Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development

Kate Forbes-Riley and Diane Litman

Learning Research and Development Ctr, Univ. of Pittsburgh, Pittsburgh, PA 15260

Abstract. We use a χ^2 analysis on our spoken dialogue tutoring corpus to investigate dependencies between uncertain student answers and 9 dialogue acts the human tutor uses in his response to these answers. Our results show significant dependencies between the tutor's use of some dialogue acts and the uncertainty expressed in the prior student answer, even after factoring out the answer's (in)correctness. Identification and analysis of these dependencies is part of our empirical approach to developing an adaptive version of our spoken dialogue tutoring system that responds to student affective states as well as to student correctness.

1 Introduction

Within research on spoken dialogue systems, promising results have been reported for automatically detecting user affect (e.g., [1–4]). The larger goal of such work is to improve dialogue system quality by automatically adapting to affect; however, to date not a lot of work has focused on the affect adaptations themselves. This difficult task involves developing appropriate responses, determining when to apply them, and evaluating the responses with real users. In some domains, it seems plausible to start with intuitively useful adaptations. For example, [5]'s health assessment system responds with empathy to user stress. [6]'s gaming system responds with sympathy and apology to user frustration. In both studies, users preferred the adaptive system over non-adaptive versions.

In contrast, in the tutoring system domain, where student learning is the primary metric of system performance, it is not clear a priori what system responses to student affective states will be most useful (for improving learning). We take an empirical approach to developing affect adaptations for our spoken dialogue tutoring system. Our approach is to develop system adaptations to student affective states based on analysis of our *human* tutor responses to those states. For this analysis, we use a previously collected human tutoring corpus that corresponds to our system corpora, except the tutor is human (Section 2.1).

We target student uncertainty as our first affective state for adaptation for two reasons. First, it occurs more often than other affective states in our tutoring dialogues [9]. Second, although most tutoring systems respond based only on student correctness, tutoring researchers are showing interest in also responding to student uncertainty, hypothesizing that uncertainty and incorrectness each create an opportunity for the student to engage in constructive learning [8, 10].

In this paper, we use the χ^2 test to investigate dependencies between uncertain student answers (Section 2.2) and 9 dialogue acts (Section 2.3) that our human tutor uses to respond to these answers. Our dialogue acts include Feedback Acts (Positive or Negative), Question Acts (Short and Hard Answer), and State Acts (Hints, Bottom Outs, Restatements, Recaps and Expansions).

Our results (Section 3) show significant dependencies between uncertain student answers and the human tutor's use of some dialogue acts in his response, even after factoring out the answer's (in)correctness. Within incorrect answers, the tutor's responses contain a Bottom Out significantly more than expected after uncertain answers. Within correct answers, the tutor's responses contain Positive Feedback significantly more than expected, and contain Expansions and Short Answer Questions significantly less than expected, after uncertain answers. This work builds on our and other prior work in this area (Section 4). Identification and analysis of these dependencies is part of our empirical approach to developing an adaptive version of our spoken dialogue tutoring system that responds to affective states[1] in student answers as well as their correctness.

2 Data and Annotations

2.1 Human Tutoring Spoken Dialogues

Our data consists of a human tutoring spoken dialogue corpus of 128 transcribed and annotated dialogues between 14 students and one human tutor. Each dialogue contains 47 student turns and 43 tutor turns on average. This corpus was collected in tandem with a computer tutoring corpus using our ITSPOKE spoken dialogue tutoring system; the human tutor and ITSPOKE performed the same task [11]. Each dialogue consists of a question-answer discussion between tutor and student about one qualitative physics problem. First, the student types an essay answer. The tutor analyzes the essay, then engages the student in a dialogue to correct misconceptions and incompleteness. The student then revises the essay, thereby ending the dialogue or causing another round of dialogue/essay revision. Figure 1 shows a dialogue excerpt from the human tutoring corpus.

The experimental procedure for collecting our corpora was as follows: each student 1) took a pre-test, 2) used a web and voice interface to work through 5 to 10 physics problems with the tutor, and 3) took a post-test. Before working the problems, students read a small document of background physics material[2].

[1] We use "affect" to cover emotions and attitudes. Some argue for separating the two, but some speech researchers find the narrow sense of "emotion" too restrictive since it excludes states in speech where emotion is not full-blown, including arousal and attitude [7]. Some tutoring researchers also combine emotion and attitude (e.g. [8]).

