# Inferring Tutorial Dialogue Structure with Hidden Markov Modeling

**Kristy Elizabeth Boyer**[a] **Eun Young Ha**[a] **Robert Phillips**[ab] **Michael D. Wallis**[ab] **Mladen A. Vouk**[a] **James C. Lester**[a]

[a]Department of Computer Science, North Carolina State University
[b]Applied Research Associates
Raleigh, NC, USA

{keboyer, eha, rphilli, mdwallis, vouk, lester}@ncsu.edu

## Abstract

The field of intelligent tutoring systems has seen many successes in recent years. A significant remaining challenge is the automatic creation of corpus-based tutorial dialogue management models. This paper reports on early work toward this goal. We identify tutorial dialogue *modes* in an unsupervised fashion using hidden Markov models (HMMs) trained on input sequences of manually-labeled dialogue acts and adjacency pairs. The two best-fit HMMs are presented and compared with respect to the dialogue structure they suggest; we also discuss potential uses of the methodology for future work.

## 1 Introduction

The field of intelligent tutoring systems has made great strides toward bringing the benefits of one-on-one tutoring to a wider population of learners. Some intelligent tutoring systems, called *tutorial dialogue systems*, support learners by engaging in rich natural language dialogue, *e.g.*, (Graesser *et al.* 2003; Zinn, Moore & Core 2002; Evens & Michael 2006; Aleven, Koedinger & Popescu 2003; Litman *et al.* 2006; Arnott, Hastings & Allbritton 2008; VanLehn *et al.* 2002). However, creating these systems comes at a high cost: it entails handcrafting each pedagogical strategy the tutor might use and then realizing these strategies in a dialogue management framework that is also custom-engineered for the application. It is hoped that the next generation of these systems can leverage corpora of tutorial dialogue in order to provide more robust dialogue management models that capture the discourse phenomena present in effective natural language tutoring.

The structure of tutorial dialogue has traditionally been studied by manually examining corpora and focusing on cognitive and motivational aspects of tutorial strategies (*e.g.*, Lepper *et al.* 1993; Graesser, Person & Magliano 1995). While these approaches yielded foundational results for the field, such analyses suffer from two serious limitations: manual approaches are not easily scalable to different or larger corpora, and the rigidity of handcrafted dialogue structure tagging schemes may not capture all the phenomena that occur in practice.

In contrast, the stochastic nature of dialogue lends itself to description through probabilistic models. In tutorial dialogue, some early work has adapted language processing techniques, namely *n*-gram analyses, to examine human tutors' responses to student uncertainty (Forbes-Riley & Litman 2005), as well as to find correlations between local tutoring strategies and student outcomes (Boyer *et al.* 2008). However, this work is limited by its consideration of small dialogue windows.

Looking at a broader window of turns is often accomplished by modeling the dialogue as a Markov decision process. With this approach,

techniques such as reinforcement learning can be used to compare potential policies in terms of effectiveness for student learning. Determining relevant feature sets (Tetreault & Litman 2008) and conducting focussed experiments for localized strategy effectiveness (Chi *et al.* 2008) are active areas of research in this line of investigation. These approches often fix the dialogue structures under consideration in order to compare the outcomes associated with those structures or the features that influence policy choice.

In contrast to treating dialogue structure as a fixed entity, one approach for modeling the progression of complete dialogues involves learning the higher-level structure in order to infer succinct probabilistic models of the interaction. For example, data-driven approaches for discovering dialogue structure have been applied to corpora of human-human task-oriented dialogue using general models of task structure (Bangalore, Di Fabbrizio & Stent 2006). Encouraging results have emerged from using a general model of the task structure to inform automatic dialogue act tagging as well as subtask segmentation.

Our current work examines a modeling technique that does not require *a priori* knowledge of the task structure: specifically, we propose to use hidden Markov models (HMMs) (Rabiner 1989) to capture the structure of tutorial dialogue implicit within sequences of tagged dialogue acts. Such probablistic inference of discourse structure has been used in recent work with HMMs for topic identification (Barzilay & Lee 2004) and related graphical models for segmenting multi-party spoken discourse (Purver *et al.* 2006). Analogously, our current work focuses on identifying dialogic structures that emerge during tutorial dialogue. Our approach is based on the premise that at any given point in the tutorial dialogue, the collaborative interaction is "in" a dialogue *mode* (Cade *et al.* 2008) that characterizes the nature of the exchanges between tutor and student; these modes correspond to the hidden states in the HMM. Results to date suggest that meaningful descriptive models of tutorial dialogue can be generated by this simple stochastic modeling technique. This paper focuses on the comparison of two first-order HMMs: one trained on sequences of dialogue acts, and the second trained on sequences of adjacency pairs.

