Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference

Recognizing Effective and Student-Adaptive Tutor Moves in Task-Oriented Tutorial Dialogue

Christopher M. Mitchell, Eun Young Ha, Kristy Elizabeth Boyer, James C. Lester

Department of Computer Science, North Carolina State University, Raleigh, NC, USA

{cmmitch2, eha, keboyer, lester}@ncsu.edu

Abstract

One-on-one tutoring is significantly more effective than traditional classroom instruction. In recent years, automated tutoring systems are approaching that level of effectiveness by engaging students in rich natural language dialogue that contributes to learning. A promising approach for further improving the effectiveness of tutorial dialogue systems is to model the differential effectiveness of tutorial strategies, identifying which dialogue moves or combinations of dialogue moves are associated with learning. It is also important to model the ways in which experienced tutors adapt to learner characteristics. This paper takes a corpus-based approach to these modeling tasks, presenting the results of a study in which task-oriented, textual tutorial dialogue was collected from remote one-on-one human tutoring sessions. The data reveal patterns of dialogue moves that are correlated with learning, and can directly inform the design of student-adaptive tutorial dialogue management systems.

1. Introduction

One-on-one tutoring is highly effective (Bloom, 1984; VanLehn et al., 2007). While the mechanisms that enable such effectiveness are not fully understood, they are explained in part by the rich interactions between students and tutors (M. T. H. Chi et al., 2001), adaptive presentation of instructional material (D'Mello, Hays, et al., 2010), motivational strategies (Lepper et al., 1993), and the exchange of rich natural language dialogue (Graesser, Person, and Magliano, 1995; Litman et al., 2009). Natural language tutorial dialogue has been studied extensively in an effort to develop tutorial dialogue systems that are highly effective, and significant progress has been made toward that goal, as evidenced by existing tutorial dialogue systems (M. Chi, VanLehn, and Litman, 2010; Di Eugenio et al., 2011; Dzikovska et al., 2010; D'Mello, Lehman, and Graesser, 2011; Forbes-Riley and Litman, 2009; Kumar et al., 2010).

Today's tutorial dialogue systems feature increasingly sophisticated dialogue, but they have not yet matched the effectiveness of expert human tutors for facilitating student learning (VanLehn et al., 2007). A promising approach for improving the effectiveness of tutorial dialogue systems is to model the association between tutoring strategies and desired outcomes such as learning gains (M. Chi et al. 2010; Di Eugenio et al. 2011; Ohlsson et al. 2007; Forbes-Riley and Litman 2009). Such an approach does not simply assume that the tutoring strategies used most frequently by humans are the most effective, but rather, identifies the most effective strategies by representing them as sets of features and building predictive models of target outcomes.

Much prior work has proceeded by assuming that the actions taken most frequently by human tutors are the most effective. Studying human tutorial dialogue in this way can yield useful insights into the collaborative patterns involved in tutorial dialogue, and into the approaches of both expert and non-expert tutors (e.g., (Boyer, Vouk, and Lester, 2007; D'Mello, Olney, and Person, 2010)). However, it may ultimately be the case that modeling differential effectiveness is the key to building models of tutorial dialogue that approach optimality.

Another aspect of effective adaptation within tutorial dialogue is to consider learner characteristics such as gender, self-efficacy, and incoming knowledge level. It is known that individual differences influence the structure of tutorial dialogue (D'Mello et al., 2009), and as such, these differences can suggest important adaptations that tutorial dialogue management systems might undertake.

This paper explores how experienced human tutors adapt to learner characteristics, and presents a correlational study between student and tutor dialogue moves and learning. The analyses are conducted on a corpus of human-human textual tutorial dialogue for introductory computer science. Dialogue moves are examined at the unigram level (individual dialogue acts) and the bigram level (pairs of adjacent dialogue acts). This work extends prior work that examined the associations between dialogue acts and tutoring effectiveness within task-oriented textual human-human tutoring of introductory computer science (Boyer et al., 2010). These analyses are part of the larger JavaTutor project, which aims to create a tutorial dialogue system that learns its behavior from corpora with experienced human tutors. The results add to the body of knowledge regarding ways in which tutorial dialogue is adapted to learner characteristics, and build on prior work that has suggested ways in which particular tutoring strategies are associated with learning.

2. Related Work

From the early days of tutorial dialogue research, it has been recognized that studying human tutoring is a promising approach to discovering effective strategies that can be used in intelligent systems (Fox 1993; D'Mello, Olney, and Person 2010b; Graesser et al. 1995; Di Eugenio et al. 2011). Because of the proven effectiveness of human tutoring, some work examined the actions of human tutors and adopted the premise that what humans did more frequently, systems ought to do as well. This has been referred to as the "code-and-count" approach (Ohlsson et al., 2007). Assuming that human tutors' actions are effective can be a reasonable step, particularly when the tutors being studied are highly experienced and have been proven effective over time (Cade, Copeland, and Person, 2008). For example, studying expert human tutors has recently yielded insights into the potential importance of off-topic conversation during tutoring (Lehman, Cade, and Olney, 2010), and has suggested ways in which tutors convey information via "collaborative lecture" (D'Mello, Hays, et al., 2010).

However, there is growing recognition that human tutors vary in their effectiveness. For example, there is sometimes not a clean distinction between the effectiveness of expert and non-expert tutors, e.g., (Cohen, Kulik, and Kulik, 1982; Di Eugenio et al., 2011; Evens and Michael, 2005). Indeed, as will be discussed in Section 3.2, the learning gains achieved by the most expert tutor in the current corpus did not exceed those of significantly less expert tutors. For this reason, it is important to model the differential effectiveness of tutoring approaches—that is, to identify which strategies or dialogue structures are associated with effective learning.

Dialogue has been found to correlate with learning at a variety of levels. These include tutors adapting to uncertainty (Forbes-Riley and Litman 2009), providing direct procedural instruction (Di Eugenio et al., 2011), eliciting information from a student (M. Chi et al., 2010), and making social dialogue moves when working with team of tutees (Kumar et al., 2010). Student moves have also been shown to correlate with learning; for example, expressions of disengagement (Forbes-Riley and Litman 2011) and negative social talk (Dzikovska et al., 2010) may be associated with decreased learning, while student utterances displaying reasoning may be correlated with increased learning (Litman and Forbes-Riley 2006).

This paper builds on prior work by examining the tutorial dialogue exchanged between experienced human tutors and novice computer science students through remote textual dialogue. The textual modality was selected because of the broader project goal of directly learning a tutorial dialogue management model for a text-based tutorial dialogue system from the collected corpus. The findings reveal correlations between unigrams and bigrams of dialogue acts with learning outcomes. Additionally, the analyses reveal ways in which learner characteristics such as self-efficacy, incoming knowledge level, and gender are associated with dialogue structure.

3. Corpus and Annotation

The corpus collected for this work consists of human-human tutorial interactions conducted within a web-based remote tutoring interface for Java programming. The dialogue is text-based. The tutoring interface (Figure 1) consists of four panes that display the interactive components of the task-oriented tutoring: the current programming task description, the student's Java code, the compilation or execution output associated with the code, and the textual dialogue messages between the student and tutor. The content of the tutor and student's interfaces were synchronized in real time. In addition to conversing via text with the tutor, the student also modified, compiled and ran programming code within the interface. The tutors' actions were constrained to conversation with the student and advancing to the next task, but they could see all student progress in real time.

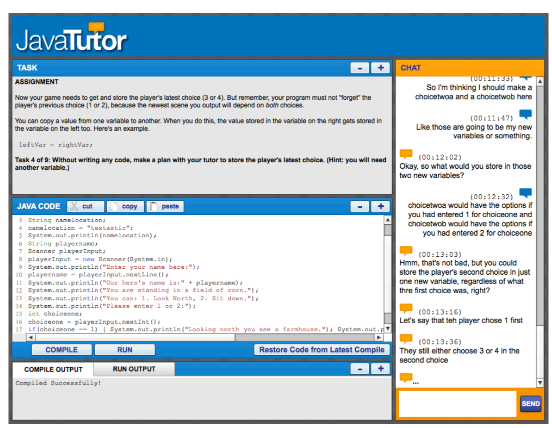

Figure 1. The JavaTutor remote tutoring interface

3.1 Study Design

The study paired 21 students with one of 4 tutors for 6 hour-long lessons on introductory Java programming over 4 weeks. The students were selected from a first-year engineering course and were pre-screened to eliminate those with significant prior programming experience. The students received full credit for one-half of their semester project in the engineering course in return for their participation. The tutors were graduate students with prior tutoring experience in Java programming. This paper reports on analyses of the first of six lessons (the JavaTutor Lesson 1 corpus) for each tutor/student pair.