[2] In the human corpus, the pretest was given before the reading. In the computer corpus, it was was moved after the reading, to measure learning gains caused only by the tutoring. Also, in the computer corpus, students work only 5 problems.

TUTOR$_{37}$: Is there any other force on the- on these objects? *[SAQ]*
STUDENT$_{38}$: Uh no, just gravity. *[non-uncertain, correct]*
TUTOR$_{39}$: Now why do you say that? *[HAQ]*

< *...further discussion ...*>

TUTOR$_{53}$: Ok, they have been dropped in vacuum, so gravity is the only force acting on them. And uh, what is the motion called when gravity is the only force, uh, acting? *[RCP, SAQ]*
STUDENT$_{54}$: ... Freefall? *[uncertain, correct]*
TUTOR$_{55}$: Uh freefall, yes, so these two balls are in freefall. *[RST, POS, RST]*
STUDENT$_{56}$: Right *[non-uncertain, NA]*
TUTOR$_{57}$: What is the essential characteristic of objects that are in freefall? *[SAQ]*
STUDENT$_{58}$: Uh... Mass? *[uncertain, incorrect]*
TUTOR$_{59}$: Mass of course is an intrinsic property of the object. Uh, but if they are in freefall then there is something which is in common in for all objects that are in freefall? *[POS, RST, SAQ]*
STUDENT$_{60}$: Ok, uh, the weight. The gravitational force. *[non-uncertain, incorrect]*
TUTOR$_{61}$: Gravitational force we have already recognized is different for both of them. It's not the same. *[HINT]*
STUDENT$_{62}$: Oh, ok, um... no, I don't know. *[uncertain, incorrect]*
TUTOR$_{63}$: Well, all objects in freefall have the same acceleration. *[BOT]*

< *...further discussion ...*>

TUTOR$_{75}$: Now, there is a letter which is used for representing acceleration due to gravity. What is that letter? *[EXP, SAQ]*
STUDENT$_{76}$: Um, A? *[uncertain, incorrect]*
TUTOR$_{77}$: No, G-G in the lower case is, uh, used for representing acceleration due to gravity. *[NEG, BOT]*

Fig. 1. Annotated Human Tutoring Corpus Excerpt (at 5.2 min. into dialogue)

2.2 Student Uncertainty and Correctness Annotations

In our human tutoring corpus, each student answer has been manually labeled by a paid annotator for uncertainty [9]. Here we distinguish two labels[3]: the *uncertain* label is used for answers expressing uncertainty or confusion about the material being learned, and the *non-uncertain* label is used for all other answers. The same annotator also manually labeled each student answer for correctness, based on the human tutor's response to the answer [11]. Here we distinguish two labels[4]: the *correct* label is used for answers the tutor considered to be wholly or partly correct, and the *incorrect* label is used for answers the tutor considered to be wholly incorrect. Labeled examples are shown in Figure 1.

[3] A second annotator labeled a subset of the human tutoring corpus (505 student turns), yielding inter-annotator agreement of 0.61 Kappa for these two labels. See [9] for further discussion of this inter-annotation. Two annotators also labeled an ITSPOKE corpus with these labels, yielding inter-annotator agreement of 0.73 Kappa.

[4] A second annotator also labeled a subset of the corpus (507 student turns) with these labels, yielding inter-annotator agreement of 0.85 Kappa.

Note that although student uncertainty and (in)correctness are related, they cannot be equated. First, prior work has shown that an uncertain answer may be correct or incorrect [8]. This is also true in our data. As discussed in Section 3.2, our dataset consists of 1985 student answers. Table 1 shows the distribution of our uncertainty and (in)correctness labels across these answers. For example, in our data 412 uncertain answers are correct, while 341 uncertain answers are incorrect. However, when we apply the χ^2 test to this data (as discussed in Section 3.2), we find a highly significant positive dependency between uncertainty and (in)correctness: uncertain answers are incorrect significantly more than expected by chance alone - or equivalently, uncertain answers are correct significantly less than expected by chance (χ^2 value = 78.47 (df=1)).

	correct	incorrect	total
uncertain	412	341	753
non-uncertain	912	320	1232
total	1324	661	1985

Table 1. Distribution of Student Answers in terms of Uncertainty and (In)Correctness

Uncertainty and incorrectness also differ in terms of what they convey to the tutor. Incorrectness conveys that there is a misconception in the student's knowledge. Uncertainty - in both a correct and a incorrect answer - conveys that the student *perceives a possible misconception* in their knowledge. If that answer is correct, a misconception does not actually exist; if that answer is incorrect, a misconception does exist. Our analyses of the dependencies presented in Section 3 suggest that our human tutor responds both to an answer's (in)correctness and to the student's perceived misconception; our system responses to uncertainty over correctness can thus be modeled based on these human tutor responses.

