Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters

Kate Forbes-Riley
Learning Research & Development Ctr
University of Pittsburgh
Pittsburgh, PA 15260
forbesk@cs.pitt.edu

Diane J. Litman
Learning Research & Development Ctr
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
litman@cs.pitt.edu

Abstract

We investigate using the PARADISE framework to develop predictive models of system performance in our spoken dialogue tutoring system. We represent performance with two metrics: user satisfaction and student learning. We train and test predictive models of these metrics in our tutoring system corpora. We predict user satisfaction with 2 parameter types: 1) system-generic, and 2) tutoring-specific. To predict student learning, we also use a third type: 3) user affect. Although generic parameters are useful predictors of user satisfaction in other PARADISE applications, overall our parameters produce less useful user satisfaction models in our system. However, generic and tutoring-specific parameters do produce useful models of student learning in our system. User affect parameters can increase the usefulness of these models.

1 Introduction

In recent years the development of *spoken dialogue* tutoring systems has become more prevalent, in an attempt to close the performance gap between human and computer tutors (Mostow and Aist, 2001; Pon-Barry et al., 2004; Litman et al., 2006). *Student learning* is a primary metric for evaluating the performance of these systems; it can be measured, e.g., by comparing student pretests taken prior to system use with posttests taken after system use.

In other types of spoken dialogue systems, the user's subjective judgments about using the system are often considered a primary system performance metric; e.g., user satisfaction has been measured via surveys which ask users to rate systems during use along dimensions such as task ease, speech input/output quality, user expectations and expertise, and user future use (Möller, 2005b; Walker et al., 2002; Bonneau-Maynard et al., 2000; Walker et al., 2000; Shriberg et al., 1992). However, it is expensive to run experiments over large numbers of users to obtain reliable system performance measures.

The PARADISE model (Walker et al., 1997) proposes instead to *predict* system performance, using parameters representing interaction costs and benefits between system and user, including task success, dialogue efficiency, and dialogue quality. More formally, a set of interaction parameters are measured in a spoken dialogue system corpus, then used in a multivariate linear regression to predict the target performance variable. The resulting model is described by the formula below, where there are *n* interaction parameters, p_i, each weighted by the analysis with a coefficient, w_i, which will be negative or positive, depending on whether the model treats p_i as a cost or benefit, respectively. The model can then be used to estimate performance during system design, with the design goals of minimizing costs and maximizing benefits.

$$\textbf{System Performance} = \sum_{i=1}^{n} \mathrm{w}_i * \mathrm{p}_i$$

We investigate using PARADISE to develop predictive models of performance in our spoken dialogue tutoring system. Although to our knowledge,

Corpus	Date	Voice	#Dialogues	#Students	#with Survey	#with Tests	#with Affect
SYN03	2003	synthesized	100	20	0	20	20
PR05	2005	pre-recorded	140	28	28	28	17
SYN05	2005	synthesized	145	29	29	29	0

Table 1: Summary of our 3 ITSPOKE Corpora

prior PARADISE applications have only used user satisfaction to represent performance, we hypothesize that other metrics may be more relevant when PARADISE is applied to tasks that are not optimized for user satisfaction, such as our spoken dialogue tutoring system. We thus use 2 metrics to represent performance: 1) a generic metric of user satisfaction computed via user survey, 2) a tutoring-specific metric of student learning computed via student pretest and posttest scores. We train and test predictive models of these metrics on multiple system corpora.

To predict user satisfaction, we use 2 types of interaction parameters: 1) system-generic parameters such as used in other PARADISE applications, e.g. speech recognition performance, and 2) tutoring-specific parameters, e.g. student correctness. To predict student learning, we also use a third type of parameter: 3) manually annotated user affect. Although prior PARADISE applications have tended to use system-generic parameters, we hypothesize that task-specific and user affect parameters may also prove useful. We emphasize that user affect parameters are still system-generic; user affect has been annotated and/or automatically predicted in other types of spoken dialogue systems, e.g. as in (Lee et al., 2002; Ang et al., 2002; Batliner et al., 2003).

Our results show that, although generic parameters were useful predictors of user satisfaction in other PARADISE applications, overall our parameters produce less useful user satisfaction models in our tutoring system. However, generic and tutoring-specific parameters do produce useful models of student learning in our system. Generic user affect parameters increase the usefulness of these models.

2 Spoken Dialogue Tutoring Corpora

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman et al., 2006) is a *speech-enabled* tutor built on top of the *text-based* Why2-Atlas conceptual physics tutor (VanLehn et al., 2002). In ITSPOKE, a student first types an essay into a web-based interface answering a qualitative physics problem. ITSPOKE then analyzes the essay and engages the student in spoken dialogue to correct misconceptions and elicit more complete explanations. Student speech is digitized from the microphone input and sent to the Sphinx2 recognizer. Sphinx2's most probable "transcription" is then sent to Why2-Atlas for syntactic, semantic and dialogue analysis. Finally, the text response produced by Why2-Atlas is converted to speech as described below, then played in the student's headphones and displayed on the interface. After the dialogue, the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision.

For this study, we used 3 ITSPOKE corpora, shown in Table 1.[1] The **SYN03** corpus was collected in 20**03** for an evaluation comparing learning in typed and spoken human and computer tutoring (Litman et al., 2006). ITSPOKE's voice was **syn**thesized with the Cepstral text-to-speech system, and its speech recognizer was trained from pilot ITSPOKE studies and Why2-Atlas evaluations. The **PR05** and **SYN05** corpora were collected in 20**05**, to evaluate the impact of tutor voice quality (Forbes-Riley et al., 2006). For these 2 corpora, ITSPOKE used an updated speech recognizer further trained on the **SYN03** corpus. For the **SYN05** corpus, ITSPOKE used the **syn**thesized tutor voice from the **SYN03** corpus; for the **PR05** corpus, ITSPOKE used a **pr**e-recorded tutor voice from a paid voice talent. Figure 1 gives an annotated (Section 3) **PR05** excerpt (**ASR** shows what ITSPOKE heard).

The same experimental procedure was used to collect all 3 ITSPOKE corpora: college students who had taken no college physics: 1) read a small document of background material, 2) took a pretest measuring initial physics knowledge, 3) work through a set of 5 problems (dialogues) with ITSPOKE, 4) took a posttest similar to the pretest.

[1] The user populations of the 2005 corpora and 2003 corpus are different, due to variation in year and recruitment method.

Subjects in the **PR05** and **SYN05** corpora also completed a survey probing user satisfaction after taking the posttest (**SYN03** corpus subjects did not). Our survey, shown in Figure 2, is essentially the same as the one used in the DARPA Communicator multi-site evaluation (Walker et al., 2002). Although tailored lexically for a tutoring system, these statements are generally applicable to spoken dialogue systems. Students rated their degree of agreement with each statement on a scale of 1 to 5.

PHYSICS PROBLEM: Suppose a lightweight car and a massive truck hit a patch of frictionless ice and have a head-on collision. Upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Defend your answers.

ESSAY UNDER DISCUSSION: The truck exerts an equal and opposite reaction force on the car. The car has a greater acceleration. The car will have a greater change in its motion.

. . . excerpt at 28.8 minutes into dialogue. . .

TUTOR$_{60}$: So, when you add forces together, besides the magnitudes of the forces, what must you also consider?

STUDENT$_{57}$: direct- direction (**ASR:** the truck in direction) *[correct, uncertain, non-frustrated/angry]*

TUTOR$_{61}$: Yes. [...] The vector sum of the forces acting on the body is also equal to the mass of the body multiplied by its acceleration. What is another term for the vector sum of forces acting on a body?

STUDENT$_{58}$: net force (**ASR:** net force) *[correct, neutral, non-frustrated/angry]*

TUTOR$_{62}$: That's super. Another term for the resultant force is the net force. In this relationship, what is the direction of the net force applied to a body and the direction of the body's acceleration?

STUDENT$_{59}$: same same same (**ASR:** seems same type seems) *[correct, certain, frustrated/angry]*

Figure 1: Annotated **PR05** Corpus Excerpt

Task Ease: It was easy to learn from the tutor.
TTS Performance: It was easy to understand the tutor.
User Expertise: I knew what I could say or do at each point in the conversations with the tutor.
Expected Behavior: The tutor worked the way I expected it to.
Future Use: Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.

