Comparing Linguistic Features for Modeling Learning in Computer Tutoring

Kate FORBES-RILEY, Diane LITMAN, Amruta PURANDARE,
Mihai ROTARU and Joel TETREAULT
University of Pittsburgh, Pittsburgh, PA, 15260, USA

Abstract. We compare the relative utility of different automatically computable linguistic feature sets for modeling student learning in computer dialogue tutoring. We use the PARADISE framework (multiple linear regression) to build a learning model from each of 6 linguistic feature sets: 1) surface features, 2) semantic features, 3) pragmatic features, 4) discourse structure features, 5) local dialogue context features, and 6) all feature sets combined. We hypothesize that although more sophisticated linguistic features are harder to obtain, they will yield stronger learning models. We train and test our models on 3 different train/test dataset pairs derived from our 3 spoken dialogue tutoring system corpora. Our results show that more sophisticated linguistic features usually perform better than either a baseline model containing only pretest score or a model containing only surface features, and that semantic features generalize better than other linguistic feature sets.

Keywords. Tutoring Dialogue Systems, Linguistics, Multiple Linear Regression

1. Introduction

Computer tutoring *dialogue* systems exploit natural language interaction in an attempt to improve student learning. A variety of features of natural language dialogue appear useful for modeling learning during tutoring. For example, longer student turns and higher percents of student words and turns were shown to correlate with learning in tutoring dialogues [1,2,3], as were specific dialogue acts (e.g., tutor feedback and question types) [1,2,4,5,6], discourse structure features, and local dialogue context features [4,7].

However, such features differ both in terms of how sophisticated they are linguistically and how easy they are to obtain. For example, turn counts represent surface linguistic properties that are easily computed by tutoring systems. Dialogue acts represent deeper pragmatic properties, but usually require manual labeling during system design or in a system corpus, even if automatic labeling is developed from the manual labeling.

We hypothesize that more sophisticated linguistic features will yield stronger models of student learning, making the extra labor to obtain them "worth the effort" from a system design point of view. These models can be used off-line to improve system dialogue design, or, if the model features are automatically computable in real time, on-line in adaptive systems that respond during tutoring. [8,9] support this hypothesis; e.g. in [8], students learned more from a system with more sophisticated dialogue feedback.

In this paper we examine this hypothesis, by comparing the utility of 6 linguistic feature sets (Section 2.2) for modeling learning in our tutoring system (Section 2.1). The feature sets represent different levels of linguistic sophistication and effort that our sys-

tem can automatically compute: 1) surface features, 2) semantic features, 3) pragmatic features, 4) discourse structure features, 5) local dialogue context features, and 6) all features combined. Prior work has not examined the relative utility of all these feature sets for modeling learning. We use the PARADISE framework [10] (Section 3), which is an application of multiple linear regression to dialogue system evaluation, to model learning. We train and test our models on 3 different train/test dataset pairs derived from our 3 spoken dialogue tutoring system corpora. Our results (Section 3.1) show that most sophisticated linguistic feature sets usually out-perform the baseline model and the surface feature model. We also find that semantic features generalize better than other linguistic feature sets, including those that are more labor-intensive to obtain.

2. Tutoring System Corpora and Linguistic Feature Sets

2.1. ITSPOKE System and Corpora

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) [4] is a *speech-enabled* version of the *text-based* Why2-Atlas tutoring system [11]. In ITSPOKE, the student first types an essay into a web interface answering a qualitative physics problem. The backend parses the essay and extracts a set of dialogue topics concerning misconceptions or incomplete explanations. ITSPOKE then engages the student in a dialogue that covers these topics. After the dialogue, the student revises the essay, thereby ending the tutoring for the problem or causing another round of dialogue/essay revision.

. . . . . spoken dialogue excerpt at 6.1 minutes after start of problem. . .

