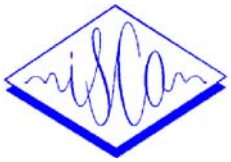

6th SIGdial Workshop on
Discourse and Dialogue
Lisbon, Portugal
September 2-3, 2005

Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus

Kate Forbes-Riley
University of Pittsburgh
Learning Research and Development Center
Pittsburgh PA, 15260, USA
forbesk@pitt.edu

Diane J. Litman
University of Pittsburgh
Department of Computer Science &
Learning Research and Development Center
Pittsburgh PA, 15260, USA
litman@cs.pitt.edu

Abstract

We use n-gram techniques to identify dependencies between student affective states of certainty and subsequent tutor dialogue acts, in an annotated corpus of human-human spoken tutoring dialogues. We first represent our dialogues as bigrams of annotated student and tutor turns. We next use χ^2 analysis to identify dependent bigrams. Our results show dependencies between many student states and subsequent tutor dialogue acts. We then analyze the dependent bigrams and suggest ways that our current computer tutor can be enhanced to adapt its dialogue act generation based on these dependencies.

1 Introduction

There has been increasing interest in *affective dialogue systems* (André et al., 2004), motivated by the belief that in human-human dialogues, conversational participants seem to be (at least to some degree) detecting and responding to the emotional states of other participants. Affective dialogue research is being pursued in many application areas, including *intelligent tutoring systems* (Aist et al., 2002; Craig and Graesser, 2003; Bhatt et al., 2004; Johnson et al., 2004; Moore et al., 2004). However, while it seems intuitively plausible that human tutors do in fact vary their responses based on the detection of student affect[1], to date this belief has largely been theoretically rather than empirically motivated. We propose using bigram-based techniques as a data-driven method for identifying relationships between student affect and tutor responses in a corpus of human-human spoken tutoring dialogues.

To investigate affect and tutorial dialogue systems, we have built ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman and Silliman, 2004), which is *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002).[2] Our long term goal is to have this system detect and adapt to student affect, and to investigate whether such an affective version of our system improves learning and other measures of performance. To date we have collected corpora of both human and computer tutoring dialogues, and have demonstrated the feasibility of annotating and recognizing student emotions from lexical, acoustic-prosodic, and dialogue features automatically extractable from these corpora (Litman and Forbes-Riley, 2004a; Litman and Forbes-Riley, 2004b; Forbes-Riley and Litman, 2004).

Here, we assume viable emotion recognition and move on to the next step: providing an empirical basis for enhancing our computer tutor to adaptively respond to student affect. We first show how to apply n-gram techniques used in other areas of computational linguistics to mine human-human dialogue corpora for dependent bigrams of student states and tutor responses. We then use our bigram analysis to show: 1) statistically-significant dependencies exist between students' emotional states and our *hu-*

[1] We use the terms "affect" and "emotion" loosely to cover emotions and attitudes believed to be relevant for tutoring.

[2] We also use ITSPOKE to examine the utility of building *spoken* dialogue tutors (e.g. (Litman et al., 2004)).

man tutor's dialogue act responses, 2) the dependent bigrams suggest empirically-motivated adaptive strategies for implementation in our *computer* tutor. This method should generalize to any domain with dialogue corpora labeled for user state and system response.

2 Spoken Tutoring Data and Annotation

2.1 The Spoken Tutoring Dialogue Corpus

Our data consists of 128 transcribed spoken dialogue tutoring sessions, between 14 different university students and one human tutor; each student participated in up to 10 sessions. The corpus was collected as part of an evaluation comparing typed and spoken human and computer dialogue tutoring (where the human tutor performed the same task as ITSPOKE) (Litman et al., 2004). The tutor and student spoke through head-mounted microphones, and were in the same room but separated by a partition.

Each session begins after a student types an essay answering a qualitative physics problem. The tutor analyzes the essay, then engages the student in dialogue to correct misconceptions and elicit more complete explanations. The student then revises the essay, thereby ending the session or causing another round of dialogue/essay revision. On average, these sessions last 18.1 minutes and contain 46.5 student and 43.0 tutor turns. Annotated (see Sections 2.2 - 2.3) excerpts[3] from our corpus are shown in Figures 1- 6 (punctuation added for clarity).

2.2 Annotating Student Certainty

Prior to the present study, each *student turn* in our corpus had been manually annotated for "certainness" (Liscombe et al., 2005)[4], as part of a larger project[5] annotating student states that are of interest in recent tutorial dialogue research (Bhatt et al., 2004; Moore et al., 2004). Our "Certainness" annotation uses one of four labels, defined in the manual as follows:

- **uncertain**: Use this label only when you feel the student is clearly uncertain about what they are saying. See Figures 1 (**STUDENT**$_{68}$) and 2 (**STUDENT**$_{17}$, **STUDENT**$_{19}$).
- **certain**: Use this label only when you feel the student is clearly certain about what they are saying. See Figures 1 (**STUDENT**$_{72}$) and 6 (**STUDENT**$_{99}$, **STUDENT**$_{101}$).
- **mixed**: Use this label if you feel that the speaker conveyed some mixture of uncertain and certain utterances within the same turn. See Figure 1 (**STUDENT**$_{75}$).
- **neutral**: Use this label when you feel the speaker conveyed no sense of certainness. In other words, the speaker seemed neither clearly uncertain nor clearly certain (nor clearly mixed). This is the default case. See Figure 1 (**STUDENT**$_{70}$, **STUDENT**$_{74}$).

PHYSICS PROBLEM: Suppose a man is running in a straight line at constant speed. He throws a pumpkin straight up. Where will it land? Explain.

. . . dialogue excerpt at 16.6 minutes into session. . .

STUDENT$_{68}$: So the- when you throw it up the acceleration will stay the same? *(UNCERTAIN)*

TUTOR$_{69}$: Acceleration uh will always be the same because there is- that is being caused by force of gravity which is not changing. *[RST, EXP]*

STUDENT$_{70}$: mm-k *(NEUTRAL)*

TUTOR$_{71}$: Acceleration is- it is in- what is the direction uh of this acceleration- acceleration due to gravity? *[SAQ]*

STUDENT$_{72}$: It's- the direction- it's downward. *(CERTAIN)*

TUTOR$_{73}$: Yes, it's vertically down. *[POS, RST]*

STUDENT$_{74}$: mm-k. *(NEUTRAL)*

. . . 53 seconds pass while student types. . .

STUDENT$_{75}$: So in relation to the velocity, when it's being thrown up the velocity increases- well no, it decreases and when it goes down it increases. *(MIXED)*

Figure 1: Annotated Corpus Excerpt #1

[3] All annotations were performed from both audio and transcription within a speech processing tool.

[4] To date, only one annotator has labeled "certainness". However, 10 dialogues were previously annotated by two other labelers using a more labor-intensive scheme, tagging certainness as well as confusion, boredom, frustration, etc. (Litman and Forbes-Riley, 2004a). Agreement across all three annotators for (un-)certain turns has an average 0.5 Kappa. We view this as a lower bound since the annotation tasks being compared are non-identical. This agreement is however similar to emotion annotation in other domains, e.g. 0.47 Kappa in (Ang et al., 2002), 0.45 and 0.48 Kappas in (Narayanan, 2002), and Kappas ranging between 0.32 and 0.42 in (Shafran et al., 2003).

[5] (Liscombe et al., 2005) show that using only acoustic-prosodic features as predictors, these student certainness annotations can be predicted with 76.42% accuracy.

PHYSICS PROBLEM: Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

. . . dialogue excerpt at 4.0 minutes into session. . .

TUTOR$_{16}$: Um, ok, so now we are thinking in terms of the factors that govern this displacement. Um, now they are- the elevator is in freefall. So does what does that tell you about the motion of the keys and the person? *[RD, LAQ]*

STUDENT$_{17}$: Um, that they're only under one force and that's gravity. *(UNCERTAIN)*

TUTOR$_{18}$:Ok, that is the force. But what does it tell you- that's right and that is about the force, but what does that tell you about their motion? *[POS, LAQ]*

STUDENT$_{19}$: Ok, uh, the motion is- oh, is they're moving in the same direction con- they're constant. *(UNCERTAIN)*

TUTOR$_{20}$: Uh, ok. They are also in freefall. *[BOT]*

Figure 2: Annotated Corpus Excerpt #2

2.3 Annotating Tutor Dialogue Acts

Also prior to the present study, each *tutor turn* in our corpus had been manually annotated for tutoring-specific dialogue acts[6] as part of a project comparing dialogue behavior in human versus computer tutoring (Forbes-Riley et al., 2005). Our tagset of "Tutor Dialogue Acts", shown in Figures 3 - 5 below, was developed based on pilot studies using similar tagsets applied in other tutorial dialogue projects[7] (Graesser and Person, 1994; Graesser et al., 1995; Johnson et al., 2004).

As shown in Figures 3 - 5, we distinguish three main types of Tutor Acts. The "Tutor Feedback Acts" in Figure 3 indicate the "correctness" of the student's prior turn.

The "Tutor Question Acts" in Figure 4 label the type of question that the tutor asks, in terms of their content and the expectation that the content presupposes with respect to the type of student answer required.

The "Tutor State Acts" in Figure 5 summarize or clarify the current state of the student's argument, based on the prior student turn(s).

- **Positive Feedback (POS)**: overt positive response to prior student turn. See Figures 1 (**TUTOR**$_{73}$), 2 (**TUTOR**$_{18}$) and 6 (**TUTOR**$_{98}$).
- **Negative Feedback (NEG)**: overt negative response to prior student turn. See Figure 6 (**TUTOR**$_{100}$).

Figure 3: Tutor Feedback Acts

- **Short Answer Question (SAQ)**: concerns basic quantitative relationships. See Figures 1 (**TUTOR**$_{71}$) and 6 (**TUTOR**$_{98}$, **TUTOR**$_{100}$).
- **Long Answer Question (LAQ)**: requires definition/interpretation of concepts. See Figure 2 (**TUTOR**$_{16}$, **TUTOR**$_{18}$).
- **Deep Answer Question (DAQ)**: requires reasoning about causes and/or effects. See Figure 6 (**TUTOR**$_{102}$).

Figure 4: Tutor Question Acts

Our corpus dialogue excerpts in Figure 1, Figure 2, and Figure 6 illustrate that most tutor turns are labeled with multiple Tutor Acts.

3 Data Analysis

We hypothesize that there are dependencies between student emotional states (as represented by the "Certainness" labels) and subsequent tutor responses (as represented by "Tutor Dialogue Act" labels), and that analyzing these dependencies can suggest ways of incorporating techniques for adapting to student emotions into our computer tutor. We test these hypotheses by extracting a bigram representation of student and tutor turns from our annotated dialogues, computing the dependencies of the bigram permutations using Chi Square analyses, and drawing conclusions from the significant results.

3.1 Dialogue Bigrams

We view the sequence: "Student Turn, Tutor Turn" as our bigram unit, whose individual elements con-

[6] While one annotator labeled the entire corpus, a second annotator labeled 776 of these turns, yielding a 0.67 Kappa.

[7] Tutoring dialogues have a number of tutoring-specific dialogue acts (e.g., hinting). Most researchers have thus used tutoring-specific rather than more domain-independent schemes such as DAMSL (Core and Allen, 1997), although (Rickel et al., 2001) present a first step towards integrating tutoring-specific acts into a more general collaborative discourse framework. Our Feedback and Question Acts have primarily backward- and forward-looking functions respectively, in DAMSL.

- Restatement **(RST)**: repetitions and rewordings of prior student statement. See Figures 1 (**TUTOR**$_{69}$, **TUTOR**$_{73}$) and 6 (**TUTOR**$_{100}$, **TUTOR**$_{102}$).
- **Recap (RCP)**: restating student's overall argument or earlier-established points. See Figure 6 (**TUTOR**$_{98}$).
- **Request/Directive (RD)**: directions summarizing expectations about student's overall argument. See Figure 2 (**TUTOR**$_{16}$).
- **Bottom Out (BOT)**: complete answer supplied after student answer is incorrect, incomplete or unclear. See Figure 2 (**TUTOR**$_{20}$).
- **Hint (HINT)**: partial answer supplied after student answer is incorrect, incomplete or unclear. See Figure 6 (**TUTOR**$_{100}$).
- **Expansion (EXP)**: novel details about student answer supplied without first being queried to student. See Figure 1 (**TUTOR**$_{69}$).

Figure 5: Tutor State Acts

stitute "words" in the bigram. In Figure 6 there are two such units: **STUDENT**$_{99}$ - **TUTOR**$_{100}$ and **STUDENT**$_{101}$ - **TUTOR**$_{102}$. Because our goal in this paper is to analyze tutor responses, we extract all and only these units from our dialogues for analysis. In other words, we do not extract bigrams of the form: "Tutor Turn, Student Turn", although we will do so in a separate future study when we analyze student responses to tutor actions. This decision is akin to disregarding word-level bigrams that cross sentence boundaries. Here, the sequence: "Student Turn, Tutor Turn" is our "dialogue sentence", and we are interested in all possible permutations of our student and tutor turn annotations in our data that combine to produce these dialogue sentences.

After extracting the annotated "Student Turn, Tutor Turn" bigrams, we sought to investigate the dependency between student emotional states and tutor responses. Although each of our student turns was labeled with a single "Certainty" tag, frequently our tutor turns were labeled with multiple "Tutor Act"

PHYSICS PROBLEM: Two closed containers look the same, but one is packed with lead and the other with a few feathers. How could you determine which had more mass if you and the containers were floating in a weightless condition in outer space? Explain.

. . . dialogue excerpt at 16.5 minutes into session. . .

TUTOR$_{98}$: Yes, we are all learning. Ok, so uh now uh you apply the same push for the same amount of time for on both the containers. Then what would you compare to distinguish between them? *[POS, RCP, SAQ]*

STUDENT$_{99}$: I would be comparing their rate of velocity. *(CERTAIN)*

TUTOR$_{100}$: Not rate. You will be comparing their velocity, you see, rate will imply that something is changing which there is no change, velocity is constant. So you will surely compare their velocities- which one will be faster? *[NEG, HINT, RST, SAQ]*

STUDENT$_{101}$: The feathers. *(CERTAIN)*

TUTOR$_{102}$: The feathers- why? *[RST, DAQ]*

Figure 6: Annotated Corpus Excerpt #3

tags, as shown in Figures 1-6. Because there are 11 "Tutor Act" tags, and no limits on tag combinations per turn, it is not surprising that in our 4921 extracted bigrams, we found 478 unique tag combinations in the tutor turns, 294 of which occurred only once. Treating each tagged tutor turn as a unique "word" would thus yield a data sparsity problem for our analysis of bigram dependencies. Due to this data sparsity problem, a question we can ask instead, is: is the tutor's *inclusion* of a particular Tutor Act in a tutor turn dependent on the student's certainness in the prior turn?

That is, we decided to approach the dependency analysis by considering the presence or absence of each Tutor Act tag separately. In other words, we performed 11 different analyses, one for each Tutor Act tag T, each time asking the question: is there a dependency between student emotional state and a tutor response containing T? More formally, for each analysis, we took our set of "Student Turn, Tutor Turn" bigrams, and replaced all annotated tutor turns containing T with only *T*, and all not containing T with *not T*. The result was 11 different sets of 4921 "Student Turn, Tutor Turn" bigrams. As an example, we show below how the tutor turns in Figure 6 are converted within the "POS" analysis:

TUTOR$_{98}$: *[POS, RCP, SAQ]* $\longrightarrow$ *[POS]*

TUTOR$_{100}$: *[NEG, HINT, RST, SAQ]* $\longrightarrow$ *[not-POS]*

TUTOR$_{102}$: *[RST, DAQ]* $\longrightarrow$ *[notPOS]*

The benefit of these multiple analyses is that we can ask specific questions directly motivated by what our computer tutor can do. For example, in the POS analysis, we ask: should student emotional state impact whether the computer tutor generates positive feedback? Currently, there is no emotion adaptation by our computer tutor - it generates positive feedback independently of student emotional state, and independently of any other Tutor Acts that it generates. The same is true for each of the Tutor Acts generated by our computer tutor.

3.2 Chi Square (χ^2) Analyses

We analyzed bigram dependency using the Chi Square (χ^2) test.[8] In this section we illustrate our analysis method, using the set of "Certainness" - "POS/notPOS" bigrams. In Section 3.3 we discuss the results of performing this same analysis on all 11 sets of "Student Certainness - Tutor Act" bigrams.

χ^2 tests the statistical significance of the relationship between two variables in a dataset. Our observed "Certainness" - "POS" bigram permutations are reported as a bivariate table in Table 1. For example, we observed 252 **neutral - POS** bigrams, and 2517 **neutral - notPOS** bigrams. Row totals show the number of bigrams containing the first bigram "word" (e.g., 2769 bigrams contained "neutral" followed by "POS" or "notPOS"). Column totals show the number of bigrams containing the second bigram "word" (e.g., 781 bigrams containing "POS" as the second token).

	POS	**notPOS**	**Total**
neutral	252	2517	2769
certain	273	831	1104
uncertain	185	631	816
mixed	71	161	232
Total	781	4140	4921

Table 1: Observed Student "Certainness" - Tutor "Positive Feedback" Bigrams

χ^2 compares these observed counts with the counts that would be expected if there were no relationship at all between the two variables in a larger population (the *null* hypothesis). For each cell c in Table 1, the expected count is computed as: (c's row total * c's column total)/(total bigrams). Expected counts for Table 1 are shown in Table 2.

	POS	**notPOS**	**Total**
neutral	439.46	2329.54	2769
certain	175.21	928.79	1104
uncertain	129.51	686.49	816
mixed	36.82	195.18	232
Total	781	4140	4921

Table 2: Expected Student "Certainness" - Tutor "Positive Feedback" Bigrams

A χ^2 value assesses whether the differences between observed and expected counts are large enough to conclude that a statistically significant relationship exists between the two variables. The χ^2 value for the table is computed by summing the χ^2 value for each cell, which is computed as follows: (observed value - expected value)^2/expected value. The total χ^2 value for Table 1 is 225.92. χ^2 would be 0 if observed and expected counts were equal. However some variation is required (the "critical χ^2 value"), to account for a given table's degree of freedom and one's chosen probability of exceeding any sampling error. For Table 1, which has 3 degrees of freedom, the critical χ^2 value at a 0.001 probability of error is 16.27.[9] Our χ^2 value of 225.92 greatly exceeds this critical value. We thus conclude that there is a statistically significant dependency between Certainness and Positive Feedback.

We can look more deeply into this overall dependency by calculating the statistical significance of the dependencies between each specific "Certainness" tag and the Positive Feedback tag. The freely available Ngram Statistics Package (NSP) (Banerjee and Pedersen, 2003) computes these χ^2 values automatically when we input each set of our "Student Certainness - Tutor Act" bigrams. Figure 7 shows the resulting NSP output for the POS/notPOS analysis. Each row shows: 1) the bigram, 2) its rank (according to highest χ^2 value), 3) its χ^2 value, 4) the number of occurrences of this bigram, 5) the number of times the first token in this bigram occurs first in

[8] A good tutorial for using the χ^2 test is found here: www.georgetown.edu/facultyballc/webtools/web_chi_tut.html

[9] Degrees of freedom is computed as (#rows - 1) * (#columns - 1). Critical χ^2 values are listed in most statistics textbooks.

any bigram, 6) the number of times the second token in this bigram occurs last in any bigram.

neutral - POS	1	217.35	252	2769	781
certain - POS	2	83.63	273	1104	781
mixed - POS	3	39.58	71	232	781
uncertain - POS	4	33.88	185	816	781

Figure 7: NSP Output: **Certainness - POS** Bigrams

Each row in Figure 7 can alternatively be viewed as a 2 X 2 table of observed counts. For example, the table for the **neutral - POS** bigram has a "neutral" row (identical to that in Table 1) and a "non-neutral" row (computed by summing all the non-neutral rows in Table 1). This table has 1 degree of freedom; the critical χ^2 value at $p < 0.001$ is 10.83. As shown, all of the bigrams in Figure 7 have χ^2 values exceeding this critical value. We thus conclude that there are statistically significant dependencies between each of the Certainness tags and Positive Feedback.[10] In Section 3.3 we will see cases where there is an *overall* significant dependency, but significant dependencies only for *a subset* of the four Certainness tags.

Finally, we can compare the difference between observed and expected values for the statistically significant dependent bigrams identified using NSP. For example, by comparing Tables 1 and 2, we see that the human tutor responds with positive feedback more than expected after emotional turns, and less than expected after neutral turns. This suggests that our computer tutoring system could adapt to non-neutral emotional states by generating more positive feedback (independently of whether the Certainness value is certain, uncertain, or mixed).

3.3 Results and Discussion

In essence, for each of the 11 Tutor Acts described in Section 2.3, the first part of our χ^2 analysis determines whether or not there is an overall dependency between Student Certainness and that specific Tutor Act. The second part then determines how this dependency is distributed across individual Student Certainness states. In this section, we present and discuss our results of the χ^2 analysis across all 11 sets of our "Certainness - Tutor Act" bigrams. Note that the tables present only our best results, where the χ^2 value exceeded the critical value at $p < 0.001$ (16.27 and 10.83 for 3 and 1 degrees of freedom, respectively). If a bigram's χ^2 value did not exceed this critical value, it is not shown.

Table 3 presents our best results across our 2 sets of "Certainness - Feedback Act" bigrams. Each set's results are separated by a double line. The last column shows the χ^2 value for each bigram. The first row for each set shows the χ^2 value for the overall dependency between Certainness and Feedback (e.g. 225.92 for **CERT - POS**). The remaining rows per set are ranked according to the χ^2 values for the specific dependencies between each "Certainness" tag and the "Feedback" tag (e.g. 217.35 for **neutral - POS**).[11] Note that, while all bigrams shown are statistically significant at $p < 0.001$, as the χ^2 values increase above the critical value, the results become more significant. Each row also shows the observed (Obs) and expected (Exp) counts of each bigram.

Bigram	**Obs**	**Exp**	χ^2
CERT - POS	781	781	225.92
neutral - POS	252	439.46	217.35
certain - POS	273	175.21	83.63
mixed - POS	71	36.82	39.58
uncertain - POS	185	129.51	33.88
CERT - NEG	196	196	135.96
neutral - NEG	34	110.29	125.67
uncertain - NEG	68	32.5	48.41
mixed - NEG	24	9.24	25.77
certain - NEG	70	43.97	20.69

Table 3: Observed, Expected, and χ^2 for Dependent "Certainness" - "Feedback" Bigrams ($p<.001$)

As shown, there are overall dependencies between Student Certainness and both Positive and Negative Tutor Feedback. There are also dependencies between every specific Certainness tag and both Positive and Negative tutor Feedback. Moreover, in both cases we see that the tutor responds with more feedback than expected after all emotional student turns

[10] Note that the χ^2 value for each of the bigrams in Figure 7 is identical to its "Certainness - notPOS" counterpart. This can be understood by observing that the 2 X 2 observed (and expected) table for each "Certainness - POS" bigram is identical to its "notPOS" counterpart, *except* that the columns are flipped. That is, "not notPOS" is equivalent to "POS".

[11] These POS results are discussed in Section 3.2; in this section we summarize the results for all 11 bigram sets.

(non-neutral), and with less feedback than expected after neutral student turns. This suggests that an increased use of feedback is a viable adaptation to non-neutral emotional states. Of course, the type of feedback adaptation (POS or NEG) must also depend on whether the student answer is correct, as will be discussed further in Section 5.

Table 4 presents our best results across our 3 sets of "Certainness - Question Act" bigrams, using the same format as Table 3. As shown, there is an overall dependency *only* between Student Certainness and Tutor Short Answer Questions that is wholly explained by the dependency of the **neutral - SAQ** bigram, where the tutor responds to student neutral turns with slightly fewer Short Answer Questions than expected. Both of these χ^2 values barely exceed the critical value however, and they are much smaller than the χ^2 values in Table 3. Moreover, there are no dependencies at all between Student Certainness and Tutor Long or Deep Answer Questions (**LAQ/DAQ**).[12] These results suggest that "Question Acts" aren't highly relevant for adaptation to Certainness; we hypothesis that they will play a more significant role when we analyze student emotional responses to tutor actions.

Bigram	Obs	Exp	χ^2
CERT - SAQ	1135	1135	18.06
neutral - SAQ	588	638.65	11.94

Table 4: Observed, Expected, and χ^2 for Dependent "Certainness" - "Question Act" Bigrams (p<.001)

Table 5 presents our best results across our 6 sets of "Certainness - State Act" bigrams. There is an overall dependency between Student Certainness and Tutor Restatements, explained by the dependencies of the **certain - RST** and **neutral - RST** bigrams. There is also an overall dependency between Student Certainness and Tutor Recaps, explained by the dependent **neutral - RCP** bigram. However, the χ^2 values for the dependent RST bigrams are much larger than those for the dependent RCP bigrams.[13] Moreover, there are no dependencies (even at p<.05) between Student Certainness and Tutor Request Directives (**RD**). Although these three Tutor State Acts all serve a summary purpose with respect to the student's argument, RCP and RD are defined as more general acts whose use is based on the overall discussion so far. Only RST addresses the immediately prior student turn; thus it's not surprising that its use shows a stronger dependency on the prior student certainness. The tutor's increased use of RST after certain turns suggests a possible adaptation strategy of increasing or maintaining student certainty by repeating information that the student has already shown certainty about.

The remaining 3 bigram sets contain Tutor Acts that clarify the prior student answer. First, there is an overall dependency between Student Certainness and Tutor Bottom Outs, which is explained by the specific dependencies of the **neutral - BOT** and **uncertain - BOT** bigrams. After uncertain turns, the tutor "Bottoms Out" (supplies the complete answer) more than expected, and after neutral turns, less than expected. This suggests a straightforward adaptive technique for student uncertainty.

There is also an overall dependency between Student Certainness and Tutor Hints, which is explained by the dependencies of the **mixed - HINT** and **neutral - HINT** bigrams. After mixed turns, the tutor "Hints" (supplies a partial answer) more than expected, and after neutral turns, less than expected. This suggests an adaptive technique similar to the BOT case, except the tutor gives less of the answer because there is less uncertainty (i.e. there is more certainty because the student turn is mixed).

Finally, there is an overall dependency between Student Certainness and Tutor Expansions, which is explained by the dependencies of the **neutral - EXP** and **uncertain - EXP** bigrams. In this case, however, the tutor responds with an "Expansion" (supplying novel details) more than expected after neutral turns, and less than expected after uncertain turns. This suggests another adaptive technique to uncertainty, whereby the tutor avoids overwhelming the uncertain student with unexpected details.[14]

[12] All the **LAQ** bigrams except **certain - LAQ** are barely significant at p< .05. Of the **DAQ** bigrams, only **CERT - DAQ** and **uncertain - DAQ** barely exceed the critical value at p<.05.

[13] Of the RCP and RST bigrams not shown, only **certain - RCP** is significant at a lower critical value (p<.01).

[14] Of the BOT, HINT, EXP bigrams not shown, only the "certain" bigrams are significant at a lower critical value (p<.05).

Bigram	Obs	Exp	χ^2
CERT - RST	1102	1102	169.18
certain - RST	402	247.23	160.96
neutral - RST	477	620.08	97.29
CERT - RCP	289	289	20.15
neutral - RCP	199	162.62	19.77
CERT - BOT	308	308	82.52
neutral - BOT	103	173.31	69.58
uncertain - BOT	97	51.07	52.82
CERT - HINT	779	779	37.07
mixed - HINT	64	36.73	25.25
neutral - HINT	383	438.34	18.98
CERT - EXP	998	998	47.08
neutral - EXP	651	561.57	40.86
uncertain - EXP	109	165.49	29.00

Table 5: Observed, Expected, and χ^2 for Dependent "Certainness" - "State Act" Bigrams ($p<.001$)

4 Related Work

While there have been other approaches to using dialogue n-grams (e.g. (Stolcke et al., 2000; Reithinger et al., 1996)), such n-grams have typically consisted of only dialogue acts, although (Higashinaka et al., 2003) propose computing bigrams of dialogue state and following dialogue act. Moreover, these methods have been used to compute n-gram probabilities for implementing statistical components. We propose a new use of these methods: to mine corpora for only the significant n-grams, for use in designing strategies for *adapting to student affect* in a computational system. Previous Ngram Statistics Package (NSP) applications have focused on extracting significant *word* n-grams (Banerjee and Pedersen, 2003), while our "dialogue" bigrams are constructed from multiple turn-level annotations of *student certainness* and *tutor dialogue acts*. Although (Shah et al., 2002) have mined a human tutoring corpus for significant "dialogue" bigrams to aid in the design of adaptive dialogue strategies, their goal is to generate appropriate tutor responses to student *initiative*. Their bigrams consist of manually labeled student initiative and tutor response in terms of mutually exclusive categories of communicative goals.

In the area of affective tutorial dialogue, (Bhatt et al., 2004) have coded (typed) tutoring dialogues for student hedging and affect. Their focus, however, has been on identifying differences in human versus computer tutoring, while our focus has been on analyzing relationships between student states and tutor responses. Conversely, (Johnson et al., 2004) have coded their tutoring dialogue corpora with tutoring-specific dialogue acts, but have not annotated student affect, and to date have performed only qualitative analyses. Finally, while our research focuses on dialogue acts, others are studying affect and different linguistic phenomena such as lexical choice (Moore et al., 2004).

5 Conclusions and Current Directions

This paper proposes an empirically-motivated approach to developing techniques for adapting to student affect in our dialogue tutorial system. Furthermore, our method of extracting and analyzing dialogue bigrams to develop adaptation techniques generalizes to other domains that seek to use user affective states to trigger system adaptation. We first extract "dialogue bigrams" from a corpus of human-human spoken tutoring dialogues annotated for student Certainness and tutor Dialogue Acts. We then use χ^2 analysis to determine which bigrams are dependent, such that there is a relationship between the use of a Tutor Act and prior Student Certainness.

Our results indicate specific human tutor emotion-adaption methods that we can implement in our computer system. Specifically, we find that there are many dependencies between student states of certainty and subsequent tutor dialogue acts, which suggest ways that our computer tutor can be enhanced to adapt dialogue act generation to student affective states. In particular, our results suggest that "Bottoming Out" and avoiding "Expansions" are viable adaptations to student uncertainty, whereas "Hinting" is a viable adaptation to a mixed student state, and adapting by "Restatements" may help maintain a state of student certainty. Positive and Negative Feedback occur significantly more than expected after all the non-neutral student states, and thus seem to be a generally "human" way of responding to student emotions.

This approach for developing adaptive strategies is currently based on one human tutor's responses across dialogues with multiple students. Clearly,

different tutors have different teaching styles; moreover, it is an open question in the tutoring community as to whether, and why, one tutor is better than any other with respect to increasing student learning. Analyzing a different tutor's responses may yield different dependencies between student emotions and tutor responses. Analyzing the responses of multiple tutors would yield a broader range of responses from which common responses could be extracted and analyzed. However, the common adaptations of multiple tutors are not necessarily better for improving student learning than the responses of a human tutor who responds differently. Moreover, such a "mix and match" approach would not necessarily yield a *consistent* generalization about adaptive strategies for student emotion. We have already demonstrated that students learned a significant amount with our human tutor (Litman et al., 2004)[15]. Thus, although it is an open question as to *why* these students learn, analyzing our tutor's responses across multiple students enables a consistent generalization about one successful tutor's adaptive strategies for student emotion.

However, it is important to note that we do not know yet if these adaptive techniques will be "effective", i.e. that they will improve student learning or improve other performance measures such as student persistence (Aist et al., 2002) when implemented in our *computer* tutor. Our next step will thus be to use these adaptive techniques as a *guideline* for implementing adaptive techniques in ITSPOKE. We can then compare the performance of the adaptive system with its non-adaptive counterpart, to see whether or not student performance is improved. Currently ITSPOKE adaptation is based only on the correctness of student turns.

We will also investigate how other factors interact with student emotional states to determine subsequent Tutor Acts. For although our results demonstrate significant dependencies between emotion and our human tutor responses, only a small amount of variance is accounted for in our results, indicating that other factors play a role in determining tutor responses. One such factor is student "correctness", which is not identical to student "certainness" (as measured by "hedging" (Bhatt et al., 2004)); for example, a student may be "certain" but "incorrect". Other factors include the dialogue act that the student is performing. We have recently completed the annotation of student turn "correctness", and we have already annotated "Student Acts" in tandem with Tutor Acts. Annotation of student "Frustration" and "Anger" categories has also recently been completed. We plan to extend the n-gram analysis by looking at other n-grams combining these new annotations of student turns with tutor responses.

In addition to using dependent bigrams to develop adaptive dialogue techniques, these results also provide features for other algorithms. We plan to use the dependent bigrams as new features for investigating learning correlations (i.e., Do students whose dialogues display more **certain - POS** bigrams learn more?), furthering our previous work in this area (Forbes-Riley et al., 2005; Litman et al., 2004).

[15]The student means for the (multiple-choice) pre- and posttests were 0.42 and 0.72, respectively.

Acknowledgments

We thank Julia Hirschberg, Jennifer Venditti, Jackson Liscombe, and Jeansun Lee at Columbia University for certainness annotation and discussion. We thank Pam Jordan, Amruta Purandare, Ted Pedersen, and Mihai Rotaru for their helpful comments. This research is supported by ONR (N00014-04-1-0108), and NSF (0325054, 0328431).

References

G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. In *Proc. Intelligent Tutoring Systems Conference*.

E. André, L. Dybkjær, W. Minker, and P. Heisterkamp, editors. 2004. *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*, volume 3068 of *Lecture Notes in Computer Science*. Springer.

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. International Conf. on Spoken Language Processing (ICSLP)*.

S. Banerjee and T. Pedersen. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proc. 4th International Conference. on Intelligent Text Processing and Computational Linguistics*.

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*.

M. G. Core and J. F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In D. Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California.

S. D. Craig and A. Graesser. 2003. Why am I confused: An exploratory look into the role of affect in learning. In A. Mendez-Vilas and J.A.Mesa Gonzalez, editors, *Advances in Technology-based Education: Towards a Knowledge-based Society Vol 3*, pages 1903–1906.

K. Forbes-Riley and D. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proc. Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL)*.

K. Forbes-Riley, D. Litman, A. Huettner, and A. Ward. 2005. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings of the International Conference on Artificial Intelligence in Education*.

A. Graesser and N. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

A. Graesser, N. Person, and J. Magliano. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9:495–522.

R. Higashinaka, M. Nakano, and K. Aikawa. 2003. Corpus-based discourse understanding in spoken dialogue systems. In *Proc. Assoc. for Computational Linguistics (ACL)*.

W. Lewis Johnson, Paola Rizzo, Wauter Bosma, Sander Kole, Mattijs Ghijsen, and Herwin van Welbergen. 2004. Generating socially appropriate tutorial dialog. In André et al. (André et al., 2004), pages 254–264.

J. Liscombe, J. Venditti, and J.Hirschberg. 2005. Detecting certainness in spoken tutorial dialogues. In *Proc. InterSpeech*.

D. Litman and K. Forbes-Riley. 2004a. