# Comparing Real-Real, Simulated-Simulated, and Simulated-Real Spoken Dialogue Corpora

**Hua Ai and Diane Litman**

University of Pittsburgh, Intelligent Systems Program
University of Pittsburgh, 5113 Sennott Square, 210 S. Bouquet St., Pittsburgh, PA 15260, USA.

**Abstract**

User simulation is used to generate large corpora for using reinforcement learning to automatically learn the best policy for spoken dialogue systems. Although this approach is becoming increasingly popular, the differences between simulated and real corpora are not well studied. We build two simulation models to interact with an intelligent tutoring system. Both models are trained on two different real corpora separately. We use several evaluation measures proposed in previous research to compare between our two simulated corpora, between the original two real corpora, and between the simulated and real corpora. We next examine the differentiating power of these measures. Our results show that although these simple statistical measures can distinguish real corpora from simulated ones, these measures cannot help us to draw a conclusion on the "reality" of the simulated corpora since even two real corpora can be very different when evaluated on the same measures.

## Introduction

Recently, Reinforcement Learning has been applied to many spoken dialogue systems to learn the best policy automatically (Henderson et al., 2005, Frampton and Lemon, 2005, English and Heeman, 2005, Singh et al., 2002, Walker, 2000). Because it is time consuming and costly to obtain sufficient data by running human subject experiments, computer simulated users have been built to interact with dialogue systems to generate larger corpora (Levin et al., 2000, Scheffler, 2002). These studies have demonstrated that simulation models of real user behaviors can be successfully trained from small corpora, but they do not evaluate the synthetic corpora systematically. Hence, it is hard to estimate how realistic the simulated corpora are or how useful the user simulations are. (Schatzmann et al., 2005) propose a set of quantitative evaluation measures to assess the quality of the simulated corpora. Although they show that these simple statistical measures are sufficient to discern simulated from real dialogues, they do not further examine the causes of the differences. In this paper, we expand their evaluation measures to investigate the potential differences between two real corpora which are generated during the interaction with the same tutoring dialogue system by two different groups of students. Our results show that two real corpora can be very different when evaluated by the same set of measures. Therefore, the differences shown by these measures are not necessarily related to whether the dialogue behaviors represented in the corpora are realistic or not. Instead, the differences could be caused by two different groups of user population. We further investigate the differentiating power of these evaluation measures by looking into the information suggested by them on two straightforward simulation models. Our results suggest that the currently used evaluation measures do not provide enough information to show why two corpora are different.



## Background

Recent simulated user models mostly work on the dialogue act level. Instead of trying to simulate fully natural utterances, these models simply generate the dialogue act of the student's next action, which in many cases is sufficient to continue the interaction with the dialogue systems. (Eckert et al., 1997) first suggest a bigram model to predict the next user's action based on the previous system's action. While this model is simple and domain-independent, it sometimes generates actions that do not make sense in the local context. (Levin et al., 2000) add constraints to the bigram model to only accept the expected dialogue acts. However, their basic assumption of making the next user's action dependent only on the system's previous action is oversimplified. (Scheffler, 2002) introduces fixed goal structures to hard-code all the possible paths of users' actions into a network. He trains the parameters of the network from training data for further prediction. (Pietquin, 2004) explicitly models the dependencies between a user's actions and his/her goal by conditioning the probabilities used by Levin et al. on a representation of the user goal. Both Scheffler's and Pietquin's work involve lots of manual work and may become infeasible when there are a large numbers of user actions. As a large number of states can make learning intractable, (Georgila et al., 2005) try to overcome this problem by exploiting commonalities between different states. They use linear combinations of shared features to express the commonalities. While most of the research is on intentional-level, (Chung, 2004) uses word-level user simulation to improve dialogue development as well as to train speech recognizer and understanding components.

To date, there are no generally accepted evaluation methodologies to assess how "realistic" a simulated corpus is. Dialogue length, goal achievement rate and goal completion length have been used in previous research (Scheffler and Young, 2001). A comprehensive set of quantitative evaluation measures is proposed by (Schatzmann et al., 2005). They consider three groups of measures that can cover the statistical properties of dialogues. The first group investigates high level dialogue features. These measures look into both how much information is transmitted in the dialogue and how active the dialogue participants are. The second group of measures analyzes the style of the dialogue in terms of the frequency of different speech acts, the proportion of goal-directed and social dialogue, and the user's degree of cooperativeness. The last group of measures examines the efficiency of the dialogues using goal achievement rates and goal completion times. As Schatzmann et al. point out, these measures are only introduced to cover a variety of dialogue properties for comparing simulated dialogues against real ones but there is no specific range of values to qualify a synthetic corpus to be sufficiently realistic.