ITSPOKE$_4$: Good. Does the packet have an acceleration? If yes please specify its direction. *[ShortAnswerQ, Advance, Depth1, GoalRep0]*

STUDENT$_4$: horizontally forward yes *[Incorrect]*

> **ITSPOKE**$_5$: Well... To find the packet's acceleration we will use Newton's second law. What is the definition of this law? *[LongAnswerQ, Push, Depth2, GoalRep0]*
>
> **STUDENT**$_5$: an object in motion tends to stay in motion until it's act by an outside force *[Incorrect]*
>
> **ITSPOKE**$_6$: Well... Newton's second law says that the net force acting on an object equals the mass times the acceleration of the object. We can define Newton's second law as F=m*a. Hence what can we determine about the acceleration of the packet given that we know the net force acting on it is not zero and points straight down? *[DeepAnswerQ, Advance, Depth2, GoalRep0]*
>
> **STUDENT**$_6$: I don't know *[Can't Answer]*

ITSPOKE$_7$: Well... Since the net force F and the acceleration a are vectors, for the equation F=m*a to hold they both need to have the same direction. <*...more...*> Does the packet have an acceleration? If yes please specify its direction. *[ShortAnswerQ, PopUp, Depth1, GoalRep1]*

Figure 1. ITSPOKE Annotated Dialogue Excerpt

We use 3 ITSPOKE corpora for this study, referred to as: **S03**, **P05**, and **S05**. The user population of **S03** is different from **P05** and **S05**, because it was collected for a different purpose in a different year using a different recruitment method; all 3 corpora use slightly different ITSPOKE versions. The same procedure was used to collect the corpora: students with no college physics: 1) read a document of background material, 2) took a pretest, 3) worked 5 problems with ITSPOKE (each problem yields 1 dialogue), 4) took a posttest. The **S03** corpus contains 20 students. The **P05** corpus contains 27 students. The **S05** corpus contains 27 students. Figure 1 shows a corpus excerpt.

2.2. Linguistic Feature Sets

We computed 207 linguistic features per student in each corpus. By "per student" we mean each feature was calculated over all 5 dialogues of the student. Although all the features can be automatically computed by our system, they represent different levels of linguistic sophistication and effort. In addition, note that we computed all student turn-based features from a human transcription of the student turns (rather than the speech recognition output), both to estimate an upper bound on these features' usefulness for modeling learning, and to enable comparison with text-based tutoring systems.

1. Surface Feature Set: We computed 21 surface linguistic features per student; some have been used to model learning in other systems [1,2,3]. 18 features represent surface measures of student and tutor dialogue contribution. For each of the 6 groups in Figure 2, we computed 3 features: *Total Words, Total Turns, Average Words per Turn*. In addition, we computed 1 feature representing a temporal measure of total contribution: *Total Elapsed Time*, and 2 features representing speech recognition quality: *Total Timeouts* and *Total Rejections*[1]. These features all require minimal effort to compute.

S: Student spoken turns (transcribed)	**ST:** Student and tutor spoken turns
T: Tutor spoken turns (transcribed)	**SE:** Student contribution (spoken and essay)
E: Student typed essay turns	**STE:** Student and tutor contribution (spoken and essay)

Figure 2. Groupings Representing Student and Tutor Contributions

2. Semantic Feature Set: We computed 38 semantic features per student. 24 features represent semantic measures of student and tutor dialogue contributions. For each of the 6 groups in Figure 2, we computed 4 features: *Total Concepts, Average Concepts per Turn, Concept-Word Ratio, Total Unique Concepts*. Our notion of "Concept" distinguishes physics content words from other words in the dialogues. Our current method of concept extraction required some additional effort to implement: we count all words (concept tokens) in the student turns that appear in an online physics dictionary. This "Total Concepts" value is used to compute the average and ratio. "Total Unique Concepts" is computed by counting the number of *different* concepts (types) over all the student turns. For example, STUDENT_5 in Figure 1 contains 2 Unique Concepts: "motion" and "force", and 3 Total Concepts: "motion", "motion" and "force".

Our remaining 14 semantic features represent the correctness of the meaning underlying various surface forms of student answers (e.g., "down", "towards earth"). ITSPOKE automatically labels the correctness of recognized answers, although here we use a version of correctness obtained by feeding the transcribed answers into ITSPOKE[2]. We use 4 Correctness labels: *Correct*, *Partially Correct*, *Incorrect*, *Can't Answer*. *Can't Answer* is used for variants of "I don't know" (STUDENT_6, Figure 1). For each of the 4 Correctness labels, we computed a *Total*, a *Percent*, and a *Ratio* to each other label.