ALMOST ALWAYS (5), OFTEN (4), SOMETIMES (3), RARELY (2), ALMOST NEVER (1)

Figure 2: ITSPOKE Survey Questionnaire

3 Interaction Parameters

3.1 Dialogue System-Generic Parameters

Prior PARADISE applications predicted user satisfaction using a wide range of system-generic parameters, which include measures of speech recognition quality (e.g. word error rate), measures of dialogue communication and efficiency (e.g. total turns and elapsed time), and measures of task completion (e.g. a binary representation of whether the task was completed) (Möller, 2005a; Möller, 2005b; Walker et al., 2002; Bonneau-Maynard et al., 2000; Walker et al., 2000; Walker et al., 1997). In this prior work, each dialogue between user and system represents a single "task" (e.g., booking airline travel), thus these measures are calculated on a per-dialogue basis.

In our work, the entire tutoring session represents a single "task", and every student in our corpora completed this task. Thus we extract 13 system-generic parameters on a per-student basis, i.e. over the 5 dialogues for each user, yielding a single parameter value for each student in our 3 corpora.

First, we extracted 9 parameters representing dialogue communication and efficiency. Of these parameters, 7 were used in prior PARADISE applications: Time on Task, Total ITSPOKE Turns and Words, Total User Turns and Words, Average ITSPOKE Words/Turn, and Average User Words/Turn. Our 2 additional "communication-related" (Möller, 2005a) parameters measure system-user interactivity, but were not used in prior work (to our knowledge): Ratio of User Words to ITSPOKE Words, Ratio of User Turns to ITSPOKE Turns.

Second, we extracted 4 parameters representing speech recognition quality, which have also been used in prior work: Word Error Rate, Concept Accuracy, Total Timeouts, Total Rejections[2].

3.2 Tutoring-Specific Parameters

Although prior PARADISE applications tend to use system-generic parameters, we hypothesize that task-specific parameters may also prove useful for predicting performance. We extract 12 *tutoring-specific* parameters over the 5 dialogues for each student, yielding a single parameter value per student, for each student in our 3 corpora. Although these parameters are specific to our tutoring system, similar parameters are available in other tutoring systems.

First, we hypothesize that the *correctness* of the students' turns with respect to the tutoring topic

[2]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 ASR output is too low.

(physics, in our case) may play a role in predicting system performance. Each of our student turns is automatically labeled with 1 of 3 "Correctness" labels by the ITSPOKE semantic understanding component: *Correct, Incorrect, Partially Correct*. Labeled examples are shown in Figure 1. From these 3 Correctness labels, we derive 9 parameters: a Total and a Percent for each label, and a Ratio of each label to every other label (e.g. Correct/Incorrect).

Second, students write and then may modify their physics essay at least once during each dialogue with ITSPOKE. We thus hypothesize that like "Correctness", the total number of essays per student may play a role in predicting system performance.

Finally, although student test scores before/after using ITSPOKE will be used as our student learning metric, we hypothesize that these scores may also play a role in predicting user satisfaction.

3.3 User Affect Parameters

We hypothesize that user affect plays a role in predicting user satisfaction and student learning. Although affect parameters have not been used in other PARADISE studies (to our knowledge), they are generic; for example, in various spoken dialogue systems, user affect has been annotated and automatically predicted from e.g., acoustic-prosodic and lexical features (Litman and Forbes-Riley, 2004b; Lee et al., 2002; Ang et al., 2002; Batliner et al., 2003).