2.3 Tutor Dialogue Act Annotations

In our human tutoring corpus, each utterance in each tutor turn has been manually labeled by a paid annotator for tutoring dialogue acts [11][5]. Our annotation is based on similar schemes from other tutorial dialogue projects (e.g. [12]). Here we distinguish 9 labels, which are defined in Figure 2 and illustrated in Figure 1. Note that some definitions relate to the prior student answer's correctness. "Feedback Acts" label feedback based on lexical items in the tutor turn. These labels often coincide with the prior student answer's correctness, but can also convey encouragement, or relate to the discourse level or to the student's earlier essay. "State Acts" summarize or clarify the current state of the student's argument, based on the prior student turns(s). "Question Acts" label the type of question asked, in terms of its content and the type of answer required.

[5] A second annotator labeled these dialogue acts in a subset of our corpus (8 dialogues containing 548 utterances), yielding agreement of 0.48 Kappa.

- **Tutor Feedback Acts**
  - Positive Feedback **(POS)**: positive feedback word/phrase present in turn
  - Negative Feedback **(NEG)**: negative feedback word/phrase present in turn
- **Tutor State Acts**
  - Restatement **(RST)**: repetitions and rewordings of prior student statement
  - Recap **(RCP)**: summarize overall argument or earlier-established points
  - Bottom Out **(BOT)**: full answer given if student answer is incorrect
  - Hint **(HINT)**: partial answer given if student answer is incorrect
  - Expansion **(EXP)**: novel details related to answer given without being queried
- **Tutor Question Acts**
  - Short Answer Question **(SAQ)**: concerns basic quantitative relationships
  - Hard Answer Question **(HAQ)**: requires definition/interpretation of concepts or reasoning about causes and/or effects

Fig. 2. Tutor Dialogue Acts

3 Student Uncertainty-Tutor Response Dependencies

3.1 Extracting Student Answers and Tutor Responses

Here we are investigating dependencies between uncertain student answers and 9 dialogue acts that our human tutor uses to respond to these answers. However, while our computer tutoring dialogues follow a strict Tutor Question-Student Answer-Tutor Response format, our human tutoring dialogues are more complex. In particular, not all student turns contain an answer to a tutor question; instead (see Figure 1, STUDENT_{56}), they may contain a backchanneling or grounding, or a clarification question, or they may be related to the situation rather than the physics content (e.g. "How do I submit my essay?"). All student turns that do not contain an answer are labeled "NA" in our correctness annotation scheme. In this study we exclude these non-answer student turns from our analysis, because they don't exist in our computer tutoring corpus. In other words, we only investigate dependencies between actual student answers (labeled correct or incorrect) and tutor responses to them. We do this by extracting from our human tutoring corpus all bigrams consisting of a student answer turn followed by a tutor response turn, yielding 1985 student answer-tutor response bigrams.

3.2 The χ^2 Analysis

We use a χ^2 analysis to investigate dependencies between uncertain student answers and each of the 9 dialogue acts that may be present in our human tutor's responses to these answers. We investigate each dialogue act separately, because most tutor turns contain multiple dialogue acts (see Figure 1) and there are no limits on their combination. Thus, treating each tag combination as a unique response would yield a data sparsity problem for our dependency analysis.

We performed 9 dependency analyses, one for each dialogue act D. For each analysis, we first took our dataset of student answer-tutor response bigrams and replaced all tutor responses containing D with only **D**, and replaced all tutor responses not containing D with only **notD**.[6] Next, we applied four χ^2 tests to this dataset. To illustrate the analysis, we refer to Table 2, which shows the results of the analysis of the BOT ("Bottom Out") dialogue act response.