3. Pragmatic Feature Set: We computed 14 pragmatic features per student. 8 features are derived from automatically labeled dialogue acts, which required significant effort to implement. In particular, in prior work we first manually labeled the tutor Ques-

[1] A Timeout occurs when ITSPOKE does not hear speech by a pre-specified time interval. A Rejection occurs when ITSPOKE's confidence score for its recognition output is too low to accept that output.

[2] These two versions of correctness produce an agreement of 91% (0.84 Kappa). We have used various versions of correctness in prior studies [7,12,13].

tion Acts in the **S03** corpus [4].[3] These dialogue acts codify the intent underlying tutor questions. Our dialogues have a Question-Answer format; every ITSPOKE turn asks a question. Our labels, *Short, Long, and Deep Answer Question*, distinguish the type of ITSPOKE question in terms of its content and the type of answer it presupposes. *Repeat Question* labels variants of "Can you repeat that?" after rejections. From our manual annotations, we created a hash table associating each ITSPOKE question with a Question Act label, and used it to automatically label the ITSPOKE questions in the other 2 corpora. For each of the 4 Question Act labels, we computed a *Total* and a *Percent*.

Our remaining 6 pragmatic features are derived from a Goal Repetition variable conceived in prior work to track how often ITSPOKE goals repeat in a dialogue [12]. Each ITSPOKE question is associated with a goal, which repeats if it is not satisfied earlier in the dialogue (see ITSPOKE_7 in Figure 1). For each student, we compute a *Total* and a *Percent* over ITSPOKE goals repeated 0, 1, or 2 times (GoalRep0-GoalRep2).

4. Discourse Structure Feature Set: We computed 35 discourse structure features per student, which derive from significant effort in our prior work showing that the discourse structure Depth and Transition for each ITSPOKE turn can be automatically labeled [7]. The Depth labels (Depth1-Depth4) represent the depth of the turn's topic in the structure. E.g., $\text{ITSPOKE}_{4,7}$ in Figure 1 have Depth 1; $\text{ITSPOKE}_{5,6}$ have Depth 2. Here we computed a *Total* and a *Percent* for each of the 4 Depth labels.

The 6 Transition labels represent the ITSPOKE turn's position in the discourse structure relative to the previous ITSPOKE turn. The first ITSPOKE turn after an essay is labeled *NewTopLevel*. Each ITSPOKE turn at the same depth as the prior ITSPOKE turn is labeled *Advance* ($\text{ITSPOKE}_{4,6}$ in Figure 1). The first ITSPOKE turn in a sub-topic dialogue (which occurs after an incorrect answer to a complex question) is labeled *Push* (ITSPOKE_5). After a sub-topic dialogue completes, ITSPOKE pops up and either asks the original question again, labeled *PopUp* (ITSPOKE_7), or moves on to the next question, labeled *PopUpAdvance*. After student turn timeouts and rejections, respectively, ITSPOKE's question is repeated or some variant of "Can you repeat that?" is used, labeled *SameGoal*. Here we computed a *Total*, a *Percent*, and a *Ratio* to each other label, for each of the 6 Transition labels, yielding our remaining 27 discourse structure features.

5. Local Dialogue Context Feature Set: We computed 99 bigram features per student, representing local dialogue context. These features were derived from prior work that examined bigrams of discourse structure and correctness [7]. The bigrams consist of the labels for two consecutive turns; in particular, bigrams of: 1) Correctness labels, 2) Transition labels, and 3) a Transition label followed by a Correctness label. For example, in Figure 1, the $\text{Student}_4 \sim \text{Student}_5$ turn pair counts as an Incorrect$\sim$Incorrect bigram. The $\text{ITSPOKE}_4 \sim \text{ITSPOKE}_5$ turn pair counts as an Advance$\sim$Push bigram. The $\text{ITSPOKE}_4 \sim \text{Student}_4$ turn pair counts as an Advance$\sim$Incorrect bigram.

Thus we have 16 possible Correctness bigrams, 36 Transition bigrams, and 24 Correctness-Transition bigrams. However, in this study we excluded rarely-occurring bigrams whose total was less than 2 on average across our corpora to reduce overfitting in our learning models. For the remaining Correctness and Transition bigrams, we computed a *Total* and a *Percent* (the denominator is total speaker turns). For the remaining Transition-Correctness bigrams, we computed a *Total*, a *Percent* (the denominator is total student turns) and a *Relative Percent* (the denominator is the Transition Total).

