.

4.1 Dialogue Profile Analyses

For each student dialogue session, the relative frequency of each dialogue act was computed, i.e., the ratio of the number of occurrences of that dialogue act to the total number of dialogue acts in the session. The population was partitioned according to three criteria: performance level (partitioning by median on pre-test), self-efficacy level (by median on efficacy scores, described below), and gender. The relative frequency of dialogue acts was then computed for high-performing and low-performing students, for high-efficacy and low-efficacy students, and for female and male students. To determine whether intra-group differences in means were significant, t-tests were performed. For each pair-wise t-test, an F-test was performed for equality of variance and the result was used to determine whether a t-test with pooled variance or unequal variance was most suitable. Table 3 summarizes the relative frequency results. Bolded values indicate statistically significant differences on pooled variance or unequal variance one-tailed t-tests ($p \leq 0.05$). It should be noted that the three partitions are not independent. For example, high performing students were more often in the higher self-efficacy group for the reported data. Similarly, the majority of females in the study were in the low performing group. Despite these confounds, we draw meaningful conclusions by examining each learner characteristic and its impact on dialogue structure individually.

Students were divided into low performing and high performing groups based on the median pre-test score of 60%. Analyses yielded the following findings:

- High performing students had a higher relative frequency of acknowledgements than low performing students.
- Low performing students requested feedback more often and made declarative statements less often than high performing students.
- Tutors paired with low performing students made more extra-domain statements and gave more prompted feedback.
- Tutors paired with low performing students made many more requests for confirmation of understanding from their students.

Following an instrument devised by Bandura to measure domain-specific self-efficacy [18], students were asked to rate their confidence in being able to complete a programming assignment under a variety of circumstances. Because the problem used for this study was drawn from a standard problem set for the course, students had an experiential basis on which to judge their ability to complete the problem. Statistically significant differences in dialogue structure emerged when students were grouped by their confidence level regarding whether they could complete a simple laboratory assignment on their own. Analyses yielded the following findings:

- Students in the high self-efficacy group made more declarative statements, or assertions, than students in the low self-efficacy group.
- Tutors paired with low self-efficacy students gave more negative feedback.
- Tutors paired with high self-efficacy students made more acknowledgements than tutors paired with low self-efficacy students.

Despite the fact that females comprised a small number of our subjects (n_{female}=7; n_{male}=28), several strong statistically significant results emerged. Analyses yielded the following findings:

- Women made more requests for feedback than men.
- Men made more declarative statements than women.
- Tutors paired with women gave more positive feedback than tutors paired with men.
- Tutors paired with women made more requests to confirm understanding.

4.2 Discussion

These findings extend those of previous studies investigating tutorial dialogue and learning effectiveness which have found correlations of dialogue structure and content with learning [11, 12, 13]. Of particular interest is a large spoken tutorial dialogue study conducted as part of the ITSPOKE project [10]. While the domains of the studies are different (qualitative physics vs. computer programming) and the purpose of the dialogues are different (detecting misconceptions and supplying additional knowledge vs. scaffolding novice programmers' computer science problem-solving activities), correlations were found in each case. The ITSPOKE study found that student utterances exhibiting reasoning and reasoning-oriented questions posed by the tutor were positively correlated with learning in a human-computer corpus, as were the introduction of new concepts in the dialogue by students in a human-human corpus. The Java Corpus study reported on here found that learner characteristics appear to significantly affect the structure of tutorial dialogue, and that both tutor and student dialogue acts appear to be affected by these differences. Tutors more often engaged in more extra-domain conversation, provided additional feedback, and more frequently engaged in discussions to gauge students' level of understanding when conversing with low performing, low efficacy, or female students. These same groups of students tended to request more feedback, make fewer declarative statements, and make fewer acknowledgements. It seems likely that learner characteristics affect (and are affected by) tutorial dialogue issues analogous to those bearing on help-seeking behaviors [19] and self-explanation [20].

Table 3: Dialogue Profile Analysis

		Relative Frequencies					
	Dialogue Act	Pre-test Performance n_{high}=17, n_{low}=18		Self-efficacy Level n_{high}=19, n_{low}=16		Gender n_{female}=7, n_{male}=28	
Tutor and Student	Student Task Question (TQ)	High	11.2%	High	11.1%	Female	12.8%
		Low	12.1%	Low	12.3%	Male	11.4%
	Tutor Task Question (TQ)	High	6.7%	High	6.2%	Female	6.5%
		Low	6.4%	Low	6.9%	Male	6.5%
	Student Concept Question (CQ)	High	0.7%	High	0.9%	Female	0.6%
		Low	0.9%	Low	0.6%	Male	0.8%
	Tutor Concept Question (CQ)	High	1.1%	High	0.8%	Female	1.1%
		Low	1.0%	Low	1.3%	Male	1.0%
	Student Answer (A)	High	6.9%	High	5.8%	Female	6.7%
		Low	5.7%	Low	6.8%	Male	6.2%
	Tutor Answer (A)	High	13.0%	High	13.2%	Female	14.8%
		Low	14.5%	Low	14.4%	Male	13.5%
	Student Acknowledgement (ACK)	**High**	**10.3%**	High	9.1%	Female	8.3%
		Low	**6.3%**	Low	7.3%	Male	8.2%
	Tutor Acknowledgement (ACK)	High	2.3%	**High**	**2.5%**	Female	1.2%
		Low	1.5%	**Low**	**1.2%**	Male	2.1%
	Student Extra Domain (EX)	High	8.8%	High	8.6%	Female	6.0%
		Low	6.1%	Low	6.0%	Male	7.8%
	Tutor Extra Domain (EX)	**High**	**6.9%**	High	8.5%	Female	8.4%
		Low	**9.4%**	Low	7.8%	Male	8.2%
Student	Request Feedback (RF)	**High**	**1.2%**	High	1.6%	**Female**	**2.8%**
		Low	**2.2%**	Low	2.0%	**Male**	**1.5%**
	Statement of Non-Understanding (SNU)	High	0.1%	High	0.2%	Female	0.3%
		Low	0.3%	Low	0.2%	Male	0.1%
	Statement (S)	**High**	**3.5%**	**High**	**3.5%**	**Female**	**1.2%**
		Low	**1.6%**	**Low**	**1.4%**	**Male**	**2.8%**
Tutor	Unprompted Positive Feedback (UPF)	High	4.6%	High	4.5%	Female	3.8%
		Low	4.3%	Low	4.4%	Male	4.6%
	Prompted Positive Feedback (PPF)	**High**	**0.5%**	High	0.9%	**Female**	**1.7%**
		Low	**1.5%**	Low	1.3%	**Male**	**0.9%**
	Unprompted Lukewarm Feedback (ULF)	High	0.8%	High	0.7%	Female	0.7%
		Low	0.6%	Low	0.6%	Male	0.7%
	Prompted Lukewarm Feedback (PLF)	**High**	**0.1%**	High	0.2%	Female	0.6%
		Low	**0.6%**	Low	0.5%	Male	0.3%
	Unprompted Negative Feedback (UNF)	High	0.6%	High	0.7%	Female	0.5%
		Low	0.5%	Low	0.4%	Male	0.5%
	Prompted Negative Feedback (PNF)	**High**	**0.1%**	**High**	**0.0%**	Female	0.3%
		Low	**0.3%**	**Low**	**0.4%**	Male	0.1%
	Hint/Advice (HA)	High	20.5%	High	20.8%	Female	20.5%
		Low	23.5%	Low	23.5%	Male	22.4%
	Request Confirmation of Understanding (RCU)	**High**	**0.1%**	High	0.3%	**Female**	**1.4%**
		Low	**0.9%**	Low	0.8%	**Male**	**0.3%**

These findings suggest that it may be possible to devise tutorial dialogue strategies that address the specific communicative needs of different groups of learners. Putting gender differences aside because of the limited data, several design implications should be considered for tutorial dialogue systems. The following recommendations suggest how the findings might be embodied in the tutorial strategies of a dialogue manager whose objectives are to improve learning effectiveness while creating a motivating problem-solving experience:

- **Encouraging Reflection:** If a student with low incoming performance is proceeding through the tutoring session while initiating few requests for feedback, the system should consider taking remedial action such as asking task questions or concept questions to assess student understanding. Such actions may encourage students to reflect on their problem-solving experience, possibly playing a similar reflective role as student requests for feedback.
- **Giving Adequate Feedback:** Systems should be prepared to give prompted feedback more often when working with low-performing or low-efficacy students. The nature and effects of this feedback, e.g., whether the student experience could be enhanced by mitigating the frequent negative feedback which may come naturally to human tutors when working with weaker students, is an important topic for further study.
- **Confirming Students' Understanding:** When working with students of low incoming performance levels, systems should be prepared to make more requests for confirmation of the students' understanding as a follow up to tutor hints or advice.
- **Making Acknowledgements:** When interacting with high-efficacy students, systems should give more acknowledgements than in their default setting; this may more accurately reflect the interaction such students would expect when working with a human tutor.
- **Maintaining Conversational Comfort:** When interacting with students who have been deemed to be low-performing prior to the tutoring session, systems should consider making slightly more extra-domain statements, which could create a more comfortable and conversational setting in which weaker students might feel more at ease.

5. Conclusion

Tutorial dialogue exhibits structural regularities that cut across learning tasks and domains. However, learner characteristics may profoundly affect the structure of tutor-student conversations. Analyses of task-oriented tutorial dialogues indicate that students' incoming performance levels, levels of self-efficacy, and gender significantly influence the structure of dialogue. The findings suggest that learner characteristics may be considered in designing tutorial dialogue strategies that more effectively target the specific needs of students with particular characteristics. While the current study provides insight into the structure of tutorial dialogue, the analyses focus exclusively on tutorial dialogue acts but do not consider other linguistic features of the utterances or the artifacts produced by the students (e.g., computer programs) in the course of completing the learning task.

The study reported here represents a first step toward understanding how learner characteristics affect the structure of tutorial dialogue. Several directions for future work appear promising. First, it will be important to explore the influence of learner characteristics on tutorial dialogue in the presence of surface level information about students' utterances. This line of investigation is of particular interest given recent results indicating that lexical cohesion in tutorial dialogue with low-performing students is found to be highly correlated with learning [21]. Second, a comparative analysis of alternate tutoring strategies on the effectiveness and efficiency of student learning will yield a clearer picture of the space of tutorial dialogue. By conducting a series of studies in which tutoring protocols are systematically varied in conjunction with targeted learner characteristics, we can observe the effects on learning outcomes and better understand which approaches to tutoring are most appropriate for specific populations of students. Third, students' motivation and frustration undoubtedly influence (and are influenced by) the structure and content of tutorial dialogue, so developing a better understanding of these interrelationships will contribute to more effective tutorial dialogue management.

Acknowledgements

The authors wish to thank Tiffany Barnes and the members of the IntelliMedia Center for Intelligent Systems at NC State University for their insightful discussions on this research, along with the software development team of August Dwight and Taylor Fondren. Special thanks to Scott McQuiggan for assistance in framing the efficacy questions of this work and for assistance in preparing the manuscript. This work was supported in part by National Science Foundation Grants CNS-0540523 and REC-0632450. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Evens and J. Michael, *One-on-One Tutoring by Humans and Computers.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 2006.

[2] C. Zinn, J. D. Moore and M. G. Core, "A 3-tier planning architecture for managing tutorial dialogue," *Intelligent Tutoring Systems, Sixth International Conference (ITS 2002),* 2002.

[3] V. Aleven, K. Koedinger and O. Popescu, "A Tutorial Dialog System to Support Self-explanation: Evaluation and Open Questions," *Proceedings of the 11th International Conference on Artificial Intelligence in Education,* pp. 39-46, 2003.