The first χ^2 test investigates the dependency between uncertain student answers and the tutor's use of D in his responses. We compute a χ^2 value for the dependency between a binary student answer variable with two values: *uncertain or non-uncertain*, and a binary tutor response variable with two values: *D or notD*.[7] For example, the first data row of Table 2 shows a significant dependency between uncertain answers and the tutor's use of BOT in his responses: the χ^2 value is 13.70, which exceeds the critical value of 3.84 ($p \leq 0.05$, df=1).[8] This row also shows the observed (112) and expected (86) counts, whose comparison determines the dependency's sign. The "+" indicates that BOT occurs significantly more than expected after uncertain answers. A "-" indicates a dependency where the observed count is significantly less than expected. An "=" indicates a non-significant dependency (observed and expected counts are nearly equal).

This first χ^2 test does not tell us whether there is also a dependency between (in)correctness and the tutor's use of D in his response. Dependencies are expected for those D labels defined in relation to correctness (Section 2.3). Our second χ^2 test thus computes a χ^2 value for the dependency between a new binary student answer variable with two values: *correct or incorrect*, and the same tutor response variable with its two values: *D or notD*. For example, the second data row of Table 2 shows that BOT occurs significantly more than expected after incorrect answers. For discussion, the third data row shows the counts for BOTs after correct answers, but both rows express the same dependency.

Because uncertainty and incorrectness co-occur significantly more than expected (Section 2), the first two χ^2 tests cannot tell us whether a dependency between uncertainty and D exists independently of (in)correctness. Thus, for our third and fourth χ^2 tests, we first factor out the answers' correctness value, and then investigate the dependency between uncertainty and the use of D in the tutor's responses. More specifically, the third χ^2 test is applied only to the incorrect answers, and the fourth χ^2 test is applied only to the correct answers. For these tests, the student answer variable and tutor response variable have the same values as in the first test. For example, the second-to-last data row of Table 2 shows that even within the incorrect answers, there is a significant positive dependency between uncertainty and the tutor's use of BOT in his responses. However, the last row shows that this is not true of correct answers.

[6] We also used this method in our prior work [9], as discussed in Section 4.

[7] The χ^2 test arrays the variables' values along the row and column axes of a table. Each cell C contains the observed count of that row and column value co-occurring. C's expected count = *(C's row total*C's column total)/(overall total)*, and the χ^2 value = *(C's observed total - C's expected total)*^{2}*/C's expected total*. The overall χ^2 value for the dependency is computed by summing the χ^2 values over all cells.

[8] The critical χ^2 value accounts for the degrees of freedom (df = (#rows-1)*(#columns-1)) between the variables and the probability of exceeding a sampling error (e.g., $p \leq 0.05$). A dependency's significance increases as its χ^2 value increases.

3.3 Results

Tables 2-5 show uncertainty dependencies that do (or do not) remain significant after factoring out *incorrectness*. First, as discussed above, Table 2 shows that the tutor uses Bottom-Outs significantly more than expected after uncertain answers, incorrect answers, and uncertain answers within the incorrect answers.

Dependency		Obs.	Exp.	χ^2
Uncertain $\sim$ BOT	+	112	86	13.70
Incorrect $\sim$ BOT	+	139	76	88.76
Correct $\sim$ BOT	-	89	152	88.76
Uncertain within Incorrect $\sim$ BOT	+	82	72	3.86
Uncertain within Correct $\sim$ BOT	=	30	28	0.30

Table 2. Student Answer $\sim$ BOT Dependencies (p$\leq$.05: critical χ^2=3.84 (df=1))

The χ^2 test is not a causal test; however, we can formulate hypotheses about the reasons underlying dependencies, to help guide the development of our system adaptations. The strong *Incorrect* $\sim$ *BOT* dependency is not surprising, given that BOT by definition is the act of supplying a complete answer after an incorrect answer[9]. It is somewhat surprising that the tutor also uses BOT more than expected after uncertain answers; one might intuitively expect a HINT here. We hypothesize that the tutor uses BOT after uncertainty (overall and within incorrects) to respond to the student's perceived misconception conveyed by his/her uncertainty. [10] argues that uncertainty and incorrectness are both types of *learning impasses*: opportunities for students to learn what they are wrong and/or uncertain about (i.e., to resolve a misconception). However, the learning event requires first perceiving and then bridging the impasse. For uncertain incorrect answers, where an impasse is already perceived, the tutor may use BOT to provide this bridge. For non-uncertain incorrect answers, he may equally employ some other technique to help students first perceive the impasse. However, the relative weakness of the *Uncertain within Incorrect* $\sim$ *BOT* dependency suggests that for some non-uncertain incorrect answers, the tutor expects BOT to enable the student to both perceive and bridge the impasse.