3.2 Data

The JavaTutor Lesson 1 corpus consists of 2564 utterances: 1777 tutor utterances and 787 student utterances. An excerpt from the corpus is displayed in Table 1. The average number of utterances per tutoring session was 122 (min=74; max=201). The average number of tutor utterances per session was 84.6 (min=51; max=137) and the average number of student utterances per session was 37.4 (min=22; max=64).

Table 1. Excerpt with dialogue act tags

Tutor: hang on :) [S]
Tutor: When we show you example code, it is not the code you need to write. [S]
Tutor: Look at the task again. [H]
Tutor: YUP [PF]
Tutor: Perfect [PF]
Tutor: OK. Go ahead and test. [DIR]
Student: And I don't need anything in the parentheses? [Q]
Tutor: Line 9 is correct. You do NOT need anything inside the parentheses. [A]
Student: Ok [ACK]
Tutor: Good. [PF]
Tutor: Moving on. [S]
Tutor advances to the next task.
Tutor: Syntactically correct. But there is a logic error [LF]
Tutor: When will the output statement display your request to the player? [Q]
Student: AFTER they put in their name [A]
Tutor: Exactly [PF]

Students completed an identical pretest and posttest for each lesson to measure learning gain. The average pretest score was 52.6% (min=23.5%; max=100%), while the average posttest score was 77.6% (min=41.1%; max=100%). Learning gain was calculated as *posttest-pretest*, and normalized learning gain for each student was calculated as in Equation (1), which features an adjustment for non-positive learning gain to avoid division by zero when posttest and pretest scores were equal (one such occurrence in the data set). This formula was derived from (Marx and Cummings, 2007).

$$norm\_gain = \begin{cases} \dfrac{posttest - pretest}{1 - pretest}, & posttest > pretest \\ \dfrac{posttest - pretest}{pretest}, & posttest \leq pretest \end{cases} \quad (1)$$

The average normalized learning gain for the JavaTutor Lesson 1 corpus was 52.2% (min=-23.5%; max=100%). Interestingly, the most experienced tutor's average learning gain of 42.3% was significantly lower than the learning gain of 72.2% for all other less experienced tutors combined (p=0.036), a finding similar to those that have been observed in other studies (Di Eugenio et al., 2011). For the remainder of the analyses in this paper, learning gains are pooled across all tutors.

Student characteristics including self-efficacy for computer science and gender were collected via a survey prior to the first tutoring session (tutors did not have access to any survey or pretest data). For the analyses reported in this paper, computer science self-efficacy was calculated as the mean of the student's responses to six Likert-scale items (Table 2). The average self-efficacy score was 3.39 (min=2.33, max=4.33) out of a possible 5.

Table 2. Domain-specific self-efficacy survey questions

Generally I have felt secure about attempting computer programming problems.
I am sure I could do advanced work in computer science.
I am sure that I can learn programming.
I think I could handle more difficult programming problems.
I can get good grades in computer science.
I have a lot of self-confidence when it comes to programming.

3.3 Dialogue Act Annotation

A dialogue act annotation protocol was devised and applied to every utterance in order to capture salient events in the textual dialogue corpus. This annotation scheme was an extension of a prior annotation scheme for task-oriented tutorial dialogue (Boyer et al., 2010). Three human annotators were trained in an iterative process that included collaborative tagging, refinement of the protocol, and independent tagging. A list of the tags in the annotation scheme is shown in Table 3.

The majority of the annotations used as a basis for the analysis reported here were completed by one annotator, with independent annotations provided by two secondary annotators. Twenty-four percent of the corpus not used during the training process was annotated independently by two of the annotators, yielding a Cohen's kappa of 0.79.

Table 3. The dialogue act annotation scheme

Tag	Description	Freq.
H	**Hint** - The tutor gives advice to help the student proceed with the task	**T:** 133 **S:** 0
DIR	**Directive** - The tutor explicitly tells the student the next step to take	**T**: 121 **S**: 0
ACK	**Acknowledgement** of previous utterance; conversational grounding	**T**: 41 **S**: 175
RC	**Request Confirmation** from the other participant (e.g., *"Make sense?"*)	**T**: 11 **S**: 0
RF	**Request for Feedback** - The student requests an assessment of his performance or his work from the tutor	**T**: 0 **S**: 7
PF	**Positive Feedback** - The tutor gives a positive assessment of the student's performance	**T**: 327 **S**: 0
LF	**Lukewarm Feedback** - The tutor gives an assessment that has both positive and negative elements	**T**: 13 **S**: 0
NF	**Negative Feedback** - The tutor gives a negative assessment of the student's performance	**T**: 1 **S**: 0
Q	**Question** - A question which does not fit into any of the above categories	**T**: 327 **S**: 120
A	**Answer** - An answer to an utterance marked Q	**T**: 96 **S**: 295
C	**Correction** of a typo in a previous utterance	**T**: 10 **S**: 6
S	**Statement** - A statement of fact which does not fit into any of the above categories	**T**: 681 **S**: 174
O	**Other** utterances, usually containing only affective content	**T**: 6 **S**: 10

