Characterizing the Effectiveness of Tutorial Dialogue with Hidden Markov Models

Kristy Elizabeth Boyer[1,*], Robert Phillips[1,2], Amy Ingram[3], Eun Young Ha[1],
Michael Wallis[1,2], Mladen Vouk[1], and James Lester[1]

[1] Department of Computer Science, North Carolina State University
[2] Applied Research Associates, Inc.
[3] Department of Mathematics and Computer Science, Meredith College
Raleigh, North Carolina, USA
keboyer@ncsu.edu

Abstract. Identifying effective tutorial dialogue strategies is a key issue for intelligent tutoring systems research. Human-human tutoring offers a valuable model for identifying effective tutorial strategies, but extracting them is a challenge because of the richness of human dialogue. This paper addresses that challenge through a machine learning approach that 1) learns tutorial strategies from a corpus of human tutoring, and 2) identifies the statistical relationships between student outcomes and the learned strategies. We have applied hidden Markov modeling to a corpus of annotated task-oriented tutorial dialogue to learn one model for each of two effective human tutors. We have identified significant correlations between the automatically extracted tutoring modes and student learning outcomes. This work has direct applications in authoring data-driven tutorial dialogue system behavior and in investigating the effectiveness of human tutoring.

Keywords: Tutorial dialogue, natural language, tutoring strategies.

1 Introduction

A key issue in intelligent tutoring systems research is identifying effective tutoring strategies to support student learning. It has been long recognized that human tutoring offers a valuable model of effective tutorial strategies, and a rich history of tutorial dialogue research has identified some components of these strategies [1-4]. An important research direction is to use dialogue corpora to create models that can assess strategies' differential effectiveness [5, 6]. There is growing evidence that tutorial dialogue structure can be automatically extracted from corpora of human tutoring, and that the resulting models can illuminate relationships between tutorial dialogue structure and student outcomes such as learning and motivation [7-11]. This paper takes a step beyond the previous work by identifying relationships between student learning and automatically extracted tutoring strategies, or *modes*. This modeling framework for extracting tutoring strategies and analyzing their differential effectiveness has

[*] Corresponding author.

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part I, LNCS 6094, pp. 55–64, 2010.
© Springer-Verlag Berlin Heidelberg 2010

direct applications in authoring data-driven tutorial dialogue system behavior and in research regarding the effectiveness of human tutors.

2 Related Work

Identifying effective tutoring strategies has long been a research focus of the intelligent tutoring systems community. Empirical studies of human and computer tutoring have revealed characteristics of novice and expert tutors [12, 13], Socratic and didactic strategies [14], collaborative dialogue patterns in tutoring [15], and interrelationships between affect, motivation, and learning [1, 16]. As a rich form of communication, tutorial dialogue is not fully understood: recent work suggests that the interactivity facilitated by human tutoring is key to its effectiveness [6], and other research indicates that students can learn effectively by watching playbacks of past tutoring sessions [17]. Such findings contribute to our understanding of tutoring phenomena, but also raise questions about the relative effectiveness of different tutoring approaches.

To shed further light on this issue, an important line of research involves modeling the specific relationships between different types of tutoring interactions and learning [5]. Some studies have investigated how shallow measures, such as average student turn length, correlate with learning in typed dialogue [18-20]. Analysis at the dialogue act and bigram levels has uncovered significant relationships with learning in spoken dialogue [7]. Recently, we have seen a growing emphasis on applying automatic techniques to investigate learning correlations across domains and modalities [21] and for devising optimal local strategies [9, 22]. Our work contributes to this line of investigation by applying hidden Markov models (HMMs) in a novel way to characterize the effectiveness of tutorial dialogue. HMMs have been applied successfully to such tasks as modeling student activity patterns [23, 24], characterizing the success of collaborative peer dialogues [25], and learning human-interpretable models of tutoring modes [8]. For tutorial dialogue, the doubly stochastic structure of HMMs (Section 5.1) is well suited to capturing local dependencies and to extracting structures whose components are distributed across entire tutoring sessions.