## 2 Corpus Analysis

The HMMs were trained on a corpus of human-human tutorial dialogue collected in the domain of introductory computer science. Forty-three learners interacted remotely with one of fourteen tutors through a keyboard-to-keyboard remote learning environment yielding 4,864 dialogue moves.

### 2.1 Dialogue Act Tagging

The tutoring corpus was manually tagged with dialogue acts designed to capture the salient characteristics of the tutoring process (Table 1).

| Tag | Act | Example |
|---|---|---|
| Q | Question | *Where should I Declare i?* |
| EQ | Evaluation Question | *How does that look?* |
| S | Statement | *You need a closing brace.* |
| G | Grounding | *Ok.* |
| EX | Extra-Domain | *You may use your book.* |
| PF | Positive Feedback | *Yes, that's right.* |
| LF | Lukewarm Feedback | *Sort of.* |
| NF | Negative Feedback | *No, that's not right.* |

Table 1. Dialogue Act Tags

The correspondence between utterances and dialogue act tags is one-to-one; compound utterances were split by the primary annotator prior to the inter-rater reliability study.[1] This dialogue act tagging effort produced sequences of dialogue acts that have been used in their un-altered forms to train one of the two HMMs presented here (Section 3).

### 2.2 Adjacency Pair Identification

In addition to the HMM trained on sequences of individual dialogue acts, another HMM was trained on sequences of dialogue act adjacency pairs. The importance of adjacency pairs is well-established in natural language dialogue (*e.g.*, Schlegoff & Sacks 1973), and adjacency pair analysis has illuminated important phenomena in tutoring as well (Forbes-Riley *et al.* 2007). The

[1] Details of the study procedure used to collect the corpus, as well as Kappa statistics for inter-rater reliability, are reported in (Boyer *et al.* 2008).

intuition behind adjacency pairs is that certain dialogue acts naturally occur together, and by grouping these acts we capture an exchange between two conversants in a single structure. This formulation is of interest for our purposes because when treating sequences of dialogue acts as a Markov process, with or without hidden states, the addition of adjacency pairs may offer a semantically richer observation alphabet.

To find adjacency pairs we utilize a $\chi^2$ test for independence of the categorical variables $act_i$ and $act_{i+1}$ for all sequential pairs of dialogue acts that occur in the corpus. Only pairs in which speaker($act_i$) ≠ speaker($act_{i+1}$) were considered. Table 2 displays a list of all dependent adjacency pairs sorted by descending (unadjusted) statistical significance; the subscript on each dialogue act tag indicates tutor (*t*) or student (*s*).

An adjacency pair joining algorithm was applied to join statistically significant pairs of dialogue acts ($p$<0.01) into atomic units according to a priority determined by the strength of the statistical significance. Dialogue acts that were "left out" of adjacency pair groupings were treated as atomic elements in subsequent analysis. Figure 1 illustrates the application of the adjacency pair joining algorithm on a sequence of dialogue acts from the corpus.

Original dialogue act sequence:

$G_s \rightarrow G_t \rightarrow Q_s \rightarrow S_t \rightarrow EQ_t \rightarrow EQ_s \rightarrow PF_t \rightarrow Q_s$

After adjacency pair joining:

$G_sG_t \rightarrow Q_sS_t \rightarrow EQ_t \rightarrow EQ_sPF_t \rightarrow Q_sS_t$

Figure 1. DA Sequence Before/After Joining

## 3 HMM of Dialogue Structure

A hidden Markov model is defined by three constituents: 1) the set of *hidden states* (dialogue modes), each characterized by its emission probability distribution over the possible *observations* (dialogue acts and/or adjacency pairs), 2) the transition probability matrix among *observations* (dialogue acts and/or adjacency pairs), 2) the transition probability matrix among hidden states, and 3) the initial hidden state (dialogue mode) probability distribution.

| $act_i$ | $act_{i+1}$ | P($act_{i+1}$\| $act_i$) | P($act_{i+1}$\| ¬$act_i$) | $\chi^2$ val | *p*-val |
|---|---|---|---|---|---|
| $EQ_s$ | $PF_t$ | 0.48 | 0.07 | 654 | <0.0001 |
| $G_s$ | $G_t$ | 0.27 | 0.03 | 380 | <0.0001 |
| $EX_s$ | $EX_t$ | 0.34 | 0.03 | 378 | <0.0001 |
| $EQ_t$ | $PF_s$ | 0.18 | 0.01 | 322 | <0.0001 |
| $EQ_t$ | $S_s$ | 0.24 | 0.03 | 289 | <0.0001 |
| $EQ_s$ | $LF_t$ | 0.13 | 0.01 | 265 | <0.0001 |
| $Q_t$ | $S_s$ | 0.65 | 0.04 | 235 | <0.0001 |
| $EQ_t$ | $LF_s$ | 0.07 | 0.00 | 219 | <0.0001 |
| $Q_s$ | $S_t$ | 0.82 | 0.38 | 210 | <0.0001 |
| $EQ_s$ | $NF_t$ | 0.08 | 0.01 | 207 | <0.0001 |
| $EX_t$ | $EX_s$ | 0.19 | 0.02 | 177 | <0.0001 |
| $NF_s$ | $G_t$ | 0.29 | 0.03 | 172 | <0.0001 |
| $EQ_t$ | $NF_s$ | 0.11 | 0.01 | 133 | <0.0001 |
| $S_s$ | $G_t$ | 0.16 | 0.03 | 95 | <0.0001 |
| $S_s$ | $PF_t$ | 0.30 | 0.10 | 90 | <0.0001 |
| $S_t$ | $G_s$ | 0.07 | 0.04 | 36 | <0.0001 |
| $PF_s$ | $G_t$ | 0.14 | 0.04 | 34 | <0.0001 |
| $LF_s$ | $G_t$ | 0.22 | 0.04 | 30 | <0.0001 |
| $S_t$ | $EQ_s$ | 0.11 | 0.07 | 29 | <0.0001 |
| $G_t$ | $EX_s$ | 0.07 | 0.03 | 14 | 0.002 |
| $S_t$ | $Q_s$ | 0.07 | 0.05 | 14 | 0.0002 |
| $G_t$ | $G_s$ | 0.10 | 0.05 | 9 | 0.0027 |
| $EQ_t$ | $EQ_s$ | 0.13 | 0.08 | 8 | 0.0042 |

Table 2. All Dependent Adjacency Pairs

### 3.1 Discovering Number of Dialogue Modes

In keeping with the goal of automatically discovering dialogue structure, it was desirable to learn *n,* the best number of hidden states for the HMM, during modeling. To this end, we trained and ten-fold cross-validated seven models, each featuring randomly-initialized parameters, for each number of hidden states *n* from 2 to 15, inclusive.[2] The average log-likelihood fit from ten-fold cross-

[2] $n$=15 was chosen as an initial maximum number of states because it comfortably exceeded our hypothesized range of 3 to 7 (informed by the tutoring literature). The Akaike Information Criterion measure steadily worsened above $n$ = 5, confirming no need to train models with $n$ > 15.

validation was computed across all seven models for each $n$, and this average log-likelihood $l_n$ was used to compute the Akaike Information Criterion, a maximum-penalized likelihood estimator that prefers simpler models (Scott 2002). This modeling approach was used to train HMMs on both the dialogue act and the adjacency pair input sequences.

### 3.2 Best-Fit Models

The input sequences of individual dialogue acts contain 16 unique symbols because each of the 8 dialogue act tags (Table 1) was augmented with a label of the speaker, either tutor or student. The best-fit HMM for this input sequence contains $n_{DA}$=5 hidden states. The adjacency pair input sequences contain 39 unique symbols, including all dependent adjacency pairs (Table 2) along with all individual dialogue acts because each dialogue act occurs at some point outside an adjacency pair. The best-fit HMM for this input sequence contains $n_{AP}$=4 hidden states. In both cases, the best-fit number of dialogue modes implied by the hidden states is within the range of what is often considered in traditional tutorial dialogue analysis (Cade *et al.* 2008; Graesser, Person & Magliano 1995).

## 4 Analysis

Evaluating the impact of grouping the dialogue acts into adjacency pairs requires a fine-grained examination of the generated HMMs to gain insight into how each model interprets the student sessions.

