Investigating the Relationship Between Dialogue Structure and Tutoring Effectiveness: A Hidden Markov Modeling Approach

Kristy Elizabeth Boyer*[a], Robert Phillips[ab], Amy Ingram[c],
Eun Young Ha[a], Michael Wallis[ab], Mladen Vouk[a] and James Lester[a]

[a] Department of Computer Science
North Carolina State University, Raleigh, North Carolina, USA

[b] Applied Research Associates, Inc., Raleigh, North Carolina, USA

[c] Department of Computer Science
University of North Carolina at Charlotte, Charlotte, North Carolina, USA

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 article addresses that challenge through a machine learning approach that 1) learns tutorial modes from a corpus of human tutoring, and 2) identifies the statistical relationships between student outcomes and the learned modes. The modeling approach utilizes hidden Markov models (HMMs) to capture the unobservable stochastic structure that is thought to influence the observations, in this case dialogue acts and task actions, that are generated by task-oriented tutorial dialogue. We refer to this unobservable layer as the hidden dialogue state, and interpret it as representing the tutor and students' collaborative intentions. We have applied HMMs to a corpus of annotated task-oriented tutorial dialogue to learn one model for each of two effective human tutors. Significant correlations emerged between the automatically extracted tutoring modes and student learning outcomes. Broadly, the results suggest that HMMs can learn meaningful hidden tutorial dialogue structure. More specifically, the findings point to specific mechanisms within task-oriented tutorial dialogue that are associated with increased student learning. This work has direct applications in authoring data-driven tutorial dialogue system behavior and in investigating the effectiveness of human tutoring.

---

* Corresponding author: keboyer@ncsu.edu

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. A rich history of tutorial dialogue research has identified some components of these strategies including adaptive cognitive scaffolding, motivational support, and collaborative dialogue patterns that support learning through tutoring (M. T. H. Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Fox, 1993; Graesser, Person, & Magliano, 2004; Lepper, Woolverton, Mumme, & Gurtner, 1993). As the field has grown in its understanding of fundamentally effective tutoring phenomena, it has also become clear that the complexities of choosing the most effective contextualized tutoring strategies are not well understood. An important research direction is to use dialogue corpora, collected from human-human or human-computer tutorial dialogue, to create models that can assess the differential effectiveness of tutoring strategies at a fine-grained level (M. Chi, Jordan, VanLehn, & Litman, 2009; Ohlsson et al., 2007). This research direction is particularly timely given the increasing attention that is being given to machine learning techniques to address tasks such as automatically authoring intelligent tutoring system behavior (Barnes & Stamper, 2010) along with creating data-driven natural language dialogue management systems for tutoring (e.g., Tetreault & Litman, 2008) and other dialogue application areas (Bangalore, Di Fabbrizio, & Stent, 2008). Because machine-learned models constitute stochastic representations of tutoring expertise, they can be used not only in a generative context to author tutoring system behavior, but also as descriptive models whose associations with student outcomes give insight into the effectiveness of tutoring.

There is growing evidence that meaningful tutorial dialogue patterns can be automatically extracted from corpora of human tutoring using machine learning techniques (e.g., M. Chi, VanLehn, & Litman, 2011; Forbes-Riley & Litman, 2009; Fossati, Di Eugenio, Ohlsson, Brown, & Chen, 2010; Kersey, Di Eugenio, Jordan, & Katz, 2009; Ohlsson et al., 2007; Tetreault & Litman, 2008). The meaningfulness of these models can be assessed in several different ways, including whether their components are *correlated* with student outcomes in an existing data set, or whether implementing them can improve the effectiveness of a tutoring system. Following the first approach, the work reported in this article explores whether models of *hidden dialogue state*, learned in an unsupervised fashion, are correlated with student learning.

An underlying premise of this work is that natural language dialogue is influenced by a layer of unobservable stochastic structure. This structure is likely comprised of cognitive, affective, and social influences. Hidden Markov models (HMMs) provide a framework for explicitly modeling unobservable structure within a layer of hidden states (Rabiner, 1989). In the current work, HMMs are learned from a corpus of annotated task-oriented tutorial dialogue, and then the layer of hidden states is interpreted as hidden dialogue state, corresponding to tutoring *modes* (Cade, Copeland, Person, & D'Mello, 2008). This interpretation is supported by previous work in which automatically extracted hidden dialogue state was found to resemble tutoring modes from the literature (Boyer, Ha et al., 2009). The current work takes a step beyond the previous work by identifying relationships between student learning and hidden dialogue state. This modeling framework for extracting tutoring modes and analyzing their differential effectiveness has direct applications in authoring data-driven tutorial dialogue system behavior and in research regarding the effectiveness of human tutors.

This article is organized as follows. First, we present related work regarding modeling the differential effectiveness of tutoring approaches. Second, we describe the human-human task-oriented tutoring study that was conducted, and the structure of the resulting corpus from which the HMMs were learned. Third, we describe the corpus annotation, both for dialogue acts and for problem-solving events. Fourth, we present an introduction to HMMs and describe the methodology by which they were applied to this corpus. Next, the best-fit models, and their correlations with student learning, are presented and discussed. Finally, we present conclusions and discuss directions for future work.