3 Tutoring Study

The corpus that serves as the basis for this work was collected during a human-human tutoring study. The goal of this study was to produce a sizeable corpus of effective tutoring from which data-driven models of task-oriented tutorial dialogue could be learned. In keeping with this goal, the study features two paid tutors who had achieved the highest average student learning gains in two prior studies [10, 26]. Tutor A was a male computer science student in his final semester of undergraduate studies. Tutor B was a female third-year computer science graduate student. An initial analysis of the corpus suggested that the tutors took different approaches; for example, Tutor A was less proactive than Tutor B [27]. As we describe below, the two tutors achieved similar learning gains.

Students were drawn from four separate sections, or modules, of the same university computer science course titled "Introduction to Programming – Java". They participated on a voluntary basis in exchange for a small amount of course credit. A total of 61 students completed tutoring sessions, constituting a participation rate of 64%. Ten of these sessions were omitted due to inconsistencies (e.g., network problems, students performing task actions outside the workspace sharing software). The first three sessions were also omitted because they featured a pilot version of the task that was modified for subsequent sessions. The remaining 48 sessions were utilized in the modeling and analysis presented here.

In order to ensure that all interactions between tutor and student were captured, participants reported to separate rooms at a scheduled time. Students were shown an instructional video that featured an orientation to the software and a brief introduction to the learning task. This video was also shown to the tutors at the start of the study. After each student completed the instructional video, the tutoring session commenced. The students and tutors interacted using software with a textual dialogue interface and a shared task workspace that provided tutors with read-only access. Students completed a learning task comprised of a programming exercise that involved applying concepts from recent class lectures including for loops, arrays, and parameter passing. The tutoring sessions ended when the student had completed the three-part programming task or one hour had elapsed.

Students completed an identical paper-based pretest and posttest designed to gauge learning over the course of the tutoring session. These free-response instruments were written by the research team and revised according to feedback from an independent panel of three computer science educators, with between three and twenty years of classroom experience. This panel assessed the difficulty of each question and the degree to which it addressed the targeted learning concepts.

According to a paired sample *t*-test, the tutoring sessions resulted in a statistically significant average learning gain as measured by posttest minus pretest (*mean*=7%; p<0.0001). There was no significant difference between the mean learning gains by tutor ($mean_A$=6.9%, $mean_B$=8.6%; p=0.569). Analysis of the pretest scores indicates that the two groups of students were equally prepared for the task: Tutor A's students averaged 79.5% on the pretest, and Tutor B's students averaged 78.9% (*t*-test *p=0.764).*

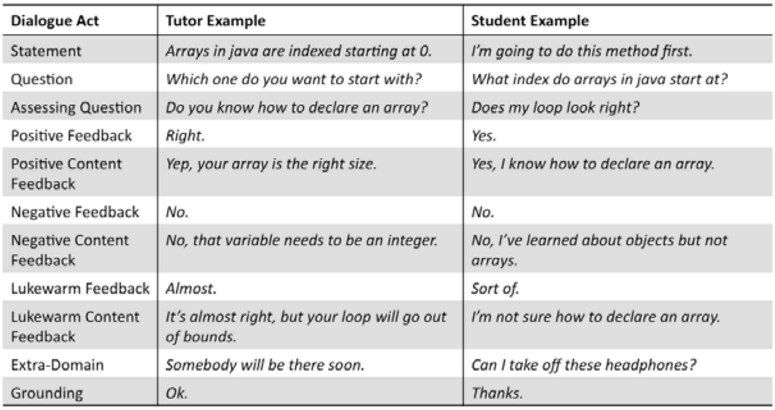
4 Corpus Annotation

The raw corpus contains 102,315 events. 4,806 of these events are dialogue messages. The 1,468 student utterances and 3,338 tutor utterances were all subsequently annotated with dialogue act tags (Section 4.1). The remaining events in the raw corpus consist of student problem-solving traces that include typing, opening and closing files, and executing the student's program. The entries in this problem-solving data stream were manually aggregated into significant student work events (Section 4.2), resulting in 3,793 tagged task actions.