As part of a larger investigation into emotion *adaptation*, we are manually annotating the student turns in our corpora for affective state. Currently, we are labeling 1 of 4 states of "Certainness": *certain, uncertain, neutral, mixed (certain and uncertain)*, and we are separately labeling 1 of 2 states of "Frustration/Anger": *frustrated/angry, non-frustrated/angry*. These affective states[3] were found in pilot studies to be most prevalent in our tutoring dialogues[4], and are also of interest in other dialogue research, e.g. tutoring (Bhatt et al., 2004; Moore et al., 2004; Pon-Barry et al., 2004) and spoken dialogue (Ang et al., 2002). Labeled examples are shown in Figure 1.[5] To date, one paid annotator has labeled all student turns in our **SYN03** corpus, and all the turns of 17 students in our **PR05** corpus.[6]

From these labels, we derived 25 **User Affect** parameters per student, over the 5 dialogues for that student. First, for each Certainness label, we computed a Total, a Percent, and a Ratio to each other label. We also computed a Total for each *sequence* of identical Certainness labels (e.g. Certain:Certain), hypothesizing that states maintained over multiple turns may have more impact on performance than single occurrences. Second, we computed the same parameters for each Frustration/Anger label.

4 Prediction Models

In this section, we first investigate the usefulness of our system-generic and tutoring-specific parameters for training models of user satisfaction and student learning in our tutoring corpora with the PARADISE framework. We use the SPSS statistical package with a stepwise multivariate linear regression procedure[7] to automatically determine parameter inclusion in the model. We then investigate how well these models generalize across different user-system configurations, by testing the models in different corpora and corpus subsets. Finally, we investigate whether generic user affect parameters increase the usefulness of our student learning models.

4.1 Prediction Models of User Satisfaction

Only subjects in the **PR05** and **SYN05** corpora completed a user survey (Table 1). Each student's responses were summed to yield a single user satisfaction total per student, ranging from 9 to 24 across corpora (the possible range is 5 to 25), with no difference between corpora ($p = .46$). This total was used as our user satisfaction metric, as in (Möller, 2005b; Walker et al., 2002; Walker et al., 2000).[8]

[3] We use "affect" and "affective state" loosely to cover student emotions and attitudes believed to be relevant for tutoring.

[4] For a full list of affective states identified in these pilot studies, see (Litman and Forbes-Riley, 2004a).

[5] Annotations were performed from both audio and transcription within a speech processing tool.

[6] In a preliminary agreement study, a second annotator labeled the entire **SYN03** corpus for *uncertain* versus *other*, yielding 90% inter-annotator agreement (0.68 Kappa).

[7] At each step, the parameter with the highest partial correlation with the target predicted variable, controlled for all previously entered parameters, is entered in the equation, until the remaining parameters do not increase R^2 by a significant amount or do not yield a significant model.

[8] Researchers have also used average score (Möller, 2005b; Walker et al., 1997); single survey statements can also be used (Walker et al., 1997). We tried these variations, and our R^2 results were similar, indicating robustness across variations.

Training Data	R^2	Predictors	Testing Data	R^2
PR05	.274	INCORRECTS, ESSAYS	SYN05	.001
SYN05	.068	TUT WDS/TRN	PR05	.018
PR05:half1	.335	PARTCORS/INCORS	PR05:half2	.137
PR05:half2	.443	STU TRNS	PR05:half1	.079
SYN05:half1	.455	STU TRNS/TUT TRNS	SYN05:half2	.051
SYN05:half2	.685	TUT WDS/TRN, STU WDS/TRN, CORRECTS	SYN05:half1	.227

Table 2: Testing the Predictive Power of User Satisfaction Models

We trained a user satisfaction model on each corpus, then tested it on the other corpus. In addition, we split each corpus in half randomly, then trained a user satisfaction model on each half, and tested it on the other half. We hypothesized that despite the decrease in the dataset size, models trained and tested in the same corpus would have higher generalizability than models trained on one corpus and tested on the other, due to the increased data homogeneity within each corpus, since each corpus used a different ITSPOKE version. As predictors, we used only the 13 system-generic and 12 tutoring-specific parameters that were available for all subjects.

Results are shown in Table 2. The first and fourth columns show the training and test data, respectively. The second and fifth columns show the user satisfaction variance accounted for by the trained model in the training and test data, respectively. The third column shows the parameters that were selected as predictors of user satisfaction in the trained model, ordered by degree of contribution[9].

For example, as shown in the first row, the model trained on the **PR05** corpus uses Total Incorrect student turns as the strongest predictor of user satisfaction, followed by Total Essays; these parameters are not highly correlated[10]. This model accounts for 27.4% of the user satisfaction variance in the **PR05** corpus. When tested on the **SYN05** corpus, it accounts for 0.1% of the user satisfaction variance.