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, such as experts' tendency to ask more questions than novice tutors (Evens & Michael, 2006). It has also been found that even in expert tutoring, "sophisticated" strategies may not be employed, yet expert tutoring is highly effective (Cade et al., 2008). Gaining insight into the sources of tutoring effectiveness is an important research direction. Phenomena such as collaborative dialogue patterns in tutoring (Graesser, Person, & Magliano, 1995), Socratic and didactic strategies (Rosé, Moore, VanLehn, & Allbritton, 2000), and interrelationships between affect, motivation, and learning (D'Mello, Taylor, & Graesser, 2007; Lepper et al., 1993) have been investigated. However, 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 (M. Chi et al., 2009), and other research indicates that students can learn effectively by watching playbacks of past tutoring sessions (M. T. H. Chi, Roy, & Hausmann, 2008). 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 (Ohlsson et al., 2007). Some studies have investigated how shallow measures, such as average student turn length, correlate with learning in typed dialogue (Core, Moore, & Zinn, 2003; Katz, Allbritton, & Connelly, 2003; Rosé, Bhembe, Siler, Srivastava, & VanLehn, 2003). Analysis at the dialogue act and bigram levels has uncovered significant relationships with learning in spoken dialogue (Litman & Forbes-Riley, 2006). Recently, we have seen a growing emphasis on applying automatic techniques to investigate learning correlations across domains and modalities (Litman, Moore, Dzikovska, & Farrow, 2009), for devising optimal local strategies (M. Chi, Jordan, VanLehn, & Hall, 2008; Tetreault & Litman, 2008), and for assessing the impact of micro-level tutorial tactics (M. Chi et al., 2010).

These lines of research do not simply acquire machine-learned models of tutoring; they assess whether the extracted structure is meaningful. The notion of whether a model is *meaningful* can be explored in many different ways. One approach is to use the model to refine the behavior of a tutoring system and determine whether its effectiveness is improved. For example, bigram analysis of dialogue turns in qualitative physics tutoring has revealed that human tutors adapt to student uncertainty (Litman & Forbes-Riley, 2006). This analysis demonstrated the statistical dependence of tutor responses on student uncertainty, and as such created a model of one aspect of tutor adaptation. However, the importance of that finding was reinforced by the fact that, when a tutorial dialogue system was engineered to adapt in a similar way, it became more effective (Forbes-Riley & Litman, 2009).

Another way to determine whether a model is capturing a meaningful aspect of tutoring is to explore whether, when treated as a descriptive model, the model's structure is correlated with outcomes. This approach is embodied in work that builds multiple regression models on annotated corpora by treating dialogue act frequencies as predictors and student learning as the predicted value (Ohlsson et al., 2007). This multiple regression approach is also widely used in natural language dialogue systems research outside of tutoring (Walker, Litman, Kamm, & Abella, 1997) and has been applied to explore user satisfaction and learning within tutoring systems (Forbes-Riley & Litman, 2006). A regression model is just one example of a model that describes relationships within a data set. The HMMs used in this work are another example, and unlike linear regression or bigram models, HMMs capture doubly stochastic structure that explicitly accounts for hidden dialogue state. HMMs have been applied successfully to such tasks as modeling student activity patterns (Beal, Mitra, & Cohen, 2007; Jeong et al., 2008), characterizing the success of collaborative peer dialogues (Soller & Stevens, 2007), and learning human-interpretable models of tutoring modes (Boyer, Ha et al., 2009).

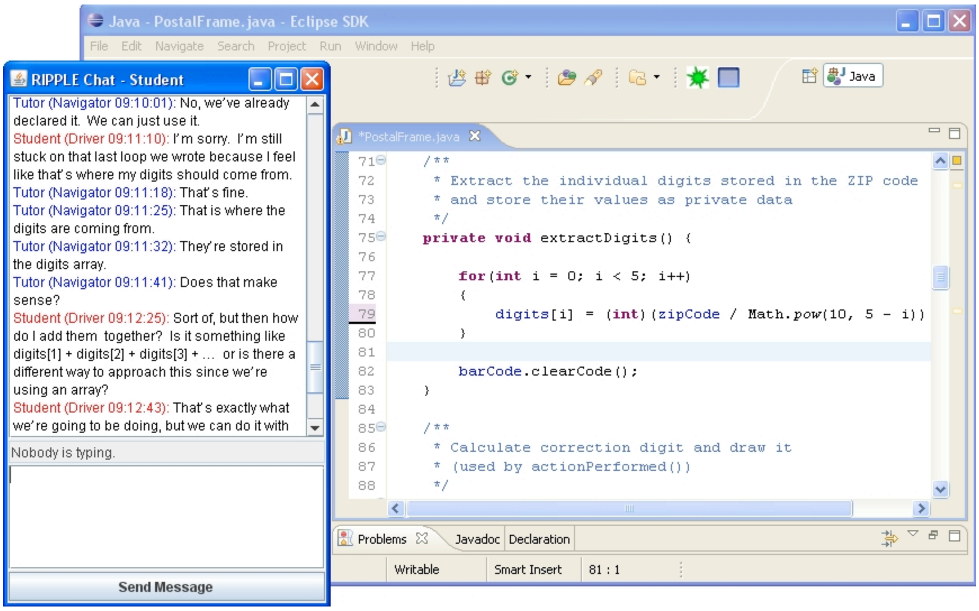
The evaluation reported here demonstrates that when hidden dialogue state is learned in an unsupervised fashion (that is, without any prior labeling of hidden dialogue state) with HMMs, the hidden structure is correlated with student learning. This finding, coupled with the finding that the constituent observed components of the hidden dialogue states were not themselves correlated with student learning, indicates that HMMs can capture a meaningful aspect of tutorial dialogue structure automatically, a finding that has theoretical and applied implications for tutorial dialogue research.

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 featured two paid tutors who had achieved the highest average student learning gains in two prior studies (Boyer, Vouk, & Lester, 2007; Boyer, Phillips, Wallis, Vouk, & Lester, 2008). 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 (Boyer, Phillips, Wallis, Vouk, & Lester, 2009). 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. The learning task that served as the focus of the tutoring sessions followed directly from students’ in-class lectures the preceding week. 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 with one another using software with a textual dialogue interface and a shared task workspace that provided tutors with read-only access (Figure 1). 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.



Tutor (Navigator 09:10:01): No, we’ve already declared it. We can just use it.
Student (Driver 09:11:10): I’m sorry. I’m still stuck on that last loop we wrote because I feel like that’s where my digits should come from.
Tutor (Navigator 09:11:18): That’s fine.
Tutor (Navigator 09:11:25): That is where the digits are coming from.
Tutor (Navigator 09:11:32): They’re stored in the digits array.
Tutor (Navigator 09:11:41): Does that make sense?
Student (Driver 09:12:25): Sort of, but then how do I add them together? Is it something like digits[1] + digits[2] + digits[3] + … or is there a different way to approach this since we’re using an array?
Student (Driver 09:12:43): That’s exactly what we’re going to be doing, but we can do it with

Nobody is typing.

```
    /**
     * Extract the individual digits stored in the ZIP code
     * and store their values as private data
     */
    private void extractDigits() {

        for(int i = 0; i < 5; i++)
        {
            digits[i] = (int)(zipCode / Math.pow(10, 5 - i))
        }

        barCode.clearCode();
    }

    /**
     * Calculate correction digit and draw it
     * (used by actionPerformed())
     */
```

Figure 1. Screenshot of human tutoring interface

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. The posttest was administered immediately after the tutoring session. According to a paired sample *t*-test, the tutoring sessions resulted in a statistically significant average learning gain as measured by posttest minus pretest scores (*mean*=7%; *variance*=0.01; $p<0.0001$). There was no statistically reliable difference between the mean learning gains by tutor ($mean_A$=6.9%, $mean_B$=8.6%; p=0.569). Analysis of the pretest scores suggests that the two groups of students were similarly prepared for the task: Tutor A's students averaged 79.5% on the pretest, and Tutor B's students averaged 78.9% (*t*-test $p=0.764$).

CORPUS ANNOTATION

The raw corpus contains 102,315 events. Of these, 4,806 are dialogue messages. The 1,468 student utterances and 3,338 tutor utterances were all subsequently annotated with dialogue act tags. 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, resulting in 3,793 tagged task actions.

Dialogue Act Annotation

One human tagger applied the dialogue act annotation scheme (Table 1) to the entire corpus.[1] Each textual dialogue message was treated as a single utterance, except where separate text spans were identified within a message that corresponded to separate dialogue acts. The primary tagger identified such spans as separate utterances, each of which corresponds to precisely one dialogue act tag. This manual segmentation was provided to the second tagger, who independently annotated a randomly selected subset containing 10% of all the utterances.[2] The resulting Kappa was 0.80, indicating *substantial* agreement.[3]

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 (Figure 2). After tagging each

---

[1] The dialogue act annotators were the second and third authors of this paper.

[2] The tutoring sessions that were used for annotation scheme development and refinement were not included in the Kappa calculation since they were annotated collaboratively by the two taggers, rather than independently.

[3] The Kappa statistic is a measure of agreement that adjusts for the agreement that would be expected by chance. Kappa values range from -1 to 1, with a Kappa of 0 being equal to chance. Throughout this article we employ a set of widely used agreement categories for interpreting Kappa values: *poor* (<0), *slight* (0-0.20), *fair* (0.21-0.40), *moderate* (0.41-0.60), *substantial* (0.61-0.80), and *almost perfect* (0.81-1.0) (Landis & Koch, 1977).

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).