4.1 Dialogue Act Annotation

One human tagger applied the dialogue act annotation scheme (Table 1) to the entire corpus. A second tagger annotated a randomly selected subset containing 10% of the utterances. The resulting Kappa was 0.80, indicating *substantial* agreement.[1]

Table 1. Dialogue act annotation scheme



Dialogue Act	Tutor Example	Student Example
Statement	*Arrays in java are indexed starting at 0.*	*I'm going to do this method first.*
Question	*Which one do you want to start with?*	*What index do arrays in java start at?*
Assessing Question	*Do you know how to declare an array?*	*Does my loop look right?*
Positive Feedback	*Right.*	*Yes.*
Positive Content Feedback	*Yep, your array is the right size.*	*Yes, I know how to declare an array.*
Negative Feedback	*No.*	*No.*
Negative Content Feedback	*No, that variable needs to be an integer.*	*No, I've learned about objects but not arrays.*
Lukewarm Feedback	*Almost.*	*Sort of.*
Lukewarm Content Feedback	*It's almost right, but your loop will go out of bounds.*	*I'm not sure how to declare an array.*
Extra-Domain	*Somebody will be there soon.*	*Can I take off these headphones?*
Grounding	*Ok.*	*Thanks.*

4.2 Task Annotation

Student task actions were recorded at a low level (i.e., individual keystrokes). A human judge aggregated these events into problem-solving chunks that occurred between each pair of dialogue utterances and annotated the student work for subtasks and correctness. The task annotation protocol was hierarchically structured and, at its leaves, included more than fifty low-level subtasks. After tagging the subtask, the judge tagged the chunk for correctness. The correctness categories were *Correct* (fully conforming to the requirements of the learning task), *Buggy* (violating the requirements of the learning task), *Incomplete* (on track but not yet complete), and *Dispreferred* (functional but not conforming to the pedagogical goals of the task).

One human judge applied this protocol to the entire corpus, with a second judge tagging 20% of the data that had been selected via random sampling stratified by tutor in order to establish reliability of the tagging scheme. Because each judge independently played back the events and aggregated them into problem-solving chunks, the two taggers often identified a different number of events in a given window. Any unmatched subtask tags were treated as disagreements. The simple Kappa statistic for subtask tagging was 0.58, indicating *moderate* agreement. However, because there is a sense of ordering within the subtask tags (i.e., the 'distance' between subtasks *1a* and *1b* is smaller than the 'distance' between subtasks *1a* and *3b*), it is also meaningful to consider the weighted Kappa statistic, which was 0.86, indicating *almost perfect* agreement. To calculate agreement on the task correctness tag, we considered all task actions for which the two judges agreed on the subtask tag. The resulting Kappa

[1] Throughout this paper we employ a set of widely used agreement categories for interpreting Kappa values: *fair*, *moderate*, *substantial*, and *almost perfect* [29].

statistic was 0.80, indicating *substantial* agreement. At the current stage of work, only the task correctness tags have been included as input to the HMMs; incorporating subtask labels is left to future work.

5 Hidden Markov Models

The annotated corpus consists of sequences of dialogue and problem-solving actions, with one sequence for each tutoring session. Our modeling goal was to extract tutoring modes from these sequences in an unsupervised fashion (i.e., without labeling the modes manually), and to identify relationships between these modes and student learning. Findings from an earlier analysis [27] suggested that the two tutors employed different strategies than each other; therefore, we disaggregated the data by tutor and learned two models. In prior work we found that identifying dependent pairs of dialogue acts and joining them into a single bigram observation during preprocessing resulted in models that were more interpretable [28]. In the current work we found that this preprocessing step produced a better model fit in terms of HMM log likelihood; the resulting hybrid sequences of unigrams and bigrams were used for training the models reported here.