The low R^2 values for both training and testing in the first two rows show that neither corpus yields a very powerful model of user satisfaction even in the training corpus, and this model does not generalize very well to the test corpus. As hypothesized, training and testing in a single corpus yields higher R^2 values for testing, as shown in the last four rows, although these models still account for less than a quarter of the variance in the test data. The increased R^2 values for training here may indicate over-fitting. Across all 6 experiments, there is almost no overlap of parameters used to predict user satisfaction.

Overall, these results show that this method of developing an ITSPOKE user satisfaction model is very sensitive to changes in training data; this was also found in other PARADISE applications (Möller, 2005b; Walker et al., 2000). Some applications have also reported similarly low R^2 values for *testing* both within a corpus (Möller, 2005b) and also when a model trained on one system corpus is tested on another system corpus (Walker et al., 2000). However, most PARADISE applications have yielded higher R^2 values than ours for *training* (Möller, 2005b; Walker et al., 2002; Bonneau-Maynard et al., 2000; Walker et al., 2000).

We hypothesize two reasons for why our experiments did not yield more useful user satisfaction models. First, in prior PARADISE applications, users completed a survey after *every* dialogue with the system. In our case, subjects completed only one survey, at the end of the experiment (5 dialogues). It may be that this "per-student" unit for user satisfaction is too large to yield a very powerful model; i.e., this measure is not fine-grained enough. In addition, tutoring systems are not designed to maximize user satisfaction, but rather, their design goal is to maximize student learning. Moreover, prior tutoring studies have shown that certain features correlated with student learning do not have the same relationship to user satisfaction (e.g. are not predictive

[9]The ordering reflects the standardized coefficients (beta weights), which are computed in SPSS based on scaling of the input parameters, to enable an assessment of the predictive power of each parameter relative to the others in a model.

[10]Hereafter, predictors in a model are not highly correlated ($R \geq .70$) unless noted. Linear regression does not assume that predictors are independent, only that they are not highly correlated. Because correlations above R =.70 can affect the coefficients, deletion of redundant predictors may be advisable.

Training Data	R^2	Predictors	Testing Data	R^2
PR05	.556	PRE, %CORRECT	SYN05	.636
SYN05	.736	PRE, INCORS/CORS, STU WDS/TRN	PR05	.472
PR05:half1	.840	PRE, PARTCORRECTS	PR05:half2	.128
PR05:half2	.575	PARTCORS/INCORS, PRE	PR05:half1	.485
SYN05:half1	.580	PRE, STU WDS/TRN	SYN05:half2	.556
SYN05:half2	.855	PRE, TIMEOUTS	SYN05:half1	.384
PR05+SYN03	.413	PRE, TIME	SYN05	.586
PR05+SYN05	.621	PRE, INCORS/CORS	SYN03	.237
SYN05+SYN03	.590	INCORS/CORS, %INCORRECT, PRE, TIME	PR05	.244

Table 3: Testing the Predictive Power of Student Learning Models with the Same Datasets

or have an opposite relationship) (Pon-Barry et al., 2004). In fact, it may be that user satisfaction is not a metric of primary relevance in our application.

4.2 Prediction Models of Student Learning

As in other tutoring research, e.g. (Chi et al., 2001; Litman et al., 2006), we use posttest score (POST) controlled for pretest score (PRE) as our target student learning prediction metric, such that POST is our target variable and PRE is always a parameter in the final model, although it is not necessarily the strongest predictor.[11] In this way, we measure student learning *gains*, not just final test score.

As shown in Table 1, all subjects in our 3 corpora took the pretest and posttest. However, in order to compare our student learning models with our user satisfaction models, our first experiments predicting student learning used the same training and testing datasets that were used to predict user satisfaction in Section 4.1 (i.e. we ran the same experiments except we predicted POST controlled for PRE instead of user satisfaction). Results are shown in the first 6 rows of Table 3.

As shown, these 6 models all account for more than 50% of the POST variance in the training data. Furthermore, most of them account for close to, or more than, 50% of the POST variance in the test data. Although again we hypothesized that training and testing in one corpus would yield higher R^2 values for testing, this is not consistently the case; two of these models had the highest R^2 values for training and the lowest R^2 values for testing (**PR05:half1** and **SYN05:half2**), suggesting over-fitting.

Overall, these results show that this is an effective method of developing a prediction model of student learning for ITSPOKE, and is less sensitive to changes in training data than it was for user satisfaction. Moreover, there is more overlap in these 6 models of parameters that are useful for predicting student learning (besides PRE); "Correctness" parameters and dialogue communication and efficiency parameters appear to be most useful overall.

Our next 3 experiments investigated how our student learning models are impacted by including our third **SYN03** corpus. Using the same 25 parameters, we trained a learning model on each set of two combined corpora, then tested it on the other corpus. Results are shown in the last 3 rows of Table 3.

As shown, these models still account for close to, or more than, 50% of the student learning variance in the training data.[12] The model trained on **PR05+SYN03** accounts for the most student learning variance in the test data, showing that the training data that is most similar to the test data will yield the highest generalizability. That is, the combined **PR05+SYN03** corpora contains subjects drawn from the same subject pool (2005) as the **SYN05** test data, and also contains subjects who interacted with the same tutor voice (synthesized) as this test data. In contrast, the combined **PR05+SYN05** corpora did not overlap in user population with the **SYN03** test data, and the combined **SYN05+SYN03** corpora did not share a tutor voice with the **PR05** test data. "Correctness" parameters

[11] In SPSS, we regress two independent variable blocks. The first block contains PRE, which is regressed with POST using the "enter" method, forcing inclusion of PRE in the final model. The second block contains all remaining independent variables, which are regressed using the stepwise method.

[12] However, INCORS/CORS and %INCORRECT are highly correlated in the **SYN05+SYN03** model, showing redundancy.

Training Data	R^2	Predictors	Testing Data	R^2
SYN03 (affect)	.644	TIME, PRE, NEUTRAL	PR05:17	.411
PR05:17 (affect)	.835	PRE, NFA:NFA, STU WDS/TRN	SYN03	.127
SYN03	.478	PRE, TIME	PR05:17	.340
PR05:17	.609	PRE, STU TRNS/TUT TRNS	SYN03	.164

Table 4: Testing the Predictive Power of Student Learning Models with User Affect Parameters

and dialogue communication and efficiency parameters are consistently used as predictors in all 9 of these student learning models.

4.3 Adding User Affect Parameters

Our final experiments investigated whether our 25 user affect parameters impacted the usefulness of the student learning models. As shown in Table 1, all 20 subjects in our **SYN03** corpus were annotated for user affect, and 17 subjects in our **PR05** corpus were annotated for user affect. We trained a model of student learning on each of these datasets, then tested it on the other dataset.[13] As predictors, we included our 25 user affect parameters along with the 13 system-generic and 12 tutoring-specific interaction parameters. These results are shown in the first two rows of Table 4. We also reran these experiments without user affect parameters, to gauge the impact of the user affect parameters. These results are shown in the last two rows of Table 4. We hypothesized that user affect parameters would produce more useful models, because prior tutoring research has shown correlations between user affect and student learning (e.g. (Craig et al., 2004)).

As shown in the first two rows, user affect predictors appear in both models where these parameters were included. The models trained on **SYN03** use pretest score and Total Time on Task as predictors; when affect parameters are included, "Neutral Certainness" is added as a predictor, which increases the R^2 values for both training and testing. However, the two models trained on **PR05:17** show no predictor overlap (besides PRE). Moreover, the **PR05:17** model that includes an affect predictor (Total Sequence of 2 Non-Frustrated/Angry turns) has the highest training R^2, but the lowest testing R^2 value.

[13] As only 17 subjects have both user affect annotation and user surveys, there is not enough data currently to train and test a user satisfaction model including user affect parameters.

5 Conclusions and Current Directions

Prior work in the tutoring community has focused on correlations of single features with learning; our results suggest that PARADISE is an effective method of extending these analyses. For the dialogue community, our results suggest that as spoken dialogue systems move into new applications not optimized for user satisfaction, such as tutoring systems, other measures of performance may be more relevant, and generic user affect parameters may be useful.

Our experiments used many of the same system-generic parameters as prior studies, and some of these parameters predicted user satisfaction both in our models and in prior studies' models (e.g., system words/turn (Walker et al., 2002)). Nonetheless, overall our user satisfaction models were not very powerful even for training, were sensitive to training data changes, showed little predictor overlap, and did not generalize well to test data. Our user satisfaction metric may not be fine-grained enough; in other PARADISE studies, users took a survey after *every* dialogue with the system. In addition, tutoring systems are not designed to maximize user satisfaction; their goal is to maximize student learning.

Our student learning models were much more powerful and less sensitive to changes in training data. Our best models explained over 50% of the student learning variance for training and testing, and both student "Correctness" parameters and dialogue communication and efficiency parameters were often useful predictors. User affect parameters further improved the predictive power of one student learning model for both training and testing.

Once our user affect annotations are complete, we can further investigate their use to predict student learning and user satisfaction. Unlike our other parameters, these annotations are not currently available, although they can be predicted automatically (Litman and Forbes-Riley, 2004b), in our sys-

tem. However, as in (Batliner et al., 2003), our prior work suggests that linguistic features reflective of affective states can replace affect annotation (Forbes-Riley and Litman, 2005). In future work we will use such features in our prediction models. Finally, we are also annotating tutor and student dialogue acts and automating the tutor act annotations; when complete we can investigate their usefulness in our prediction models; dialogue acts have also been used in prior PARADISE applications (Möller, 2005a).

Acknowledgements

NSF (0325034 & 0328431) supports this research. We thank Pam Jordan and the NLP Group.

References

J. Ang, R. Dhillon, A. Krupski, E.Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. Int. Conf. Spoken Language Processing (ICSLP)*.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40:117–143.

K. Bhatt, M. Evens, and S. Argamon. 2004. Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In *Proc. 26th Annual Meeting of the Cognitive Science Society*.

H. Bonneau-Maynard, L. Devillers, and S. Rosset. 2000. Predictive performance of dialog systems. In *Proc. Language Resources and Evaluation Conf. (LREC))*.

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

S. Craig, A. Graesser, J. Sullins, and B. Gholson. 2004. Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29:241–250.

K. Forbes-Riley and D. Litman. 2005. Correlating student acoustic-prosodic profiles with student learning in spoken tutoring dialogues. In *Proc. INTERSPEECH*.

K. Forbes-Riley, D. Litman, S. Silliman, and J. Tetreault. 2006. Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. In *Proc. FLAIRS*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2002. Combining acoustic and language information for emotion recognition. In *Proc. ICSLP*.

D. Litman and K. Forbes-Riley. 2004a. Annotating student emotional states in spoken tutoring dialogues. In *Proc. SIGdial*, pages 144–153.

D. Litman and K. Forbes-Riley. 2004b. Predicting student emotions in computer-human tutoring dialogues. In *Proc. ACL*, pages 352–359.

D. Litman, C. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *Intnl Jnl of Artificial Intelligence in Education, To Appear*.

S. Möller. 2005a. Parameters for quantifying the interactioin with spoken dialogue telephone services. In *Proc. SIGdial*.

S. Möller. 2005b. Towards generic quality prediction models for spoken dialogue systems - a case study. In *Proc. INTERSPEECH*.

J. D. Moore, K. Porayska-Pomsta, S. Varges, and C. Zinn. 2004. Generating tutorial feedback with affect. In *Proc. FLAIRS*.

J. Mostow and G. Aist. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus and P. Feltovich, editors, *Smart Machines in Education*.

H. Pon-Barry, B. Clark, E. Owen Bratt, K. Schultz, and S. Peters. 2004. Evaluating the effectiveness of SCoT: a Spoken Conversational Tutor. In *Proc. of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions*.

E. Shriberg, E. Wade, and P. Price. 1992. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proc. DARPA Speech and NL Workshop*, pages 49–54.

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. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems*.

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

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6:363–377.

M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard. 2002. DARPA communicator: Cross-system results for the 2001 evaluation. In *Proc. Internat. Conf. on Spoken Language Processing (ICSLP)*.