## Corpus

Our data consists of dialogues between students and ITSPOKE (Litman and Silliman, 2004), a spoken dialogue tutor built on top of the Why2-Atlas conceptual physics text-based tutoring system (VanLehn, 2002). In ITSPOKE, a student first types an essay answering a qualitative physics question. A tutoring dialogue is initiated by ITSPOKE after analyzing the essay to correct misconceptions and to elicit further explanations. After that, the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. These tutoring dialogues are manually authored to define the system's response to correct/incorrect/partially correct answers.

In our study, we use three corpora of tutoring dialogues collected with two different groups of subjects. The two groups were recruited on the University of Pittsburgh campus in fall 2003 and spring 2005 separately. The 2003 group consists of students recruited via posted flyers only, whereas the students in the 2005 group were recruited via flyers as well as via the instructor of a large introduction to psychology class. Subjects have never taken college-level physics. They first read a small document of background physics material, then work through 5 problems with ITSPOKE. A pre-test is given before the tutoring session and a post-test is given afterwards. The student's normalized learning gain is computed using the following formula: NLG = (postTestScore-preTestScore)/(1-preTestScore).

The main components of ITSPOKE remain the same in the 2005 experiment as in 2003, but a slightly different language model is used and some bugs are fixed as well. Also, the system uses a synthesized voice in all 2003 experiments, though in the 2005 experiments, one half of the experiments use synthesized voice and the other half use pre-recorded voice. Table 1 shows an overview of the collected corpora.

| **Corpus** | | **Student population** | **System difference** | **Number of dialogues** |
|---|---|---|---|---|
| f03 | | 2003 | Synthesized voice | 100 |
| s05 | syn | 2005 | Synthesized voice | 136 |
| | pre | 2005 | Pre-recorded voice | 135 |

Table 1: Overview of collected corpora

## Experiments

Our goal is to study the evaluation measures proposed in (Schatzmann et al., 2005) in terms of the information suggested by these measures on simple user models. For example, if the evaluation measures predict that a model which generates random utterances is very human-like, we then know that these evaluation measures are not powerful enough. We build two simple simulation models, the ProbCorrect model and the Random model. The ProbCorrect model simulates the student who generates utterances based on the dialogue context and who has an average knowledge level that can answer tutor's questions with the same probability as the average real students. On the contrary, the Random model simulates the student who gives out random answers that do not necessarily make sense in the dialogue. Relating to previous research, the ProbCorrect model is similar to the Bigram model while the Random model is similar to the Unigram model. In this section, we describe these two models and our simulation experiments. In the next section, we show our results.

### Simulation Models

We choose to generate student's utterances on the word level since generating student's dialogue acts alone does not provide sufficient information for our tutoring system to decide the next system's action. Because it is hard to generate a natural language utterance corresponding to each tutor's question, we use the answer sets from the real corpus as the candidate answer sets for the simulated students. The answer sets are extracted from the logs of human subject experiments thus the answers include misrecognized utterances. As we are not simulating student's essays at this stage, we restrict the simulated tutoring session to begin with a blank essay and stop after the first time the tutor asks for an essay revision. Actually in our real corpora, more than half of the students' essays are treated as blank essays due to the system's inability to understand them or due to the essays' low quality. In order to perform a fair comparison between the real and the simulated corpora, in the next section we only compare the real dialogues that start with "blank" essays with the simulated dialogues and we only look at the first part of these dialogues before the first essay revision.

| | |
|---|---|
| Tutor said: | Please type in your essay. |
| You typed | *(Blank essay)* |
| Tutor said: | To analyze this problem, which law of motion would you use? |
| You said: | *The second law.* |
| Tutor said: | The best law of motion to use in this case is Newton's third law. Do you recall what this law says? |
| You said: | *No.* |
| Tutor said: | Newton's third law says ... Is there a gravitational force on the sun due to the earth? |
| You said: | *Yes.* |
| Tutor said: | Good. Now, what is the relationship between the two forces' magnitudes? |
| You said: | *They're equal.* |
| Tutor said: | Yeah. What is the relationship between the two forces' directions? |
| You said: | *They're different.* |
| Tutor said: | To summarize, ... We've just walked through how to solve this problem … |