5.1 Modeling Framework

In our application of HMMs to tutorial dialogue, we treat the hidden states as tutorial strategies, or modes, whose structure is learned during model training.[2] These states are characterized by *emission probability distributions*, which map each hidden state onto the observable symbols. The *transition probability distribution* determines transitions between hidden states, and the *initial probability distribution* determines the starting state [30]. Model training is an iterative process that terminates when the model parameters have converged or when a pre-specified number of iterations have been completed. Our training algorithm varied the number of hidden states from two to twenty and selected the model size that achieved the best average log-likelihood fit across ten stratified subsets of the data.

5.2 Best-Fit HMMs

The best-fit HMM for Tutor A's dialogues features eight hidden states. Figure 1 depicts a subset of the transition probability diagram with nodes representing hidden states (tutoring modes). Inside each node is a histogram of its emission probability distribution. For simplicity, only five of the eight states are displayed in this diagram; each state that was omitted mapped to less than 5% of the observed data sequences and was not significant in the correlational analysis. We have interpreted and named each tutoring mode based on its structure. For example, State 4 is dominated by correct task actions; therefore, we name this state *Correct Student Work.* State 6 is comprised of student acknowledgements, pairs of tutor statements, some correct task

[2] The notion that tutorial dialogue strategies, or modes, constitute a portion of the underlying structure of tutorial dialogue is widely accepted. However, describing these hidden states as *tutoring modes* is an interpretive choice because the HMMs were learned in an unsupervised fashion.

actions, and assessing questions by both tutor and student; we label this state *Student Acting on Tutor Help.* The best-fit model for Tutor B's dialogues features ten hidden states. A portion of this model, consisting of all states that mapped to more than 5% of observations, is displayed in Figure 2.

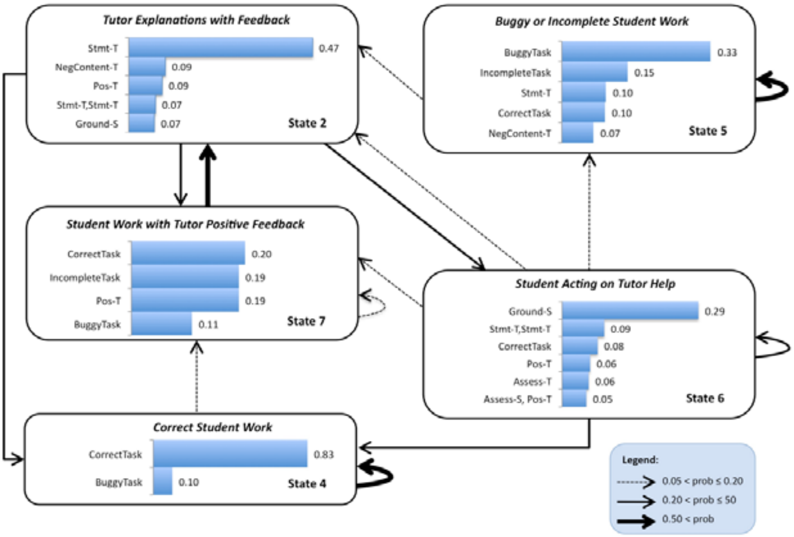


Fig. 1. Subset of HMM transition diagram for Tutor A. Histograms represent emission probability distributions. (Emission and transition probabilities < 0.05 are not displayed.).