### 4.1 Dialogue Act HMM

Figure 2 displays the emission probability distributions for the dialogue act HMM. State $0_{DA}$, *Tutor Lecture*,[3] is strongly dominated by tutor statements with some student questions and positive tutor feedback. State $1_{DA}$ constitutes *Grounding/Extra-Domain*, a conversational state consisting of acknowledgments, backchannels, and discussions that do not relate to the computer science task. State $2_{DA,}$ *Student Reflection,* generates student evaluation questions, statements, and positive and negative feedback. State $3_{DA}$ is comprised of tutor utterances, with positive feedback occurring most commonly followed by statements, grounding, lukewarm feedback, and negative feedback. This state is interpreted as a *Tutor Feedback* mode. Finally, State $4_{DA}$, *Tutor Lecture/Probing*, is characterized by tutor statements and evaluative questions with some student grounding statements.

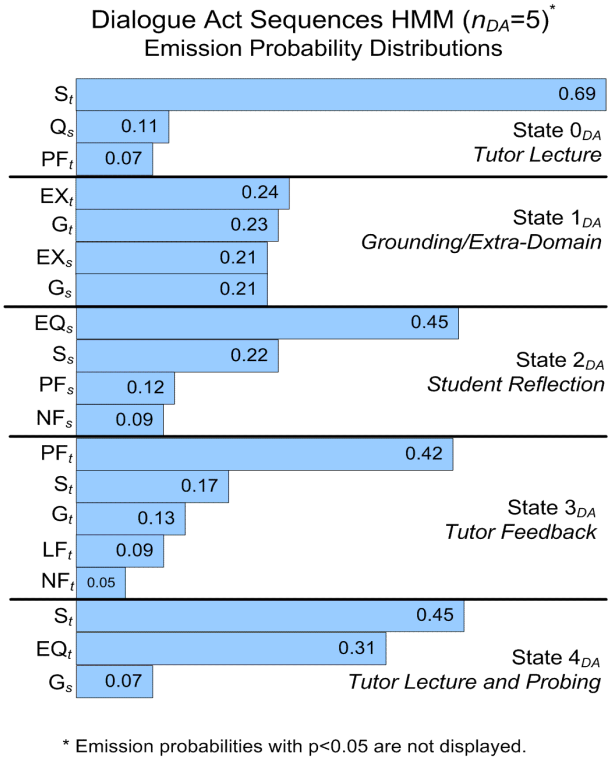


Figure 2. Emission Probability Distributions for Dialogue Act HMM

The state transition diagram (Figure 3) illustrates that *Tutor Lecture* ($0_{DA}$) and *Grounding/Extra-Domain* ($1_{DA}$) are stable states whose probability of self-transition is high: 0.75 and 0.79, respectively. Perhaps not surprisingly, *Student Reflection* ($2_{DA}$) is most likely to transition to *Tutor Feedback* ($3_{DA}$) with probability 0.77. *Tutor Feedback* ($3_{DA}$) transitions to *Tutor Lecture* ($0_{DA}$) with probability 0.60, *Tutor Lecture/Probing* ($4_{DA}$) with probability 0.26, and *Student Reflection* ($2_{DA}$) with probability 0.09. Finally, *Tutor Lecture/Probing* ($4_{DA}$) very often transitions to *Student Reflection* ($2_{DA}$) with probability 0.82.

[3] For simplicity, the states of each HMM have been named according to an intuitive interpretation of the emission probability distribution.