4. Results

The primary goal of the analysis is to determine which dialogue moves, or pairs of them, are associated with student learning. It is also desirable to examine relationships between dialogue structure and learner characteristics such as domain-specific self-efficacy. This section presents correlational analyses between dialogue moves and learning, and between learner characteristics and dialogue structure.

These correlations were computed using the relative frequency of each dialogue act, including its speaker identifier. This relative frequency represents a move's percentage of overall dialogue acts in a session. For example, the relative frequency of tutor ACKNOWLEDGEMENTs in a given dialogue was computed as number of tutor ACKNOWLEDGEMENTs divided by the total number of moves in that dialogue.

First, correlations were computed between the relative frequency of individual (unigram) dialogue acts and the factors of interest: normalized learning gain, self-efficacy and incoming knowledge level as measured by pretest score (Table 4). Two tutor dialogue moves were negatively correlated with learning: tutor DIRECTIVES (*e.g., "Now press the compile button."*) and tutor REQUESTS FOR CONFIRMATION (*e.g., "Make sense?"*). Two other tutor moves were correlated with learner characteristics: tutors gave more ANSWERS to student with lower pretest scores, and gave more HINTS to students with lower self-efficacy. No student moves were associated with learning gain, but two were correlated with learner characteristics. Students with low pretest score made more REQUESTS FOR FEEDBACK, and students with low self-efficacy made more OTHER moves.

Table 4. Dialogue act unigram correlations
(*n* = total counts in corpus, *r* = correlation coefficient)

Factor	Dialogue Act	*r*	*p*	*n*
Normalized Learning Gain	DIR (Tutor)	-.54	.010	121
	RC (Tutor)	-.45	.040	11
Pretest Score	C (Student)	.46	.033	6
	RF (Student)	-.44	.048	7
	A (Tutor)	-.46	.036	96
Self-Efficacy	O (Student)	.19	.040	10
	H (Tutor)	-.41	.006	133

In order to gain further insight into the structure of the tutorial dialogues and its association with learning and learner characteristics, the same analysis was performed using the relative frequencies of pairs of adjacent dialogue acts (bigrams) (Table 5). Two bigrams were negatively correlated with learning: a student ANSWER followed by a tutor HINT and a tutor DIRECTIVE followed by another tutor DIRECTIVE. A tutor was more likely to give POSITIVE FEEDBACK after providing an ANSWER to a student's question if the student's pretest score for that lesson was lower (the tutors did not have access to these pretest scores). Additionally, the relative frequencies of three bigrams were found to correlate with a lower domain-specific self-efficacy: tutor ANSWER followed by a DIRECTIVE, tutor HINT followed by another tutor HINT, and a tutor HINT followed by a tutor STATEMENT. An excerpt containing of each of these bigrams is shown in Table 6.

Finally, we partitioned students by gender to identify differences between the dialogue structures of male and female students. Female students tended to request more feedback than males. The average relative frequency for RF student dialogue moves for women was 0.82%, versus 0.09% for men (p=0.045). Male learners made more acknowledgements (relative frequency = 7.3% compared to 5.1% for women; p=0.036). Finally, males also corrected their own utterances in a subsequent utterance more, a common behavior in textual instant messaging (relative frequency 0.38% for males compared with 0% for women; p=0.03).

Table 5. Dialogue act bigram correlations (DA = dialogue act, St = student, Tu = tutor, *n* = total counts in corpus, *r* = correlation coefficient)

Factor	DA 1	DA 2	*r*	*p*	*n*
Norm. Learning Gain	A (St)	H (Tu)	-.52	.016	19
	DIR (Tu)	DIR (Tu)	-.60	.004	13
Pretest	A (Tu)	PF (Tu)	-.48	.037	11
Self-Efficacy	A (Tu)	DIR (Tu)	-.50	.022	5
	H (Tu)	H (Tu)	-.48	.026	42
	H (Tu)	S (Tu)	-.49	.023	9

5. Discussion

A promising approach to developing effective tutorial dialogue management strategies is to identify dialogue structures that are associated with learning. With this goal, the analysis presented in this paper explored correlations between learning and unigrams or bigrams of dialogue acts. Several significant correlations emerged.

Table 6. Excerpts from the corpus including bigrams from Table 5 (note: typographical errors originated in corpus)

Excerpt A	**Tutor:** ready? [Q] **Student:** yep [A] *Tutor advances to next task.* **Tutor:** compare with teh example [H] **Tutor:** you need a + [H]
Excerpt B	**Tutor:** comment out line 8. [DIR] **Tutor:** now compile. [DIR]
Excerpt C	**Student:** should I complile it before running it again? [Q] **Tutor:** yes. [A] **Tutor:** excellent. [PF]
Excerpt D	**Tutor:** well, java is waiting at the nextLIne() [A] **Tutor:** so type anything now [DIR]
Excerpt E	**Tutor:** on right side of = should be exactly like what is given. [H] **Tutor:** java will stop when it gets to the nextLine() and wait for user to type something [S]

As shown in Table 5, the {ANSWER (Student), HINT (Tutor)} bigram was negatively correlated with normalized learning gain. Nearly all instances of this bigram occurred immediately before or after the tutor advanced to the next task, as shown in Table 6, Excerpt A. In these cases, the ANSWER was a confirmation that the student was ready to proceed to the next task, and the HINT was providing advice on the newly presented task. These hints often occurred a full minute or more after the task was advanced, indicating that the student was being allowed to work independently on the task without engaging with the tutor. In contrast, the student and tutor might engage in a dialogue discussing the new task after it was presented, or a tutor could provide POSITIVE FEEDBACK on student progress, indicating that the student had completed the task without the tutor's assistance. These results suggest that providing unsolicited hints may be less effective than discussing the task with the student or providing tutorial advice only when requested by the student. This explanation is also supported by the stronger negative correlation between learning and tutor DIRECTIVEs in both unigram (Table 4) and bigram (Table 5) analyses. A DIRECTIVE was normally a stronger version of a HINT, and allowed little room for interaction or initiative from the student (Table 6, Excerpts B and D).

In addition to revealing correlations between dialogue structure and learning as discussed above, the analyses indicated ways in which learner characteristics such as self-efficacy, gender, and knowledge level are associated with dialogue structure. Students with lower self-efficacy receive more tutor hints (both unigrams and bigrams). While it is the case that pretest score and self-efficacy are associated in this corpus, which can partly explain this finding, it also suggests that the tutors were sensitive to the confidence level of the students and provided extra cognitive support for these students. The OTHER utterances observed more often with high self-efficacy students were mainly affective utterances (*e.g.* "hahaha" or "lol"). It is plausible that increased self-efficacy related to the subject matter translated into heightened comfort level during tutoring, allowing students to engage in conversational behaviors such as these. Interestingly, these utterances display significant regularities that may facilitate automatic recognition by an intelligent system, potentially leading to automatic inference of a learner's self-efficacy based on the dialogue.

6. Conclusion and Future Work

Building an automated tutor with the same effectiveness as an expert human tutor remains a major goal in ITS research. A highly promising approach is to model the ways in which human tutorial dialogue moves are correlated with learning, to enable subsequent implementation of the most effective strategies within an intelligent system. This paper has described the collection and annotation of a corpus of textual task-oriented tutorial dialogue, and explored several significant correlations between dialogue and learning. Additionally, the analyses revealed ways in which learner characteristics influence the structure of tutorial dialogue. These findings pave the way for confirmatory future investigations across tutorial domains and for larger populations of students.

Important directions for future work include investigating the impact of the affective properties of tutor and student dialogue on learning gains. This direction holds particular promise when a tutor and student interact repeatedly, allowing the tutor to form a long-term student model and establishing a robust rapport. Finally, the strategies discovered in this and other analyses of human tutoring must be tested in the context of human-computer tutoring. It is hoped that these lines of investigation will enable the creation of highly effective tutorial dialogue systems by modeling the differential effectiveness of tutoring strategies, and enabling fine-grained adaptation to learner characteristics.