Figure 1: Simulated dialogue by the ProbCorrect Model

**ProbCorrect Model.** The ProbCorrect model is designed to give a correct/incorrect answer with the same probability as the average of real students. For each tutor's question, we automatically compute the average correctness rate of real student answers from the system logs. Then, a correct/incorrect answer is randomly chosen from the correct/incorrect answer sets for this question. We implement a back-off mechanism to count possible answers that do not appear in the real corpus. An excerpt of the simulated corpus by the ProbCorrect model is given in Figure 1.

**Random Model.** The Random model, as a simple unigram model, randomly picks a student's utterance from the real corpus as the answer to a tutor's question, neglecting which question it is. An excerpt of the simulated corpus by the Random model is given in Figure 2.

## Simulation

We train the Random model and the ProbCorrect model on the real corpora f03 and s05 separately to generate four simulated corpora: 03Random, 03ProbCorrect, 05Random, 05ProbCorrect. We generate 30[1] simulated students with each model. Each student goes through 5 problems (dialogues). In total, there are 150 dialogues generated with each model.

[1] This provides us with simulated corpora of comparable size to those used in Schatzmann et al.'s work

| | |
|---|---|
| Tutor said: | Please type in your essay. |
| You typed | *(Blank essay)* |
| Tutor said: | To analyze this problem, which law of motion would you use? |
| You said: | *More.* |
| Tutor said: | The best law of motion to use in this case is Newton's third law. Do you recall what this law says? |
| You said: | *After all forces.* |
| Tutor said: | Newton's third law says ... Is there a gravitational force on the sun due to the earth? |
| You said: | *Down.* |
| Tutor said: | We have just seen that Newton's third law says... Now, what is the relationship between the two forces' magnitudes? |
| You said: | *Yes.* |
| Tutor said: | Well… According to Newton's third law… What is the relationship between the two forces' directions? |
| You said: | *It's decreasing.* |
| Tutor said: | To summarize, ... We've just walked through how to solve this problem … |

Figure 2: Simulated dialogue by the Random Model

# Results

## Evaluation Measures

Our evaluation consists of two parts. We adapt the evaluation measures proposed in (Schatzmann et al., 2005) to assess the reality level (described below) of the generated dialogue behaviors. We also add several measures to estimate student learning gain, which is an important property in tutoring systems. In this paper, we use reality level to describe how human-like the behaviors are. We assume that all real corpora have the same reality level. A simulated corpus has lower reality level when it is less human-like. In our case, we expect the ranking of reality level to be (from high to low): the real corpora, corpora simulated by the ProbCorrect model, and corpora simulated by the Random model.

Table 2 shows the changes we make to apply the previous measures to our domain. For example, in the first column of high-level dialogue features, Schatzmann et al. use number of turns as dialogue length; we are also using the number of turns but we split it into the number of student turns (S_turn) and the number of tutor turns (T_turn).

When we compare the real corpora, we apply all the measures shown in Table 2. As we are not simulating student's essays at this stage, the simulated corpora only include dialogues of discussions before the first time the tutor asks for an essay revision. As a result, dialogue style features and dialogue success rate features are not available in simulated corpora. We are not simulating student's learning gains either; correctRate is the only learning feature used.

| Schatzmann et al. | Our measures | Abbreviation |
|---|---|---|
| **High-level dialogue features** | | |
| Dialogue length (number of turns) | Number of student/tutor turns | S_turn, T_turn |
| Turn length (number of actions per turn) | Total words per student/tutor turn | S_wordRate, T_wordRate |
| Participant activity (ratio of system and user actions per dialog) | Ratio of system and user words per dialog | WordRatio |
| **Dialogue style and cooperativeness** | | |
| Proportion of goal-directed actions vs. others | Proportion of goal-directed turns vs. others | Phy/non * |
| Number of times a piece of information is re-asked | Number of times a physics concept is re-discussed | repeatConcept* |
| **Dialogue Success Rate and Efficiency** | | |
| Average goal/subgoal achievement rate | Average number of essay submissions | essayRevision * |
| **Learning features** | | |
| None | Percentage of correct answers | correctRate |
| None | Learning gains | Learning * |

Table 2: Mapping between evaluation measures
* the feature is not available in the simulated corpora