Table 1. Dialogue act annotation scheme

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

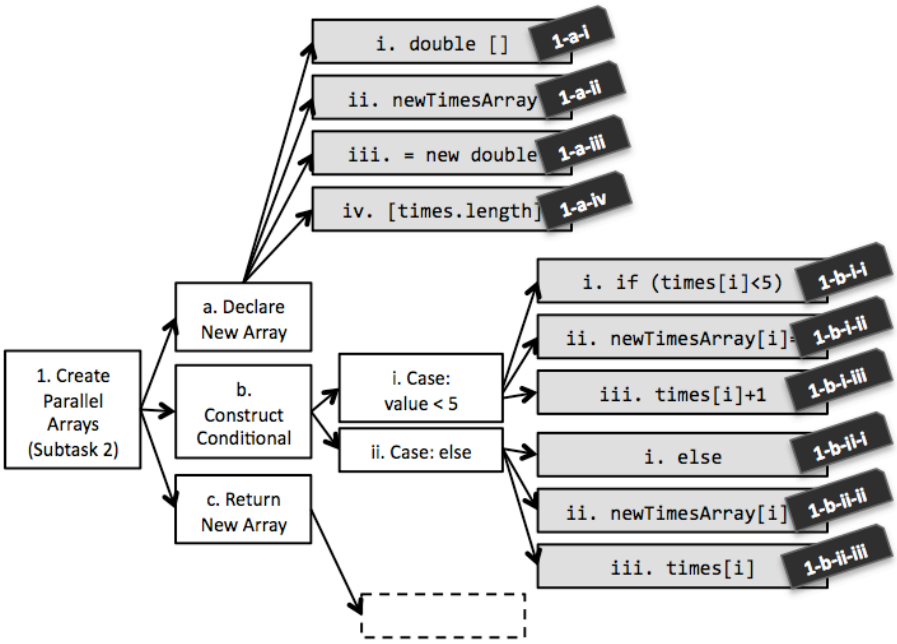


Figure 2. Portion of task action annotation scheme that was applied to manually segmented student task actions

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 balanced 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, each of which applied to a manually segmented group of student task actions, 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 (Cohen, 1968), 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 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.

Joining Adjacency Pairs

The annotation described above produced sequences of dialogue act and task action tags that capture the events of the tutoring sessions. Although these sequences could be used directly as input for learning HMMs, prior work has found that identifying dependent pairs of dialogue acts, or *adjacency pairs*, and joining them into a single bigram observation during preprocessing resulted in models that were more interpretable, in part because joining the adjacency pairs prevents the HMM from switching to a new state between the two dependent dialogue acts (Boyer, Phillips et al., 2009).

This approach first utilizes Chi-square dependency analysis to determine whether the probability of a particular dialogue act occurring is positively statistically dependent on the occurrence of a preceding dialogue act. If so, this pair of dialogue acts is considered to be an *adjacency pair*, and co-occurrences of the two acts are joined in a preprocessing step prior to building the HMMs. 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 therefore used for training the models reported here. An example of the result of the adjacency-pair joining algorithm is presented in Figure 3. The adjacency-pair joining algorithm was applied to the data prior to disaggregating by tutor.

Original dialogue act sequence:

Greeting$_S$ → Greeting$_T$ → Question$_S$ → Stmt$_T$ → AssessQ$_T$ → AssessQ$_S$ → PosFdbk$_T$

After adjacency pair joining:

Greeting$_S$Greeting$_T$ → **Question$_S$Stmt$_T$** → AssessQ$_T$ → **AssessQ$_S$PosFdbk$_T$**

Figure 3. Adjacency-pair joining

A partial list of statistically dependent dialogue acts is provided in Table 2. In this table, act_t and act_{t+1} are the dialogue acts or task actions that occur adjacent to each other in the corpus. The *p*-values indicate strength of statistical significance based on a 2×2 *Chi-square* test, with the Bonferroni *p*-value adjusted for the number of statistical tests that were performed. This adjustment accounts for the probability of finding a statistically significant result by chance. The frequency indicates how often this adjacency pair occurred in the corpus, while the expected frequency is how often the pair would be expected to occur if the two acts in the pair were independent.

HIDDEN MARKOV MODELS

The annotated corpus consists of sequences of dialogue and problem-solving actions, with one sequence for each tutoring session. Our goal was to extract a model of hidden dialogue state from these sequences in an unsupervised fashion (i.e., without labeling the hidden dialogue states manually), and to identify relationships between the tutoring modes and student learning. Findings from earlier work (Boyer, Ha et al., 2009) suggested that the two tutors employed different strategies than each other; therefore, we disaggregated the data by tutor and learned two models. This section provides an introduction to the HMM framework and then describes the modeling approach that was applied.

Table 2. Subset of statistically dependent adjacency pairs

act_t	**act_{t+1}**	***p*** **(Bonferroni)**	***p*** **(Chi-square)**	**Freq.**	**Expected freq.**
CorrectTask	CorrectTask	6.7E-257	2.1E-259	1318	675
ExtraDomain-S	ExtraDomain-T	6.2E-249	1.9E-251	44	2
Greeting-S	Greeting-T	2.7E-147	8.2E-150	81	8
AssessQuestion-T	PosFdbk-S	3.9E-138	1.2E-140	65	6
AssessQuestion-S	PosFdbk-T	1.4E-108	4.3E-111	110	18
Question-T	Stmt-S	2.6E-108	8.1E-111	37	2
AssessQuestion-T	Stmt-S	4.0E-96	1.2E-98	74	10
ExtraDomain-T	ExtraDomain-S	2.1E-82	6.6E-85	26	2
Question-S	Stmt-T	2.0E-53	6.2E-56	93	24
IncompleteTask	IncompleteTask	1.8E-51	5.5E-54	94	23
PosFdbk-S	Greeting-T	2.3E-40	7.1E-43	26	3
NegFdbk-S	Greeting-T	4.7E-34	1.4E-36	17	2
BuggyTask	BuggyTask	7.9E-32	2.4E-34	181	81
ExtraDomain-T	ExtraDomain-T	7.2E-29	2.2E-31	18	2
DispreferredTask	DispreferredTask	8.7E-27	2.7E-29	9	1
AssessQuestion-S	LkwmContent-T	2.5E-26	7.6E-29	24	4
AssessQuestion-T	PosContent-S	1.8E-23	5.4E-26	14	1
AssessQuestion-T	LkwmFdbk-S	4.9E-21	1.5E-23	12	1
AssessQuestion-S	NegContent-T	9.6E-21	3.0E-23	43	11
Stmt-T	Greeting-S	6.5E-20	2.0E-22	128	60
AssessQuestion-S	PosContent-T	1.4E-18	4.2E-21	20	3
BuggyTask	NegContent-T	1.6E-16	5.1E-19	77	31
Stmt-T	Stmt-T	4.0E-16	1.2E-18	324	215
AssessQuestion-S	LkwmFdbk-T	5.9E-16	1.8E-18	18	3
LkwmFdbk-T	Stmt-T	3.3E-14	1.0E-16	42	14
AssessQuestion-S	NegFdbk-T	2.6E-13	8.1E-16	13	2
IncompleteTask	Stmt-T	5.2E-10	1.6E-12	124	71
AssessQuestion-T	NegFdbk-S	6.7E-10	2.1E-12	15	3