Table 3 suggests what this other technique for non-uncertain incorrect answers may be. The tutor uses Hints significantly more than expected after incorrect answers, but not after uncertain answers (overall or within incorrects). Again, the strong *Incorrect* $\sim$ *HINT* dependency is not surprising; HINT is defined as the act of supplying help after an incorrect answer. Taken with the BOT results, the HINT results suggest that HINTs, unlike BOTs, aren't used as a specific response to uncertainty because uncertain students already perceive a learning impasse; rather, the tutor may often employ a HINT to ensure incorrect students first perceive (and possibly bridge) a particular impasse.

[9] BOTs, HINTs and NEGs after correct answers are cases where the answer was only partly correct or the tutor felt it should be filled out in some way.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ HINT	=	170	160	1.26
Incorrect ∼ HINT	+	227	141	101.32
Correct ∼ HINT	-	195	281	101.32
Uncertain within Incorrect ∼ HINT	=	115	117	0.12
Uncertain within Correct ∼ HINT	=	55	61	0.91

Table 3. Student Answer ∼ HINT Dependencies ($p \leq .05$: critical χ^2=3.84 (df=1))

Table 4 shows that the tutor uses Negative Feedback significantly more than expected only after uncertain answers and incorrect answers overall. Thus the uncertainty dependency is wholly accounted for by the stronger incorrectness dependency. The tutor may use NEG equally after all incorrect answers to assert his recognition of a learning impasse, e.g. before using a BOT or HINT.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ NEG	+	87	64	14.39
Incorrect ∼ NEG	+	154	56	278.09
Correct ∼ NEG	-	15	113	278.09
Uncertain within Incorrect ∼ NEG	=	81	79	0.08
Uncertain within Correct ∼ NEG	=	6	5	0.56

Table 4. Student Answer ∼ NEG Dependencies ($p \leq .05$: critical χ^2=3.84 (df=1))

Table 5 shows that the tutor uses Restatements significantly less than expected after uncertain answers and incorrect answers overall, but not after uncertain within incorrect answers. Again, the uncertainty dependency is wholly accounted for by the stronger incorrectness dependency. It is not surprising that the tutor is unlikely to restate or reword an incorrect answer; the tutor's increased use of RST appears to be a response to all types of correct answers.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ RST	-	199	252	26.88
Incorrect ∼ RST	-	71	221	229.58
Correct ∼ RST	+	593	443	229.58
Uncertain within Incorrect ∼ RST	=	30	37	2.78
Uncertain within Correct ∼ RST	=	169	184	3.44

Table 5. Student Answer ∼ RST Dependencies ($p \leq .05$: critical χ^2=3.84 (df=1))

Tables 6-8 show uncertainty dependencies that are significant even after factoring out *correctness*. The tutor uses Positive Feedback significantly more than expected after uncertain answers, correct answers, and uncertain within incorrect answers; the latter two dependencies are very strong. We hypothesize that the human tutor uses POS to respond to uncertainty over correctness as a direct method of bridging the perceived learning impasse: Positive Feedback asserts that the students' perceived misconception does not exist.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ POS	+	241	211	9.86
Incorrect ∼ POS	-	20	185	305.88
Correct ∼ POS	+	535	370	305.88
Uncertain within Incorrect ∼ POS	=	12	10	0.58
Uncertain within Correct ∼ POS	+	229	166	57.20

Table 6. Student Answer ∼ POS Dependencies (p≤.05: critical χ^2=3.84 (df=1))

Table 7 shows that the tutor uses Expansions significantly less than expected after uncertain answers, incorrect answers, and uncertain answers within the correct answers. It is not surprising that the tutor is more likely to expand on a correct answer than an incorrect one. However, the *Uncertain within Correct* ∼ *EXP* dependency suggests the tutor's strategy may be to address perceived (false) misconceptions without adding novel (and possibly confusing) information to his response.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ EXP	-	126	146	5.31
Incorrect ∼ EXP	-	96	128	14.77
Correct ∼ EXP	+	288	256	14.77
Uncertain within Incorrect ∼ EXP	=	51	50	0.11
Uncertain within Correct ∼ EXP	-	75	90	4.42

Table 7. Student Answer ∼ EXP Dependencies (p≤.05: critical χ^2=3.84 (df=1))

Table 8 shows that the tutor uses Short Answer Questions significantly less than expected after uncertain within correct answers. Like Expansions, this dependency suggests that the tutor's strategy may be to address perceived (false) misconceptions without asking for further basic information in his response.

Dependency		Obs.	Exp.	χ^2
Uncertain ∼ SAQ	=	211	226	2.21
Incorrect ∼ SAQ	=	206	198	0.67
Correct ∼ SAQ	=	389	397	0.67
Uncertain within Incorrect ∼ SAQ	=	107	106	0.01
Uncertain within Correct ∼ SAQ	-	104	121	4.94

Table 8. Student Answer ∼ SAQ Dependencies (p≤.05: critical χ^2=3.84 (df=1))

Finally, we found no significant dependencies involving the tutor's use of Hard Answer Questions (HAQs) or Recaps (RCPs). Overall our results suggest that some dialogue acts used in our tutor's response do depend on the prior student answer's uncertainty after factoring out correctness, but that these are not the only factors governing their use. For example, his use of Recaps and Hard Answer questions may instead depend on larger units of discourse structure, such as the number or type of topics covered so far, and/or larger units of uncertainty, such as the total overall uncertainty (within (in)correctness) seen so far. Moreover, the content of these dialogue acts may also depend on uncertainty; e.g. when the tutor uses a Short Answer Question that moves on to a new topic, versus when he uses a Short Answer Question that further queries the current topic. Although the current analysis is a step towards identifying human tutor responses to uncertainty, it does not capture such additional factors.

4 Related Work

This work builds on our prior work [9], where we also used χ^2 to investigate dependencies between uncertain student answers and dialogue acts used in the human tutor's response to these answers. There are numerous differences in this current work. First, here we distinguish uncertain and non-uncertain answers, because our current focus is on developing adaptations only for uncertain answers. In [9] we also distinguished neutral, certain, and mixed answers. Second, here our dataset is restricted to student answer turns, because in our ITSPOKE dialogues all student turns answer a tutor question. In [9] we included all student turns as a preliminary analysis. Third, in [9] we examined only uncertainty dependencies; here we also examine (in)correctness dependencies and uncertainty dependencies after factoring out (in)correctness. Fourth, in [9] we used a different version of our dialogue act tags. The version used here better corresponds to the dialogue acts used by our computer tutor. Here we also further develop the conclusions from this prior work. In particular, in [9] we also found that BOTs occur more than expected after uncertainty; here we showed this is actually only the case for uncertainty within incorrectness. Similarly, in [9] we also found that EXPs occur less than expected after uncertainty; here we showed this is actually only the case for uncertainty within correctness. We found similar clarifications of the tutor's use of the other dialogue acts by factoring out (in)correctness.

Our work also builds on related tutoring research developing system adaptations to student affect based on human tutor dialogue act responses. For example, [8] used a frequency analysis to extract two tutor responses to uncertain answers from a human tutoring corpus, then implemented and evaluated them in the SCoT-DC tutor. These adaptations, "paraphrasing" after uncertain+correct answers and "referring back to past dialogue" after uncertain+incorrect answers, were found to increase learning when used after all correct and incorrect answers, but not when used only after the uncertain answers. Other examples include researchers who have focused on developing computer tutor Feedback Acts that respond to affect as well as correctness. [13] developed a set of positive feedback responses based on a frequency analysis of the human tutor's responses in a spoken tutoring dialogue corpus, which included praising acknowledgments after uncertain+correct student turns, and implemented these responses in their Memory Game computer tutor. Students rated the system that used these positive feedback responses more highly than a version without. Such research suggests that human tutors adapt the content and presentation of their response to student uncertainty over and above correctness, and that these human tutor responses can be mined to develop effective computer tutor responses. However, none of these adaptations have yet shown a positive impact on student learning, which suggests that further research on affect adaptations is worthwhile. Moreover, our approach differs in that we use a statistical method of determining significant differences in how the human tutor responds to uncertainty, rather than a less rigorous frequency analysis.