[3] Our Acts are based on similar labels in related work [5]. Two annotators labeled these Acts in 8 dialogues in a parallel human tutoring corpus, yielding agreement of 90% (0.75 Kappa).

3. Student Learning Models

The PARADISE framework [14] uses multiple linear regression (MLR) to model a dialogue system performance metric (M) from an input set of dialogue features. The resulting model predicts M as the sum of n model parameters, p_i, each weighted by a coefficient, w_i. The model can then be used in system design, by treating p_i as a benefit to be maximized if w_i is positive, or as a cost to be minimized if w_i is negative.

Many PARADISE applications model "user satisfaction" (e.g., [10,14,13]). We have also used PARADISE to model learning [7,13], and MLR is used to model learning in other tutoring research (e.g., [6]). We model learning gain by predicting posttest and forcing pretest as the first model parameter; the stepwise procedure automatically selects other features for inclusion in the model. However, as in [14], the procedure can only select features that correlate with posttest controlled for pretest (at $p \leq 0.1$); this can help prevent overfitting and reduce the average error of the model.

Here we train a learning model on each of our 5 linguistic feature sets, and from all feature sets combined, to examine their relative usefulness. As a baseline, we train a model with only pretest. Further, we train the 7 models on 3 different datasets corresponding to all combinations of 2 corpora: **S03+P05**, **S03+S05**, and **P03+S05**. In each case, we test the models on the third corpus. This allows us to investigate how well the models perform and generalize with different user populations and system versions.

3.1. Results

Table 1 shows our 7 learning models trained on the **S03+P05** corpora and tested on the **S05** corpus. The first column shows the feature set used to build the model. The second column shows how many features in each set correlated and were included in the stepwise procedure. The third column presents the trained model: each selected parameter and its weight; larger weights indicate parameters with greater relative predictive power.[4] The last two columns show how well the model performed on the training and test data, respectively, in terms of the amount of Posttest variance accounted for by the model (R^2). For example, the model trained on Surface features contains Pretest and Total Elapsed Time, and outperforms the Pretest model on the training data, but not on the test data.

Considering training, all models outperform the Pretest model. The Semantic, Local Context[5], and All models outperform the Surface model. The All model performs best, and only contains Semantic and Local Context features. The Pragmatic and Discourse Structure feature sets performed worst. No Pragmatic features correlated at $p \leq 0.1$; one feature correlated at p=.12 and was forced into the model for comparison.

Considering testing, only the Semantic model outperforms the Pretest model, and Pretest is the strongest predictor in all models. This suggests both that Pretest is our most generalizable predictor of Posttest, and that of our linguistic features, Semantic features are the most generalizable. However, all sophisticated linguistic models outperform the Surface model. This suggests that sophisticated linguistic features are "worth the effort" and that it would be worthwhile to expend further effort extending these feature sets to find more generalizable features. Finally, the Semantic model outperforms the All model, suggesting that other procedures may be better than stepwise for developing an "absolute" best learning model.

[4] These weights are the standardized coefficients (beta weights), i.e. based on scaling of the input features.

[5] A fourth parameter in this model was excluded, because it was highly correlated with another parameter ($R \geq .70$); inclusion of highly correlated parameters can affect the coefficients [14].

Table 1. Learning Models Trained on S03+P05 (47 Students) and Tested on S05 (27 Students)

Feature Set	# Feats	Learning Model	train R^2	test R^2
Pretest	n/a	0.57*Pretest	0.319	0.377
+Surface	4	0.70*Pretest + 0.31*Elapsed_Time#	0.396	0.293
+Semantic	20	0.48*Pretest + 0.33*S_Concept/Word + 0.22*E_Unique_Concept#	0.501	0.386
+Pragmatic	1 (p=.12)	0.62*Pretest + 0.20*LongAnswerQ#	0.356	0.323
+Disc. Str.	3	0.46*Pretest - 0.27*PopUp/Advance	0.378	0.323
+Local	15	0.58*Pretest - 0.34*PopUp∼Incorrect% + 0.34*Advance∼Advance#	0.531	0.330
+All	42	0.51*Pretest + 0.41*Advance∼Advance# - 0.33*PopUp∼Incorrect# - 0.31*PopUp∼Advance% + 0.29*STE_Concept/Turn	0.673	0.321

Table 2 shows our results for training on **S03+S05** and testing on **P05**. This training set and the prior one combine the 2003 and 2005 user populations; thus we expected them to yield similar results and more generalizable models.

Table 2. Learning Models Trained on S03+S05 (47 Students) and Tested on P05 (27 Students)

Feature Set	# Feats	Learning Model	train R^2	test R^2
Pretest	n/a	0.54*Pretest	0.288	0.452
+Surface	3	0.56*Pretest + 0.23*SE_Words#	0.341	0.356
+Semantic	18	0.51*Pretest + 0.32*S_Concept/Word	0.392	0.524
+Pragmatic	1 (p=.11)	0.48*Pretest - 0.21*RepeatQ#	0.330	0.500
+Disc. Str.	6	0.58*Pretest - 0.26*PopUp/Push	0.352	0.278
+Local	20	0.60*Pretest - 0.32*PopUp∼Incorrect% + 0.30*Push∼Push%	0.482	0.304
+All	47	0.64*Pretest - 0.33*PopUp∼Incorrect% + 0.26*STE_Concept/Word + 0.23*Push∼Push%	0.541	0.375

Considering training, we see considerable similarity with the prior table. Again all linguistic models outperform the Pretest model, and the Semantic, Local Context, and All models outperform the Surface model; however, here the Discourse Structure model does too. Again the All model is best and contains only Semantic and Local Context features. The Pragmatic set again had one feature forced into the model for comparison.

Considering testing, we continue to see considerable similarity with the prior table. The Semantic model again outperforms the Pretest model, and contains the two of the same parameters as its counterpart in the prior table. However, here the Pragmatic model also outperforms the Pretest model, which suggests that using only correlated features may not produce an "absolute" best model. Again no other models outperform the Pretest model. Again the Semantic, Pragmatic, and All models again outperform the Surface model, although here the Discourse Structure and Local Context models do not. How-

ever, the All model contains the features of the Local Context model plus one additional Semantic feature, and these Local Context features are derived from bigrams of Discourse Structure features. Overall our results in these tables suggest both that Semantic features are our most generalizable linguistic predictors of Posttest, and that our other more sophisticated linguistic feature sets are probably "worth the effort" too.

Table 3 shows our results for training on **P05+S05** and testing on **S03**. Since this training set includes only 2005 students, we expected less similarity and generalizability.

Table 3. Learning Models Trained on P05+S05 (54 Students) and Tested on S03 (20 Students)

Feature Set	# Feats	Learning Model	train R^2	test R^2
Pretest	n/a	0.64*Pretest	0.412	0.214
+Surface	1	0.59*Pretest - 0.21*Rejections#	0.452	0.259
+Semantic	8	-0.40*Incorrect/Correct + 0.35*Pretest + 0.24 S_Concept/Word	0.585	0.338
+Pragmatic	2	0.53*Pretest + 0.27*GoalRep0%	0.470	0.135
+Disc. Str.	7	0.55*Pretest - 0.23*Depth4%	0.455	0.145
+Local	28	0.40*Pretest - 0.31*PopUp∼Incorrect# + 0.28*Can'tAnswer∼Correct%rel	0.567	0.284
+All	46	-0.40*Incorrect/Correct + 0.35*Pretest + 0.24 S_Concept/Word	0.585	0.338

Considering training, we see more similarity than expected. Again all linguistic models outperform the Pretest model; they all also outperform the Surface model. Again the All model performs best; here however it is identical to the Semantic model and a Semantic feature is the strongest predictor. Furthermore, Student Concept-Word Ratio was consistently selected by the Semantic models in all tables, suggesting it is our most generalizable linguistic feature. The Local Context model is again second best.

Considering testing, performance as expected is generally lower as compared to the other tables, but we continue to see similarities. Although here all but 2 models (Pragmatic and Discourse Structure) outperform both the Pretest and Surface models, as in the prior tables, the Semantic model continues to performs best on testing.

4. Conclusions and Current Directions

This paper examined the relative usefulness of linguistic feature sets for modeling learning in computer dialogue tutoring. We hypothesized that more sophisticated linguistic feature sets would yield stronger learning models, making them "worth the effort" to obtain. We built models from 6 different linguistic feature sets on 3 different subsets of our 3 ITSPOKE corpora, then compared how the models generalized to new test sets.

Our results support our hypothesis. The Semantic models consistently outperformed all models on testing, yielding our most generalizable linguistic features, with Student Concept-Word Ratio present in every Semantic model and some Semantic feature(s) in every All model. The Semantic and Local Context models consistently outperformed the Surface model on testing. The performance of the Discourse Structure and Pragmatic models was more variable, but the Pragmatic model outperformed the Surface model on testing in two datasets, while in every dataset either the Discourse Structure model outperformed the Surface model on training, or these features appeared in models (Local or All) that did so. Moreover, no Surface features were selected in any of the All models.

Although our linguistic features can all be computed automatically, more sophisticated features required more effort to conceive and implement than surface features. Our results suggest it would be worthwhile to expend further effort to find more generalizable sophisticated features. For example, we are extending our Semantic feature set to include de-stemmed and synonymous concept words and phrases. We are extending our Local Context features to include other bigrams. We will also add features based on recognized student speech, as well as para-linguistic features, such as affective states.

We will also try other model-building procedures. We used PARADISE based on our prior work in dialogue systems; this can be viewed as one method of data mining for learning-related features. Although our models do not imply causality between model parameters and learning, they suggest hypotheses about how learning can be increased, i.e. by modifying the system to try to minimize negative parameters (costs) and maximize positive parameters (benefits). These hypotheses can be tested by evaluating the new system. For example, the Surface Feature model in Table 3 suggests rejections are a cost. We can test this by lowering ITSPOKE's recognition threshold so it rejects answers less frequently. Incorporating our most generalizable parameters into system design will likely have the most impact on learning in new user populations. However, the hypotheses suggested by these parameters do not always suggest obvious system changes. For example, we cannot directly manipulate the Student Concept-Word Ratio; increasing this ratio may involve changing the system's response during the tutoring.

Acknowledgements

NSF (#0325054 and #0328431) and ONR (N00014-04-1-0108) support this research.

References

[1] M. G. Core, J. D. Moore, and C. Zinn. The role of initiative in tutorial dialogue. In *Proc. EACL*, 2003.

[2] C. P. Rosé, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn. The role of why questions in effective human tutoring. In *Proceedings of AIED*, 2003.

[3] S. Katz, D. Allbritton, and J. Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *Internat. Jnl of Artificial Intelligence and Education*, 13, 2003.

[4] D. J. Litman and K. Forbes-Riley. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Journal of Natural Language Engineering*, 12(2):161–176, 2006.

[5] G. Jackson, N. Person, and A. Graesser. Adaptive tutorial dialogue in AutoTutor. In *Proceedings Workshop on Dialog-based Intelligent Tutoring Systems at ITS*, 2004.

[6] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471–533, 2001.

[7] M. Rotaru and D. Litman. Exploiting discourse structure for spoken dialogue performance analysis. In *Proceedings of EMNLP*, Sydney, Australia, 2006.

[8] B. Di Eugenio, D. Fossati, D. Yu, S. Haller, and M. Glass. Natural language generation for intelligent tutoring systems: a case study. In *Proceedings of AIED*, The Netherlands, 2005.

[9] M. Evens and J. Michael. *One-on-One Tutoring by Humans and Computers*. Lawrence Erlbaum,, 2006.

[10] M. Walker, D. Litman, C. Kamm, and A. Abella. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings ACL/EACL*, pages 271–280, 1997.

[11] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of ITS*, 2002.

[12] J. Tetreault and D. Litman. Comparing the utility of state features in spoken dialogue using reinforcement learning. In *NAACL*, New York City, 2006.

[13] K. Forbes-Riley and D. Litman. Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings HLT-NAACL*, 2006.

[14] M. Hajdinjak and F. Mihelic. The PARADISE evaluation framework: Issues and findings. *Computational Linguistics*, 32(2):263–272, June 2006.