Modeling Framework

In an HMM, the *observation symbols* are constituents of the input sequences from which the model is learned. In our application, the observation symbols are annotated dialogue acts, task actions, or joined pairs of these as described above. Each observation symbol is said to be *generated* by a *hidden state* according to that hidden state's *emission probability distribution*, which maps each hidden state onto the observable symbols. In the current application, hidden states are interpreted as tutoring modes and given names pursuant to that interpretation, such as "Student Acting on Tutor Help," which is characterized by a probability distribution of dialogue acts and task actions that would be observed when the tutorial dialogue is in that hidden state. The HMM's *transition probability distribution* determines transitions between hidden states, and the *initial probability distribution* determines the starting state. More details about HMMs and inference on them can be found in a tutorial by Rabiner (1989).

Learning an HMM involves training its emission, transition, and initial probability distributions to maximize the probability of seeing the observed data given the model. Model training is an iterative process that terminates when the model parameters have converged or when a pre-specified number of iterations has been completed. The fit of a model is measured with log-likelihood, which has a monotonic relationship with likelihood and is less susceptible to numerical underflow. The iterative training process operates over a particular number N of hidden states. Our training approach uses the Baum-Welch iterative training algorithm (Rabiner, 1989) and incorporates a meta-level learning procedure that varies the number of hidden states from two to twenty and selects the model size that achieves the best average log-likelihood fit in ten-fold cross-validation.

HMMs can be graphically depicted in several different ways. Each depiction explicitly shows some components of the model while omitting others for simplicity. Perhaps the most common depiction is a time-slice view, in which each time step is displayed along with its associated hidden state and observation. Missing from such a depiction is the probability of transitioning from one hidden state to the next, because in a time slice view each transition *did* occur. In this article, the best-fit HMMs are displayed in a summary, or Bayesian, view as shown in Figure 4. This view depicts the hidden states as nodes within a graph, and the emission probability distribution of each hidden state as a histogram within the node. Arrows indicate transition probability distributions between hidden states, with heavier arrow weight indicating higher probability. Since this is not a time-slice view, omitted from this diagram are the observations that are generated at each time step.

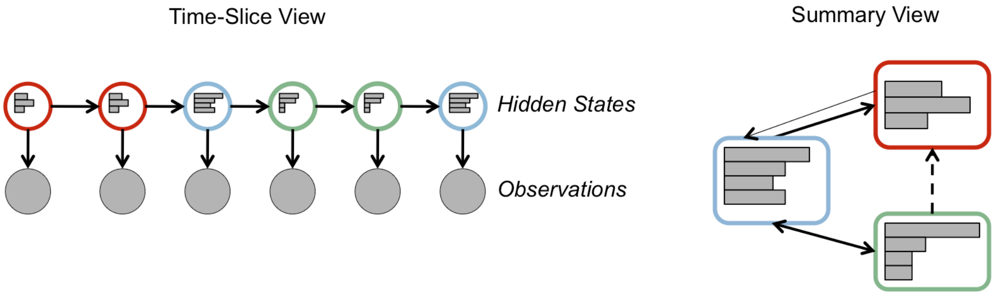


Figure 4. Time-slice and summary views of example HMM

Best-fit HMMs

The HMM that was fitted to the annotated tutoring sessions conducted with Tutor A features eight hidden states. Figure 5 depicts a subset of this HMM in a summary view with nodes representing hidden states. For simplicity, only those states that mapped to more than 5% of the observed data sequences are included. Each hidden state was interpreted as a tutoring mode and named based on its structure. For example, State 4 is dominated by correct task actions; therefore, this state is referred to as *Correct Student Work.* State 6 is comprised of student acknowledgements, pairs of tutor statements, some correct task 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 6.