5 Current Directions

We used χ^2 tests to identify and analyze dependencies between uncertain student answers and 9 dialogue acts the human tutor uses in his response to these answers. Within incorrect answers, we found that the tutor gives a Bottom Out significantly more than expected after uncertain answers. Within correct answers, the tutor gives Positive Feedback significantly more, and gives Expansions and ShortAnswer Questions significantly less, than expected after uncertain answers. We hypothesized that these dependencies reflect tutor methods of resolving learning impasses after students express perceived misconceptions.

Our next steps will be to develop and implement responses to uncertainty over correctness in ITSPOKE, based on analyses such as discussed here, and also on our recent work investigating contexts in our ITSPOKE corpora that are strongly associated with uncertainty [14]. For example, in [14] we found that uncertainty occurs significantly more than expected after Hard Answer Questions. Here we found BOT used significantly more than expected after uncertain and incorrect answers. Together these results suggest that ITSPOKE can use BOTs to respond to uncertain and incorrect answers to HAQs. This context-dependent approach to affect adaptation is described further in [14]. After developing and implementing our uncertainty adaptations, we will conduct a controlled experiment to test whether they improve student learning. More generally, our empirical approach can be used to develop adaptations for other dialogue systems and other user affective states, such as frustration, by analyzing dependencies between those states and human responses in annotated human-human dialogue corpora.

In general, it is common for dialogue system researchers to model systems on human behavior; as discussed in Section 4, this is particularly true for tutoring systems. Of course, this approach assumes that the human tutor's behavior can have a positive impact on the learning process. In our case, this assumption is supported by the fact that our human tutor had significant prior tutoring experience and our students learned significantly with our human tutor [11]. In future work we can examine this assumption further by running correlations between learning and the dependencies analyzed here (as in [9]). Although we have only analyzed the behavior of one human tutor, as discussed in [9], tutors have different teaching styles and skill levels; thus studying multiple tutors will not necessarily yield consistent generalizations about the "best" adaptive strategies. More generally, it is still an open question in the tutoring literature as to the "best" method of responding to uncertainty (and other affective states) [8, 10].

Acknowledgements

This material is based upon work supported by NSF (0328431 & 0631930) and ONR (N00014-04-1-0108). We thank the ITSPOKE Group.

References

1. Litman, D., Forbes-Riley, K.: Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication **48**(5) (2006) 559–590
2. Lee, C.M., Narayanan, S.: Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing **13**(2) (2005)
3. Vidrascu, L., Devillers, L.: Detection of real-life emotions in dialogs recorded in a call center. In: Proceedings of INTERSPEECH, Lisbon, Portugal (2005)
4. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: How to find trouble in communication. Speech Communication **40** (2003) 117–143
5. Liu, K., Picard, R.W.: Embedded empathy in continuous, interactive health assessment. In: CHI Workshop on HCI Challenges in Health Assessment. (2005)
6. Klein, J., Moon, Y., Picard, R.: This computer responds to user frustration: Theory, design, and results. Interacting with Computers **14** (2002) 119–140
7. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Communication **40** (2003) 5–32
8. Pon-Barry, H., Schultz, K., Bratt, E.O., Clark, B., Peters, S.: Responding to student uncertainty in spoken tutorial dialogue systems. International Journal of Artificial Intelligence in Education **16** (2006) 171–194
9. Forbes-Riley, K., Litman, D.: Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In Dybkjaer, L., Minker, W., eds.: Recent Trends in Discourse and Dialogue. Springer (2007) To Appear.
10. VanLehn, K., Siler, S., Murray, C.: Why do only some events cause learning during human tutoring? Cognition and Instruction **21**(3) (2003) 209–249
11. Litman, D.J., Forbes-Riley, K.: Correlations between dialogue acts and learning in spoken tutoring dialogues. Journal of Natural Language Engineering: Special Issue on Educational Applications **12**(2) (2006) 161–176
12. Graesser, A., Person, N., Magliano, J.: Collaborative dialog patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology **9** (1995) 495–522
13. Tsukahara, W., Ward, N.: Responding to subtle, fleeting changes in the user's internal state. In: Proceedings of the SIG-CHI on Human factors in computing systems, Seattle, WA, ACM (2001) 77–84
14. Forbes-Riley, K., Rotaru, M., Litman, D., Tetreault, J.: Exploring affect-context dependencies for adaptive system development. In: Proc. NAACL-HLT. (2007)