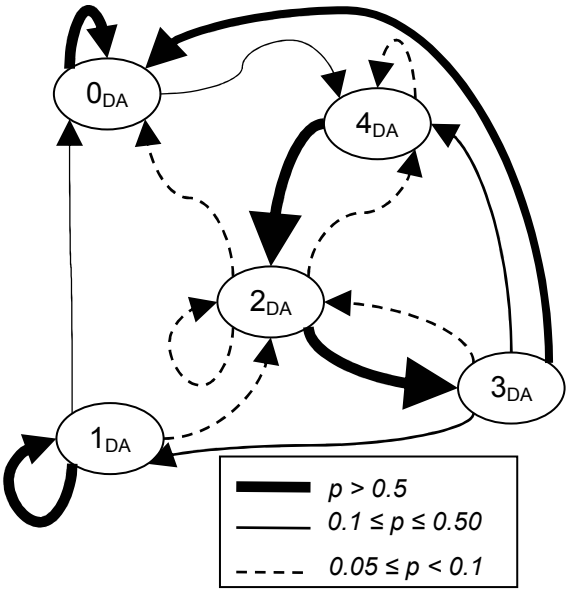


Figure 3. Transition diagram for dialogue act HMM

### 4.2 Adjacency Pair HMM

Figure 4 displays the emission probability distributions for the HMM that was trained on the input sequences of adjacency pairs. State $0_{AP}$, *Tutor Lecture*, consists of tutorial statements, positive feedback, and dialogue turns initiated by student questions. In this state, student evaluation questions occur in adjacency pairs with positive tutor feedback, and other student questions are answered by tutorial statements. State $1_{AP}$, *Tutor Evaluation*, generates primarily tutor evaluation questions, along with the adjacency pair of tutorial statements followed by student acknowledgements. State $2_{AP}$ generates conversational grounding and extra-domain talk; this *Grounding/Extra-Domain* state is dominated by the adjacency pair of student grounding followed by tutor grounding. State $3_{AP}$ is comprised of several adjacency pairs: student questions followed by tutor answers, student statements with positive tutor feedback, and student evaluation questions followed by positive feedback. This *Question/Answer* state also generates some tutor grounding and student evaluation questions outside of adjacency pairs.

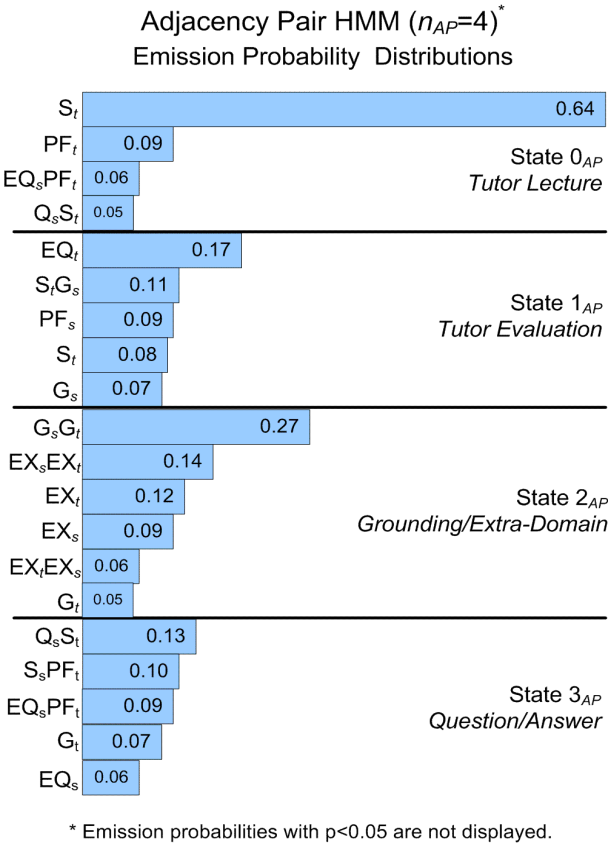


Figure 4. Emission Probability Distributions for Adjacency Pair HMM

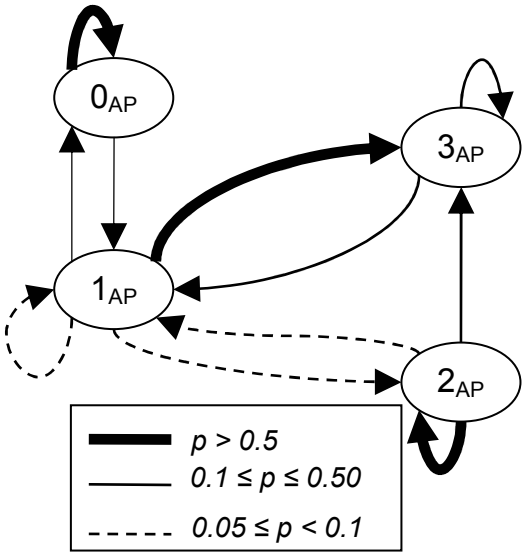


Figure 5. Transition diagram for adjacency pair HMM

### 4.3 Dialogue Mode Sequences

In order to illustrate how the above models fit the data, Figure 6 depicts the progression of dialogue modes that generate an excerpt from the corpus.