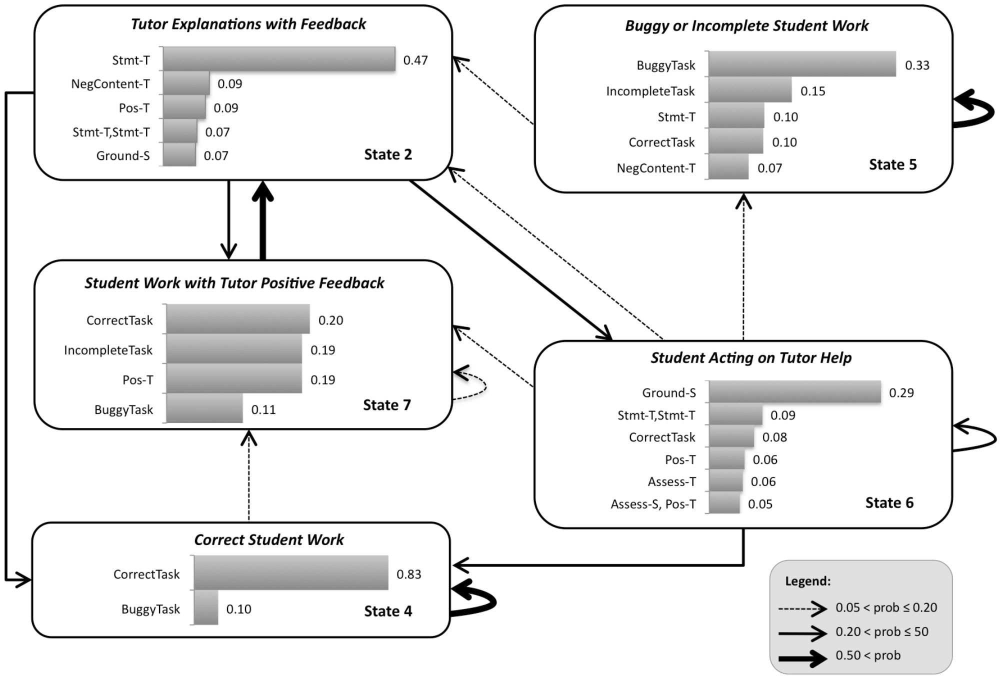


Figure 5. Portion of learned HMM for Tutor A

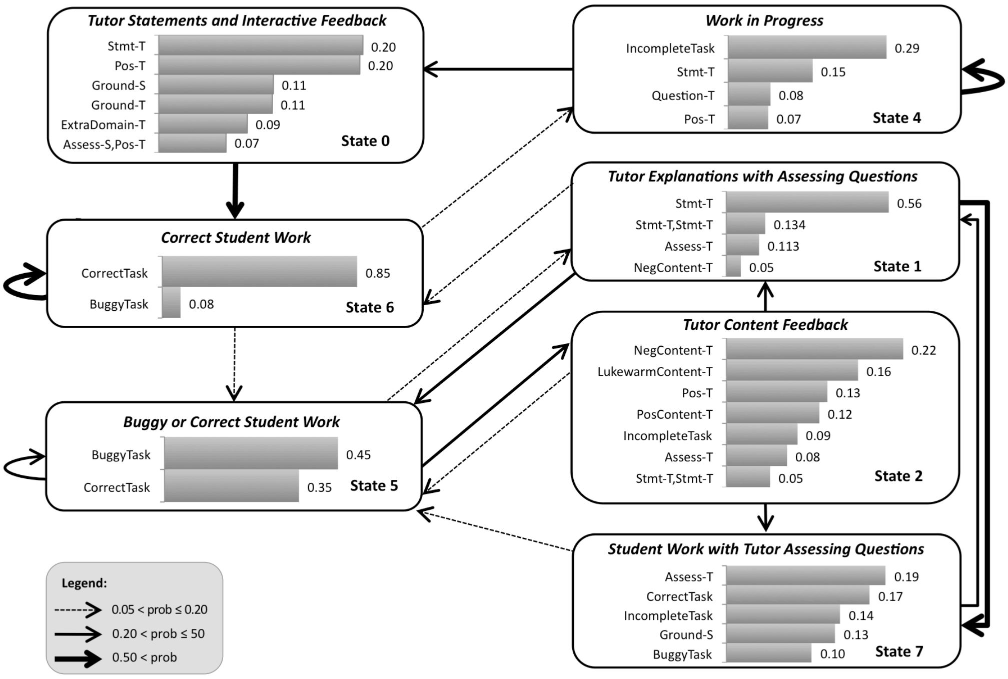


Figure 6. Portion of learned HMM for Tutor B

Model Interpretation

Some tutoring modes with similar structures were identified by both models. The frequency distribution across tutoring modes for each model is displayed in Figure 7.