Acknowledgments

This work is supported in part by the National Science Foundation through Grants IIS-0812291, DRL-1007962 and CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-one Tutoring. *Educational Researcher*, *13*(6), 4–16.

Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M., and Lester, J. (2010). Characterizing the Effectiveness of Tutorial Dialogue with Hidden Markov Models. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 55–64).

Boyer, K. E., Vouk, M. A., and Lester, J. C. (2007). The Influence of Learner Characteristics on Task-oriented Tutorial Dialogue. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 365–372).

Cade, W., Copeland, J., and Person, N. (2008). Dialogue Modes in Expert Tutoring. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 470–479).

Chi, M. T. H., Siler, S. a., Jeong, H., Yamauchi, T., and Hausmann, R. G. (2001). Learning from Human Tutoring. *Cognitive Science*, *25*(4), 471-533.

Chi, M., VanLehn, K., and Litman, D. (2010). Do Micro-Level Tutorial Decisions Matter: Applying Reinforcement Learning To Induce Pedagogical Tutorial Tactics. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 224-234).

Cohen, P. A., Kulik, J. A., and Kulik, C.-L. C. (1982). Educational Outcomes of Tutoring: A Meta-analysis of Findings. *American Educational Research Journal*, *19*(2), 237-248.

Di Eugenio, B., Chen, L., Fossati, D., Ohlsson, S., and Cosejo, D. (2011). Exploring Effective Dialogue Act Sequences in One-on-one Computer Science Tutoring Dialogues. *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 65-75).

Dzikovska, M., Steinhauser, N., Moore, J. D., Campbell, G. E., Harrison, K. M., and Taylor, L. S. (2010). Content, Social, and Metacognitive Statements: An Empirical Study Comparing Human-human and Human-computer Tutorial Dialogue. *Proceedings of the European Conference on Technology Enhanced Learning* (pp. 93-108).

D'Mello, S. K., Hays, P., Williams, C., Cade, W., Brown, J., and Olney, A. (2010). Collaborative Lecturing by Human and Computer Tutors. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 178–187).

D'Mello, S. K., Lehman, B., and Graesser, A. C. (2011). A Motivationally Supportive Affect-Sensitive AutoTutor. In R. A. Calvo and S. K. D'Mello (Eds.), *New Perspectives on Affect and Learning Technologies* (pp. 113-126).

D'Mello, S. K., Olney, A., and Person, N. (2010). Mining Collaborative Patterns in Tutorial Dialogues. *Journal of Educational Data Mining*, *2*(1), 1–37.

D'Mello, S. K., Williams, C., Hays, P., and Olney, A. (2009). Individual Differences as Predictors of Learning and Engagement. *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 308-313).

Evens, M. W., and Michael, J. (2005). *One-on-one tutoring by humans and computers*.

Forbes-Riley, K., and Litman, D. (2009). Adapting to Student Uncertainty Improves Tutoring Dialogues. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 33-40).

Forbes-Riley, K., and Litman, D. (2011). When Does Disengagement Correlate with Learning in Spoken Dialog Computer Tutoring? *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 81–89).

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems* (1st ed.).

Graesser, A. C., Person, N., and Magliano, J. P. (1995). Collaborative Dialogue Patterns in Naturalistic One-to-one Tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522.

Kumar, R., Ai, H., Beuth, J., and Rosé, C. P. (2010). Socially Capable Conversational Tutors Can Be Effective in Collaborative Learning Situations. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 156–164).

Lehman, B., Cade, W., and Olney, A. (2010). Off Topic Conversation in Expert Tutoring: Waste of Time or Learning Opportunity? *Proceedings of the International Conference on Educational Data Mining* (pp. 101–110).

Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J.-L. (1993). Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-based Tutors. In S. P. Lajoie and S. J. Derry (Eds.), *Computers as Cognitive Tools* (pp. 75-105).

Litman, D., Moore, J. D., Dzikovska, M., and Farrow, E. (2009). Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains Modalities. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 149–156).

Litman, D., and Forbes-Riley, K. (2006). Correlations Between Dialogue Acts and Learning in Spoken Tutoring Dialogues. *Natural Language Engineering*, *12*(2), 161-176.

Marx, J. D., and Cummings, K. (2007). Normalized Change. *American Journal of Physics*, *75*(1), 87.

Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., and Kershaw, T. C. (2007). Beyond the Code-and-count Analysis of Tutoring Dialogues. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 349–356).

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., and Rosé, C. P. (2007). When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science*, *30*(1), 3-62.