[4] K. VanLehn, P. W. Jordan, C. P. Rose, D. Bhembe, M. Bottner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg and A. Roque, S. Siler, R. Srivastava "The architecture of Why2-Atlas: A coach for qualitative physics essay writing," *Proceedings of the 6th International Conference on Intelligent Tutoring Systems,* pp. 158–167, 2002.

[5] D. J. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe and S. Silliman, "Spoken Versus Typed Human and Computer Dialogue Tutoring," *International Journal of Artificial Intelligence in Education,* vol. 16, pp. 145-170, 2006.

[6] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark and S. Peters, "Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems," *International Journal of Artificial Intelligence in Education,* vol. 16, pp. 171-194, 2006.

[7] H. C. Lane and K. VanLehn, "Teaching the Tacit Knowledge of Programming to Novices with Natural Language Tutoring," *Computer Science Education,* vol. 15, pp. 183-201, 2005.

[8] A. Graesser, G. Jackson, E. Mathews, H. Mitchell, A. Olney, M. Ventura, P. Chipman, D. Franceschetti, X. Hu, M.M. Louwerse, N.K. Person, and the Tutoring Research Group, "Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog," *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society,* pp. 474-479, 2003.

[9] J. Marineau, P. Wiemer-Hastings, D. Harter, B. Olde, P. Chipman, A. Karnavat, V. Pomeroy, S. Rajan, A. Graesser, and the Tutoring Research Group, "Classification of speech acts in tutorial dialog," in *Proceedings of the Workshop on Modelling Human Teaching Tactics and Strategies of ITS-2000,* pp. 65-71, 2000.

[10] K. Forbes-Riley, D. Litman, A. Huettner and A. Ward, "Dialogue-learning correlations in spoken dialogue tutoring," *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005),* pp. 225-232, 2005.

[11] M. G. Core, J. D. Moore and C. Zinn, "The role of initiative in tutorial dialogue," *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics,* pp. 67-74, 2003.

[12] C. Rose, D. Bhembe, S. Siler, R. Srivastava and K. VanLehn, "The Role of Why Questions in Effective Human Tutoring," *Proceedings of Artificial Intelligence in Education,* 2003.

[13] S. Katz, D. Allbritton and J. Connelly, "Going Beyond the Problem Given: How Human Tutors Use Post-Solution Discussions to Support Transfer," *International Journal of Artificial Intelligence in Education,* vol. 13, pp. 79-116, 2003.

[14] J. Liscombe, J. Hirschberg and J. Venditti, "Detecting Certainness in Spoken Tutorial Dialogues," *Proceedings of Interspeech,* 2005.

[15] K. Forbes-Riley and D. Litman, "Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus," *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue,* 2005.

[16] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann, "Learning from human tutoring," *Cognitive Science,* vol. 25, pp. 471-533, 2001.

[17] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics,* vol. 26, pp. 339-373, 2000.

[18] A. Bandura, "Guide for constructing self-efficacy scales," in *Self-Efficacy Beliefs of Adolescents,* F. Pajares and T. Urdan, Eds. Greenwich, Connecticut: Information Age Publishing, 2006, pp. 307-337.

[19] V. Aleven, B. McLaren, I. Roll and K. Koedinger, "Toward tutoring help seeking: applying cognitive modeling to meta-cognitive skills," *Proceedings of the 7th International Conference on Intelligent Tutoring Systems,* pp. 227-239, 2004.

[20] M. T. H. Chi, N. Leeuw, M. H. Chiu and C. LaVancher, "Eliciting Self-Explanations Improves Understanding," *Cognitive Science,* vol. 18, pp. 439-477, 1994.

[21] A. Ward and D. Litman, "Cohesion and learning in a tutorial spoken dialog system," *Proceedings of the 19th International FLAIRS (Florida Artificial Intelligence Research Society) Conference,* 2006.