## Comparisons between real corpora

We first compare the real corpora. Figure 3 illustrates the mean values of each evaluation measure for each corpus. The error bars show standard deviations of the mean values. Since our system changed slightly in 2005 from 2003, it is not appropriate to compare the corpora collected in different years directly[2] using a strict statistic test (such as t-test). Instead, we choose to show the differences graphically. In the graph, x-axis shows the evaluation measures, y-axis shows the mean for each corpus normalized to the f03 mean. For instance, when comparing S_wordRate, the mean of S_wordRate for f03 is scaled to "1", and the means for the 2005 corpus are normalized accordingly. We can tell how different two corpora are from the overlapping between the error bars. The less overlapping are the error bars, the greater is the difference between the two corpora. If the error bars do not overlap at all, the differences are likely to be significant if tested in a statistical test (we call these clear differences here).

We can see by studying the first eight groups of bars that the two corpora syn and pre from 2005 are very different from f03 using most of the measures proposed by Schatzmann et al. There are no clear differences between syn and pre, which suggests that the only major changes in the system from 2003 to 2005, the type of the system's voice, does not cause systematical differences. Given that the system is almost the same and there are no differences in the reality level of the two, the clear differences between the 2003 and 2005 corpora are most likely due to the different population of subjects. In other words, the differences caught by the above measures may be due to the different subject populations represented in the corpora, instead of the differences in the reality level. As a result, the differences shown by using the above measures are not sufficient to support the conclusion on the reality level of a simulated corpus. If the real corpus used to train the simulation model represents the entire population, also a successful simulated corpus represents the same population; the differences shown by the above measures might then be interpreted as the differences in reality level. However, if the real data is skewed and only represent a small part of the entire population while a successful simulation represents the entire population, these two corpora might be shown to be very different using the above measures. Nevertheless, these differences do not indicate that the simulated corpus has a low reality level.

Interestingly, we do not see clear differences between any two of the three corpora using learning gain features (shown by the last two groups of bars). This could be a positive sign that different groups of students do learn from the interaction with the tutoring system. However, as human learning is a complex behavior which may not be fully described with learning gains alone, we need to verify this in the future work.

To briefly summarize our observation in the real corpora comparisons, we find that using the evaluation measures shown in Figure 3 we can distinguish real corpora from different populations.

## Comparisons between simulated and real corpora

Now we compare the subsets of the real corpora f03 and s05 with the simulated corpora generated from them. Since in the previous section we show no clear difference between syn and pre, we use s05 as a whole to train simulation models instead of using the two parts of the s05 corpus separately. The system used in the interaction with the simulation models is a slightly newer version than those used in the 2003 and 2005 human subject experiments, which uses the same dialogue manager but is changed a little bit in the way of handling the student's second mistake to the same question. Thus we do not apply statistical tests, but use the same type of graph to show the differences between corpora as we do in the previous real-real comparisons.

[2] This is a prevalent problem when comparing corpora generated by different systems (for example, the comparisons done by previous studies on data from the DARPA Communicator project).