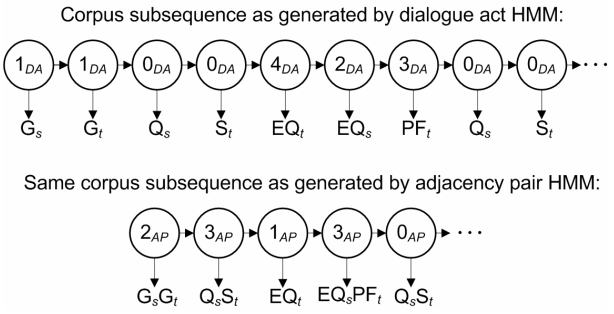


Figure 6. Best-fit sequences of hidden states

In both models, the most commonly-occurring dialogue mode is *Tutor Lecture*, which generates 45% of observations in the dialogue act model and around 60% in the adjacency pair model. Approximately 15% of the dialogue act HMM observations are fit to each of states *Student Reflection*, *Tutor Feedback*, and *Tutor Lecture/Probing*. This model spends the least time, around 8%, in *Grounding/Extra Domain.* The adjacency pair model fits approximately 15% of its observations to each of *Tutor Evaluation* and *Question/Answer*, with around 8% in *Grounding/Extra-Domain*.

### 4.4 Model Comparison

While the two models presented here describe the same corpus, it is important to exercise caution when making direct structural comparisons. The models contain neither the same number of hidden states nor the same emission symbol alphabet; therefore, our comparison will be primarily qualitative. It is meaningful to note, however, that the adjacency pair model with $n_{AP}$=4 achieved an average log-likelihood fit on the training data that was 5.8% better than the same measure achieved by the dialogue act model with $n_{DA}$=5, despite the adjacency pair input sequences containing greater than twice the number of unique symbols.[4]

[4] This comparison is meaningful because the models depicted here provided the best fit among all sizes of models trained for the same input scenario.

Our qualitative comparison begins by examining the modes that are highly similar in the two models. State $2_{AP}$ generates grounding and extra-domain statements, as does State $1_{DA}$. These two states both constitute a *Grounding/Extra-Domain* dialogue mode. One artifact of the tutoring study design is that all sessions begin in this state due to a compulsory greeting that signaled the start of each session. More precisely, the initial state probability distribution for each HMM assigns probability 1 to this state and probability 0 to all other states.

Another dialogue mode that is structurally similar in the two models is *Tutor Lecture*, in which the majority of utterances are tutor statements. This mode is captured in State 0 in both models, with State $0_{AP}$ implying more detail than State $0_{DA}$ because it is certain in the former that some of the tutor statements and positive feedback occurred in response to student questions. While student questions are present in State $0_{DA}$, no such precise ordering of the acts can be inferred, as discussed in Section 1.

Other states do not have one-to-one correspondence between the two models. State $2_{DA}$, *Student Reflection,* generates only student utterances and the self-transition probability for the state is very low; the dialogue usually visits State $2_{DA}$ for one turn and then transitions immediately to another state. Although this aspect of the model reflects the fact that students rarely keep the floor for more than one utterance at a time in the corpus, such quick dialogue mode transitions are inconsistent with an intuitive understanding of tutorial dialogue modes as meta-structures that usually encompass more than one dialogue turn. This phenomenon is perhaps more accurately captured in the adjacency pair model. For example, the dominant dialogue act of State $2_{DA}$ is a student evaluation question ($EQ_s$). In contrast, these dialogue acts are generated as part of an adjacency pair by State $3_{AP}$; this model joins the student questions with subsequent positive feedback from the tutor rather than generating the question and then transitioning to a new dialogue mode. Further addressing the issue of frequent state transitions is discussed as future work in Section 6.

## 5 Discussion and Limitations

Overall, the adjacency pair model is preferable for our purposes because its structure lends itself more readily to interpretation as a set of dialogue modes each of which encompasses more than one dialogue move. This structural property is guaranteed by the inclusion of adjacency pairs as atomic elements. In addition, although the set of emission symbols increased to include significant adjacency pairs along with all dialogue acts, the log-likelihood fit of this model was slightly higher than the same measure for the HMM trained on the sequences of dialogue acts alone. The remainder of this section focuses on properties of the adjacency pair model.

One promising result of this early work emerges from the fact that by applying hidden Markov modeling to sequences of adjacency pairs, meaningful dialogue modes have emerged that are empirically justified. The number of these dialogue modes is consistent with what researchers have traditionally used as a set of hypothesized tutorial dialogue modes. Moreover, the composition of the dialogue modes reflects some recognizable aspects of tutoring sessions: tutors teach through the *Tutor Lecture* mode and give feedback on student knowledge in a *Tutor Evaluation* mode. Students ask questions and state their own perception of their knowledge in a *Question/Answer* mode. Both parties engage in "housekeeping" talk containing such things as greetings and acknowledgements, and sometimes, even in a controlled environment, extra-domain conversation occurs between the conversants in the *Grounding/Extra-Domain* mode.

Although the tutorial modes discovered may not map perfectly to sets of handcrafted tutorial dialogue modes from the literature (*e.g.,* Cade *et al.* 2008), it is rare for such a perfect mapping to exist even between those sets of handcrafted modes. In addition, the HMM framework allows for succinct probabilistic description of the phenomena at work during the tutoring session: through the state transition matrix, we can see the back-and-forth flow of the dialogue among its modes.

## 6 Conclusions and Future Work

Automatically learning dialogue structure is an important step toward creating more robust tutorial dialogue management systems. We have presented two hidden Markov models in which the hidden states are interpreted as *dialogue modes* for task-oriented tutorial dialogue. These models were learned in an unsupervised fashion from manually-labeled dialogue acts. HMMs offer concise stochastic models of the complex interaction patterns occurring in natural language tutorial dialogue. The evidence suggests this methodology, which as presented requires only a sequence of dialogue acts as input, holds promise for automatically discovering the structure of tutorial dialogue.

Future work will involve conducting evaluations to determine the benefits gained by using HMMs compared to simpler statistical models. In addition, it is possible that more general types of graphical models will prove useful in overcoming some limitations of HMMs, such as their arbitrarily frequent state transitions, to more readily capture the phenomena of interest. The descriptive insight offered by these exploratory models may also be increased by future work in which the input sequences are enhanced with information about the surface-level content of the utterance. In addition, knowledge of the task state within the tutoring session can be used to segment the dialogue in meaningful ways to further refine model structure.

It is also hoped that these models can identify empirically-derived tutorial dialogue structures that can be associated with measures of effectiveness such as student learning (Soller & Stevens 2007). These lines of investigation could inform the development of next-generation natural language tutorial dialogue systems.

## Acknowledgments

Thanks to Marilyn Walker and Dennis Bahler for insightful early discussions on the dialogue and machine learning aspects of this work, respectively. This research was supported by the National Science Foundation under Grants REC-0632450, IIS-0812291, CNS-0540523, and GRFP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References


Aleven, V., K. Koedinger, and O. Popescu. 2003. A tutorial dialog system to support self-explanation: Evaluation and open questions. *Proceedings of the 11th International Conference on Artificial Intelligence in Education*: 39-46.

Arnott, E., P. Hastings, and D. Allbritton. 2008. Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavioral Research Methods* 40(3): 694-698.

Bangalore, S., Di Fabbrizio, G., and Stent, A. 2006. Learning the structure of task-driven human-human dialogs. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*: 201-208.

Barzilay, R., and Lee, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of NAACL HLT*: 113–120.

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

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

Chi, M., Jordan, P., VanLehn, K., and Hall, M. 2008. Reinforcement learning-based feature selection for developing pedagogically effective tutorial dialogue tactics. *Proceedings of the 1st International Conference on Educational Data Mining*: 258-265.

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

Forbes-Riley, K., and Litman, D. J. 2005. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*: 87-96.

Forbes-Riley, K., Rotaru, M., Litman, D. J., and Tetreault, J. 2007. Exploring affect-context dependencies for adaptive system development. *Proceedings of NAACL HLT*: 41-44.

Graesser, A., G. Jackson, E. Mathews, H. Mitchell, A. Olney, M. Ventura, and P. Chipman. 2003. Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*: 1-6.

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

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

Litman, D. J., C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education* 16(2): 145-170.

Purver, M., Kording, K. P., Griffiths, T. L., and Tenenbaum, J. B. 2006. Unsupervised topic modelling for multi-party spoken discourse. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*: 17-24.

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

Schlegoff, E., and H. Sacks. 1973. Opening up closings. *Semiotica* 7(4): 289-327.

Scott, S. L. 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* 97(457): 337-352.

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

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

VanLehn, K., P. W. Jordan, C. P. Rose, D. Bhembe, M. Bottner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, and A. Roque. 2002. The architecture of Why2-atlas: A coach for qualitative physics essay writing. *Proceedings of Intelligent Tutoring Systems Conference*: 158–167.

Zinn, C., Moore, J. D., and Core, M. G. 2002. A 3-tier planning architecture for managing tutorial dialogue. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*: 574-584.