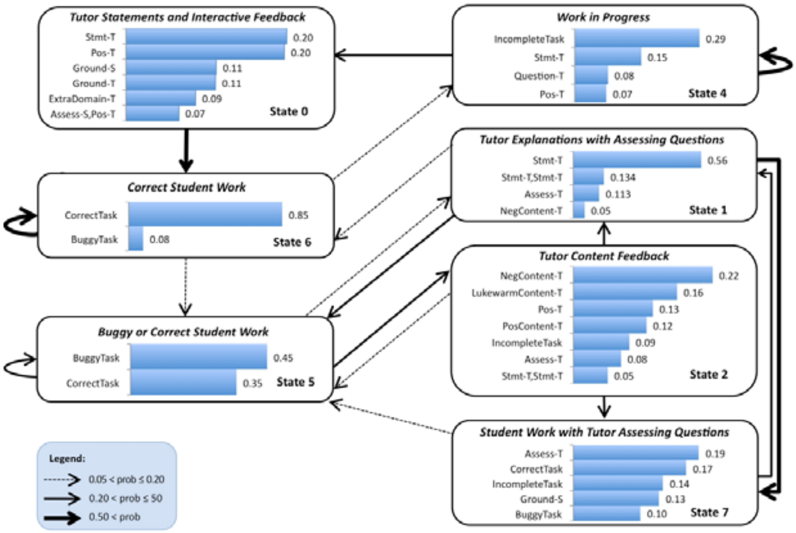


Fig. 2. Subset of HMM transition diagram for Tutor B. Histograms represent emission probability distributions. (Emission and transition probabilities < 0.05 are not displayed).

5.3 Model Interpretation

Some tutoring modes with similar structures were identified by both models. Both models feature a *Correct Student Work* mode characterized by the student's successful completion of a subtask. This state maps to 38% of observations with Tutor A and 29% of observations with Tutor B. In both cases the *Correct Student Work* mode occurs more frequently than any other mode. Each of the next three most frequently occurring modes maps onto 10-15% of the observations. For Tutor A, one such mode is *Tutor Explanations with Feedback*, while for Tutor B a corresponding mode is *Tutor Explanations with Assessing Questions*. In both cases, the mode involves tutors explaining concepts or task elements. A key difference is that with Tutor A, the explanation mode includes frequent negative content feedback or positive content-free feedback, while for Tutor B the explanation mode features questions in which the tutor aims to gauge the student's knowledge. A similar pattern emerges with each tutor's next most frequent mode: for Tutor A, this mode is *Student Work with Tutor Positive Feedback*; for Tutor B, the mode is *Student Work with Tutor Assessing Questions*. These corresponding modes illuminate a tendency for Tutor A to provide feedback in situations where Tutor B chooses to ask the student a question. For Tutor A, the only mode that featured assessing questions was *Student Acting on Tutor Help*, which as we will discuss, was positively correlated with student learning.

5.4 Correlations with Student Outcomes

With the learned models in hand, the next goal was to identify statistical relationships between student learning and the automatically extracted tutoring modes. The models presented above were used to map each sequence of observed dialogue acts and task actions onto the set of hidden states (i.e., tutoring modes) in a maximum likelihood fashion. The transformed sequences were used to calculate the frequency distribution of the modes that occurred in each tutoring session (e.g., State 0 = 32%, State 1 = 15%...State 8 = 3%). For each HMM, correlations were generated between the learning gain of each student session and the relative frequency vector of tutoring modes for that session to determine whether significant relationships existed between student learning and the proportion of discrete events (dialogue and problem solving) that were accounted for by each tutoring mode. For Tutor A, the *Student Acting on Tutor Help* mode was positively correlated with learning (*r=0.51; p<0.0001*). For Tutor B, the *Tutor Content Feedback* mode was positively correlated with learning (*r=0.55; p=0.01*) and the *Work in Progress* mode was negatively correlated with learning (*r=-0.57; p=0.0077).*

6 Discussion

We have identified significant correlations between student learning gains and the automatically extracted tutoring modes modeled in the HMMs as hidden states. While students who worked with either tutor achieved significant learning on average, each group of students displayed a substantial range of learning gains. The correlational analysis leveraged this data spread to gain insight into which aspects of the tutorial interaction were related to higher or lower learning gains.

For Tutor A, the relative frequency of the *Student Acting on Tutor Help mode* was positively correlated with student learning. This mode was characterized primarily by student acknowledgments and also featured tutor explanations, correct student work, positive tutor feedback, and assessing questions from both tutor and student. The composition of this tutoring mode suggests that these observed events possess a synergy that, in context, contributed to student learning. In a learning scenario with novices, it is plausible that only a small subset of tutor explanations were grasped by the students and put to use in the learning task. The *Student Acting on Tutor Help* mode may correspond to those instances, in contrast to the *Correct Student Work* mode in which students may have been applying prior knowledge.