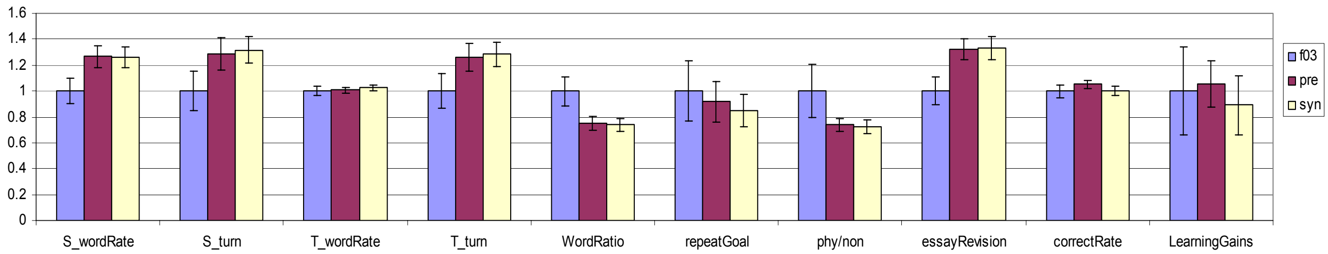


Figure 3: Comparisons between real corpora

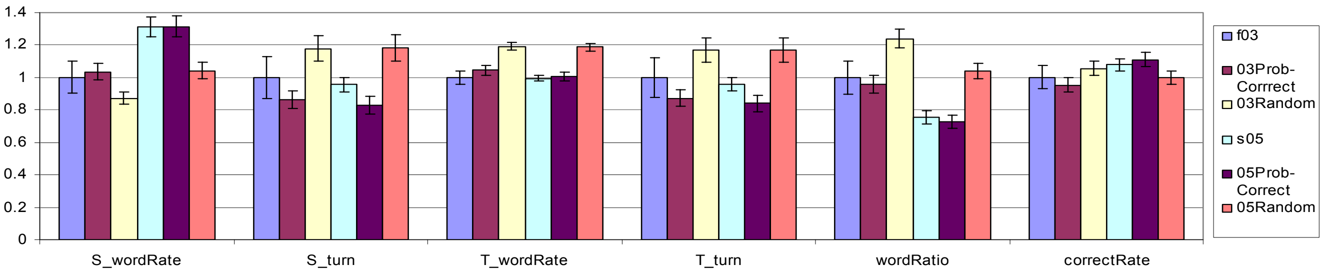


Figure 4: Comparisons between simulated and real corpora

These evaluation measures demonstrate that the ProbCorrect model can generate corpora more similar to the real ones than the Random model. In Figure 4, we observe that f03 is different from 03Random when looking at S_wordRate, S_turn, T_wordRate, T_turn, and wordRatio. The same results hold for s05 and 05Random. The difference between f03 and 03Random on correctRate is not clear. However, 05Random has lower correctRate than s05. When comparing 03ProbCorrect with f03 and 05ProbCorrect with s05 respectively, we notice that both of the simulated corpora are not clearly different from their training corpora in terms of T_wordRate, S_wordRate, wordRatio, and correctRate. Yet, there are quite big differences on T_turn and S_turn.

Similar to the work of Schatzmann et al., we find that the model which has more random-like behaviors has extremely long turn length while the model with more restricted behaviors produces more reasonable results. On the contrary, the Bigram model in the work of Schatzmann et al. finishes dialogue far too early while here the Random model produces longer dialogue than the other. This is due to the different properties of the dialogues. They are simulating users to perform in the travel booking domain. In those cases, the Bigram model simulates the users who are very uncooperative, causing the system to finish the dialogue earlier before completing any booking. In contrast, in our domain, the computer tutor is programmed to discuss every concept that the student does not know before finishing a tutoring session. As a result, the Random model generates students who need more discussion and have longer dialogues.

To sum up, these measures can show the expected ranking of the reality level of the corpora, which is (from high to low): the real corpora, corpora simulated by the ProbCorrect model, and the corpora simulated by the Random model. Although the ProbCorrect model can imitate reasonable and consistent student behaviors, it is still a simple model which does not take the learning behaviors or the student emotions into account. It is surprising to see that this model generates corpora with a high reality level. One reason might be that the evaluation measures we applied are not powerful enough to distinguish real corpora and highly-real simulated corpora. More measures are needed to investigate this issue.

## Comparisons between simulated corpora

Next, we compare the simulated corpora. As shown in Figure 4, most of the measures (except for correctRate) are able to catch the differences between the Random model and the ProbCorrect model clearly. Schatzmann et al. also find it is very easy to distinguish the corpus generated from the Bigram model and the one from Levin's model. Although these shallow measures are able to distinguish two simulated models, we cannot conclude on the differentiating power of these measures due to the very big differences between the two models.

As we expect, the model trained on one corpus is more similar to its training corpus than to the other real corpora. However, 03ProbCorrect and 05ProbCorrect perform similarly on half of the measures while performing quite differently on the other half of the measures. We thus cannot draw conclusions on whether the two corpora are different or not here. This may be because these measures are not powerful enough to grasp the differences between the same models which are trained from different corpora, or because the limited structure of tutoring dialogues constrains the models' performance.

## Conclusions & Future work

We use several evaluation measures proposed in previous research to carry out comparisons between real corpora, between simulated corpora, and between simulated and real corpora. While these measures can highlight some differences between corpora, our results indicate that these measures do not provide enough information to figure out why two corpora are different and to what extent the two corpora are different. We observe that two real corpora can be very different when measured by these measures, so the differences shown by these measures are not necessarily related to the reality level of the users. As a consequence, even if these measures demonstrate that a simulated corpus is different from a real corpus, we cannot conclude that the simulated corpus is not realistic enough.

We built two straightforward models to estimate the differentiating power of the evaluation measures. Our results show that these evaluation measures can, to some extent, distinguish real from simulated corpora, distinguish two simulated corpora generated by different models trained on the same real corpus, as well as distinguish two simulated corpora generated by the same model trained on two different real corpora. However, we do not have enough information to draw conclusions on the differentiating power of these measures.

Our conclusions are preliminary due to both that we are only using shallow word-level features and that we only simulate the first part of the real dialogue so the dialogue structure is very limited. In the future, we plan to improve our simulation models by simulating the whole tutoring session so that we can try some deep evaluation measures such as dialogue acts, dialogue styles, dialogue success rate, etc. We also want to simulate more characteristics of students to enrich the simulated corpus (for example, to add student emotions and learning features). As our goal is to use the simulated corpus for improving reinforcement learning (Tetreault and Litman, 2006), we would also like to test whether using simulated corpora in reinforcement learning will provide us with different policies than using real corpora.

## Acknowledgements

NSF (0325054) supports this research. The authors wish to thank Pam Jordan, Mihai Rotaru, and Joel Tetreault for their valuable suggestions, Scott Silliman for his support on building the simulation system, and the anonymous reviewers for their insightful comments.

## References

Chung, G. 2004. Developing a Flexible Spoken Dialog System Using Simulation. In *Proc. of ACL 04*.

Eckert, W.; Levin, E.; and Pieraccini, R. 1997. User Modeling for Spoken Dialogue System Evaluation. In *Proc. of IEEE workshop on ASRU*.

English, M., and Heeman, P. 2005. Learning Mixed Initiative Dialog Strategies By Using Reinforcement Learning On Both Conversants. In *Proc. Of HLT/EMNLP*.

Frampton, M., and Lemon, O. 2005. Reinforcement Learning of Dialogue Strategies Using The User's Last Dialogue Act. In *Proc. of IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 83-90.

Georgila, K.; Henderson, J.; and Lemon, O. 2005. Learning user simulations for information state update dialogue systems. In *Proc. of Eurospeech '05*.

Henderson, J.; Lemon, O.; and Georgila. K. 2005. Hybrid Reinforcement/Supervised Learning For Dialogue Policies From COMMUNICATOR Data. In *Proc. of IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 68-75.

Levin, E.; Pieraccini, R.; and Eckert, W. 2000. A Stochastic Model of Human-Machine Interaction For learning Dialogue Strategies. *IEEE Trans. On Speech and Audio Processing*, 8(1):11-23.

Litman, D., and Silliman, S. 2004. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In *Companion Proc. of the Human Language Technology: NAACL*.

Pietquin, O. 2004. A Framework for Unsupervised Learning of Dialogue Strategies. Ph.D. diss, Faculte Polytechnique de Mons.

Scheffler, K., and Young, S. J. 2001. Corpus-based Dialogue Simulation for Automatic Strategy Learning and Evaluation. In *Proc. of NAACL Workshop on Adaptation in Dialogue Systems*, 64-70.

Scheffler, K. 2002. Automatic Design of Spoken Dialogue Systems. Ph.D. diss., Cambridge University.

Schatzmann, J.; Georgila, K.; and Young, S. 2005. Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. In *Proc .of 6th SIGdial Workshop on Discourse and Dialogue*, 45-54.

Singh, S.; Litman, D.; Kearns, M.; and Walker, M. 2002. Optimizing Dialog Management with Reinforcement Learning: Experiments with the NJFun System. In *Journal of Artificial Inttelligence Research*, 16: 105-133.

Tetreault, J. and Litman, D. 2006. Using Reinforcement Learning to Build a Better Model of Dialogue State. In *Proc. of 11th Conference of the European Chapter of the Association for Computational Linguistics*.

VanLehn, K.; Jordan, P. W.; Ros´e, C. P.; Bhembe, D.; B¨ottner, M.; Gaydos, A.; Makatchev, M.; Pappuswamy, U.; Ringenberg, M.; Roque, A.; Siler, S.; Srivastava, R.; and Wilson, R. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems Conference*.

Walker, M. A. 2000. An Application of Reinforcement Learning to Dialog Strategy Selection in a Spoken Dialog System for Email. In *Journal of Artificial Intelligence Research*, 12:387-416.