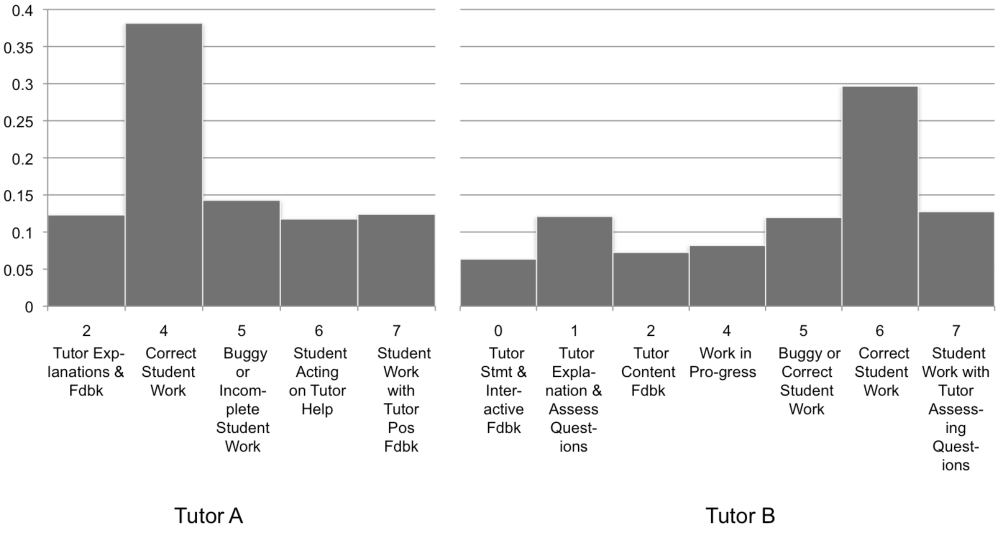


Figure 7. Relative frequency of hidden dialogue states across tutoring sessions

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. The next three most frequently occurring modes each map 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*. As we will discuss, increased frequency of this mode was positively correlated with student learning.

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 HMMs mapped 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 (Figure 8, Excerpt 1) was positively correlated with learning ($r=0.51; p<0.0001$). For Tutor B, the *Tutor Content Feedback* mode (Figure 8, Excerpt 2) was positively correlated with learning ($r=0.55; p=0.01$) and the *Work in Progress* mode was negatively correlated with learning ($r=-0.57; p=0.0077$).

DISCUSSION

Effectiveness of Tutoring Modes

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 leverages this data spread to gain insight into which aspects of the tutorial interaction were related to higher or lower learning gains.

Excerpt 1: Tutor A, State 6 (*Student Acting on Tutor Help*)

Tutor: We're going to need to create a new array to hold the new information before we start the loop [STMT]

Student: Thanks for pointing that out. [GROUND/ACK]

Student: [CORRECT TASK ACTION; SUBTASK 2 a i]

Excerpt 2: Tutor B, State 2 (*Tutor Content Feedback*)

Student: [BUGGY TASK ACTION; SUBTASK 2 c iii]

Tutor: You want to return the array [NEGCONTENTFDBK]

Tutor: Not a double [NEGCONTENTFDBK]

Student: [CORRECT TASK ACTION; SUBTASK 2 c iii]

Figure 8. Excerpts illustrating hidden dialogue states

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 previously, 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 actionable 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 seem to operationalize easily 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.

Meaningfulness of Hidden Dialogue State

The statistically significant correlations found between hidden dialogue state and student learning provide support for the notion that in this context, HMMs can learn meaningful tutorial dialogue structure in an unsupervised fashion. This finding is reinforced by the fact that the relative frequencies of the observations themselves (dialogue acts and task actions) were not significantly correlated with learning gain. In this case, the occurrence of the hidden state, an aggregate structure over observation symbols, contributed to characterizing the effectiveness of the tutorial dialogue better than the observations themselves.

Some limitations of the approach stem from the rich nature of human-human dialogue, and the uncertainty it entails. Even within HMMs' doubly stochastic framework, it is not possible to fully capture all relevant variance in the observed sequences. The fact that only a subset of the hidden states was correlated with student learning is, in part, an artifact of this limitation. Specifically, if the HMM could capture all patterns that influence learning, we might expect all hidden states' frequencies (not just a subset of them) to be correlated either positively or negatively with learning.

Another limitation stems from the extent to which the annotation scheme, both for dialogue acts and for task actions, is able to capture pedagogically relevant aspects of the interaction. There is no agreed upon dialogue act annotation scheme for tutorial dialogue, and researchers must devise tagging schemes based on the structure of their corpus, related learning and dialogue theories, and the particular aspects of the dialogue that are of interest for their work. It is possible that with any annotation of tutorial dialogue at this granularity, some pedagogically relevant phenomena are not explicitly represented within the tags. Because the hidden dialogue states are aggregate structures defined by a probability distribution over the tags, the tagging scheme has great impact on the results that can be obtained. For this reason, applying the methodology presented here on a corpus tagged in several different ways will be an important direction for future work.

CONCLUSION AND FUTURE WORK

Modeling the structure of tutorial dialogue is an important research direction that can contribute to investigating the sources of tutoring effectiveness as well as to the automatic authoring of tutorial dialogue system behavior. Tutorial dialogue is thought to have a layer of unobservable stochastic structure that includes cognitive, affective, and social dimensions. HMMs are a promising modeling framework because they explicitly model hidden structure. The work reported here provides evidence that HMMs can capture meaningful tutorial dialogue structure in an unsupervised fashion, as evidenced by the fact that the HMMs' hidden states were correlated with student learning outcomes, while the states' constituent observation symbols were not themselves correlated with learning.

The results suggest that with novice computer programming students, receiving and acting on tutor help, and receiving specific content feedback, were productive tutoring modes associated with higher learning gains. In contrast, a work-in-progress mode characterized by incomplete student work and tutor moves that contained no specific task feedback was associated with lower learning gains. The results extend findings that have correlated learning with highly localized structures such as unigrams and bigrams of dialogue acts (Boyer et al., 2008; Ohlsson et al., 2007; Litman & Forbes-Riley, 2006). 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. Promising directions include explicitly capturing the task structure within the dialogue model, such as with hierarchical HMMs, and exploring whether the difficulty of particular subtasks is related to the effectiveness of tutoring strategies, as emerging evidence suggests (Chi, VanLehn, & Litman, 2011). 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. Continued work on rich dialogue act and task annotation is also an important direction for future work, since limitations of the annotation schemes impose limitations on the learned models. Finally, deep linguistic analysis of dialogue utterances is an important area for future work, as are other directions that have the potential to eliminate manual annotation from the pipeline and produce fully unsupervised models of tutorial dialogue.

Acknowledgments. The authors wish to thank Tiffany Barnes for insightful discussions on the educational data mining aspects of this work. 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, IIS-0812291 and DRL-1007962. 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

Bangalore, S., Di Fabbrizio, G., & Stent, A. (2008). Learning the structure of task-driven human-human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(7), 1249-1259.

Barnes, T., & Stamper, J. (2010). Automatic hint generation for logic proof tutoring using historical data. *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems,* Pittsburgh, PA. 3-14.

Beal, C., Mitra, S., & Cohen, P. R. (2007). Modeling learning patterns of students with a tutoring system using hidden Markov models. *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* Marina del Rey, California. 238-245.

Boyer, K. E., Ha, E. Y., Wallis, M. D., Phillips, R., Vouk, M. A., & Lester, J. C. (2009). Discovering tutorial dialogue strategies with hidden Markov models. *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* Brighton, U.K. 141-148.

Boyer, K. E., Phillips, R., Ha, E. Y., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2009). Modeling dialogue structure with adjacency pair analysis and hidden Markov models. *The North American Association for Computational Linguistics Human Language Technologies Conference (NAACL-HLT) Short Papers,* 49-52.

Boyer, K. E., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* Montreal, Canada. 239-249.

Boyer, K. E., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2009). The impact of instructor initiative on student learning through assisted problem solving. *Proceedings of the 40th SIGCSE Technical Symposium on Computer Science Education,* Chattanooga, Tennessee. 14-18.

Boyer, K. E., Vouk, M. A., & Lester, J. C. (2007). The influence of learner characteristics on task-oriented tutorial dialogue. *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* Marina del Rey, California. 365-372.

Cade, W., Copeland, J., Person, N., & D'Mello, S. (2008). Dialog modes in expert tutoring. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* Montreal, Canada. 470-479.

Chi, M., Jordan, P., VanLehn, K., & Hall, M. (2008). Reinforcement learning-based feature selection for developing pedagogically effective tutorial dialogue tactics. *The 1st International Conference on Educational Data Mining,* Montreal, Canada. 258-265.

Chi, M., Jordan, P., VanLehn, K., & Litman, D. (2009). To elicit or to tell: Does it matter? *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 197-204.

Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems,* 224-234.

Chi, M., VanLehn, K., & Litman, D. (2011). An Evaluation of Pedagogical Tutorial Tactics for a Natural Language Tutoring System: A Reinforcement Learning Approach. *International Journal of Artificial Intelligence in Education, Special Issue on "The Best of ITS 2010".*

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32*(2), 301-341.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics,* Budapest, Hungary. 67-74.

D'Mello, S., Taylor, R. S., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. *Proceedings of the 29th Annual Cognitive Science Society,* 203-208.

Evens, M., & Michael, J. (2006). *One-on-one tutoring by humans and computers*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. *Proceedings of the 14th International Conference on Artificial Intelligence and Education,* 33-40.

Forbes-Riley, K., & Litman, D. (2006). Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, 264-271.

Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., & Chen, L. (2010). Generating proactive feedback to help students stay on track. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems,* 315-317.

Fox, B. A. (1993). *The human tutorial dialogue project*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495–522.

Graesser, A. C., Person, N., & Magliano, J. (2004). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Journal of Applied Cognitive Psychology, 9*, 269-306.

Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008). Using hidden Markov models to characterize student behaviors in learning-by-teaching environments. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* Montreal, Canada. 614-625.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education, 13*(1), 79-116.

Kersey, C., Di Eugenio, B., Jordan, P., & Katz, S. (2009). KSC-PaL: A peer learning agent that encourages students to take the initiative. *Proceedings of the NAACL*

HLT Workshop on Innovative use of NLP for Building Educational Applications, Boulder, Colorado. 55-63.

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. L. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie, & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75-105). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Litman, D., & Forbes-Riley, K. (2006). Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering, 12*(2), 161-176.

Litman, D., Moore, J., Dzikovska, M., & Farrow, E. (2009). Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 149-156.

Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., & Kershaw, T. C. (2007). Beyond the code-and-count analysis of tutoring dialogues. *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* 349-356.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Rosé, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2000). *A comparative evaluation of Socratic versus didactic tutoring* No. #LRDC-BEE-1)

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). The role of why questions in effective human tutoring. *Proceedings of the International Conference on Artificial Intelligence in Education,* 55-62.

Soller, A., & Stevens, R. (2007). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 217-253) Information Age Publishing.

Tetreault, J. R., & Litman, D. J. (2008). A reinforcement learning approach to evaluating state representations in spoken dialogue systems. *Speech Communication, 50*(8-9), 683-696.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics,* 271-280.