For Tutor B, the *Tutor Content Feedback* mode was positively correlated with student learning. This mode was relatively infrequent, mapping to only 7% of tutoring events. However, as noted in Section 5.3, providing direct feedback represents a departure from this tutor's more frequent approach of asking assessing questions of the student. Given the nature of the learning task and the corresponding structure of the learning instrument, students may have identified errors in their work and grasped new knowledge most readily through this tutor's direct feedback.

For Tutor B, the *Work in Progress* mode was negatively correlated with learning. This finding is consistent with observations that in this tutoring study, students did not easily seem to operationalize new knowledge that came through tutor hints, but rather, often needed explicit constructive feedback. The *Work in Progress* mode features no direct tutor content feedback. Tutor questions and explanations (which are at a more abstract level than the student's solution) in the face of incomplete student work may not have been an effective tutoring approach in this study.

7 Conclusion and Future Work

We have collected a corpus of human-human tutorial dialogue, manually annotated it with dialogue acts and task actions, and utilized HMMs to extract the tutoring modes present in the corpus in an unsupervised fashion. We have examined two by-tutor HMMs and identified correlations between these models and student learning. This work extends findings that have correlated learning with highly localized structures such as unigrams and bigrams of dialogue acts [7, 10]. Using HMMs, we have correlated student learning with automatically extracted tutoring modes whose structure was learned from tutoring sessions. This work takes a step toward fully automatic extraction of tutorial strategies from corpora, a contribution that has direct application in human tutoring research. The approach also has application in tutorial dialogue system development, for example, by producing a data-driven library of system strategies.

A promising direction for future work involves learning models that more fully capture the tutorial phenomena that influence learning. There seems to be significant room for improvement in this regard, as evidenced by the fact that relatively few of the automatically extracted tutorial dialogue modes were correlated with learning. Continuing work on rich dialogue act and task annotation and deep linguistic analysis of dialogue utterances are important directions. Additionally, future work should leverage details of the task structure to a greater extent by considering regularities within tasks and subtasks as part of an augmented model structure in order to more fully capture details of the tutorial interaction.

Acknowledgments. The authors wish to thank Tiffany Barnes for insightful discussions. Thanks to Chris Eason, Lauren Hayward, and Dan Longo for expert review of the learning instruments. This work is supported in part by the NCSU Department of Computer Science along with the National Science Foundation through a Graduate Research Fellowship and Grants CNS-0540523, REC-0632450 and IIS-0812291. 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

1. Lepper, M.R., Woolverton, M., Mumme, D.L., et al.: Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors. In: Lajoie, S.P., Derry, S.J. (eds.) Computers as Cognitive Tools. Lawrence Erlbaum Associates, Hillsdale (1993)
2. Fox, B.A.: The Human Tutorial Dialogue Project. Lawrence Erlbaum Associates, Hillsdale (1993)
3. Graesser, A.C., Person, N., Magliano, J.: Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. Jl. of Applied Cog. Psych. 9, 269–306 (2004)
4. Chi, M.T.H., Siler, S.A., Jeong, H., et al.: Learning from Human Tutoring. Cog. Sci. 25, 471–533 (2001)
5. Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., Kershaw, T.C.: Beyond the Code-and-Count Analysis of Tutoring Dialogues. In: 13th International Conference on AI in Education, pp. 349–356 (2007)
6. Chi, M., Jordan, P., VanLehn, K., Litman, D.: To Elicit Or to Tell: Does it Matter? In: 14th International Conference on AI in Education, pp. 197–204 (2009)
7. Litman, D., Forbes-Riley, K.: Correlations between Dialogue Acts and Learning in Spoken Tutoring Dialogues. Nat. Lang. Eng. 12, 161–176 (2006)
8. Boyer, K.E., Ha, E.Y., Wallis, M.D., Phillips, R., Vouk, M.A., Lester, J.C.: Discovering Tutorial Dialogue Strategies with Hidden Markov Models. In: 14th International Conference on AI in Education, pp. 141–148 (2009)
9. Chi, M., Jordan, P., VanLehn, K., Hall, M.: Reinforcement Learning-Based Feature Selection for Developing Pedagogically Effective Tutorial Dialogue Tactics. In: 1st International Conference on Educational Data Mining, pp. 258–265 (2008)
10. Boyer, K.E., Phillips, R., Wallis, M.D., Vouk, M.A., Lester, J.C.: Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 239–249. Springer, Heidelberg (2008)
11. Forbes-Riley, K., Litman, D.: Adapting to Student Uncertainty Improves Tutoring Dialogues. In: 14th International Conference on AI and Education, pp. 33–40 (2009)
12. Evens, M., Michael, J.: One-on-One Tutoring by Humans and Computers. Lawrence Erlbaum Associates, Mahwah (2006)
13. Cade, W., Copeland, J., Person, N., D'Mello, S.: Dialog Modes in Expert Tutoring. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 470–479. Springer, Heidelberg (2008)
14. Rosé, C.P., Moore, J.D., VanLehn, K., et al.: A Comparative Evaluation of Socratic Versus Didactic Tutoring. #LRDC-BEE-1 (2000)

15. Graesser, A.C., Person, N.K., Magliano, J.P.: Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. Applied Cog. Psych. 9, 495–522 (1995)
16. D'Mello, S., Taylor, R.S., Graesser, A.: Monitoring Affective Trajectories during Complex Learning. In: 29th Annual Cognitive Science Society, pp. 203–208 (2007)
17. Chi, M.T.H., Roy, M., Hausmann, R.G.M.: Observing Tutorial Dialogues Collaboratively: Insights about Human Tutoring Effectiveness from Vicarious Learning. Cog. Sci. 32, 301–341 (2008)
18. Core, M.G., Moore, J.D., Zinn, C.: The Role of Initiative in Tutorial Dialogue. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 67–74 (2003)
19. Katz, S., Allbritton, D., Connelly, J.: Going Beyond the Problem Given: How Human Tutors use Post-Solution Discussions to Support Transfer. International Journal of Artificial Intelligence in Education 13, 79–116 (2003)
20. Rosé, C., Bhembe, D., Siler, S., Srivastava, R., VanLehn, K.: The Role of Why Questions in Effective Human Tutoring. In: International Conference on AI in Education, pp. 55–62 (2003)
21. Litman, D., Moore, J., Dzikovska, M., Farrow, E.: Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities. In: 14th International Conference on AI in Education, pp. 149–156 (2009)
22. Tetreault, J.R., Litman, D.J.: A Reinforcement Learning Approach to Evaluating State Representations in Spoken Dialogue Systems. Speech Comm. 50, 683–696 (2008)
23. Beal, C., Mitra, S., Cohen, P.R.: Modeling Learning Patterns of Students with a Tutoring System using Hidden Markov Models. In: 13th International Conference on AI in Education, pp. 238–245 (2007)
24. Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., Schwartz, D.: Using Hidden Markov Models to Characterize Student Behaviors in Learning-by-Teaching Environments. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 614–625. Springer, Heidelberg (2008)
25. Soller, A., Stevens, R.: Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment. In: Hancock, G.R., Samuelsen, K.M. (eds.) Advances in Latent Variable Mixture Models, pp. 217–253. Information Age Publishing (2007)
26. Boyer, K.E., Vouk, M.A., Lester, J.C.: The Influence of Learner Characteristics on Task-Oriented Tutorial Dialogue. In: Proceedings of the 13th International Conference on AI in Education, pp. 365–372 (2007)
27. Boyer, K.E., Phillips, R., Wallis, M.D., Vouk, M.A., Lester, J.C.: The Impact of Instructor Initiative on Student Learning through Assisted Problem Solving. In: 40th Tech. Symposium on Computer Science Education, pp. 14–18 (2009)
28. Boyer, K.E., Phillips, R., Ha, E.Y., Wallis, M.D., Vouk, M.A., Lester, J.C.: Modeling Dialogue Structure with Adjacency Pair Analysis and Hidden Markov Models. In: NAACL HLT, Short Papers, pp. 49–52 (2009)
29. Landis, J.R., Koch, G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159–174 (1977)
30. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77, 257–286 (1989)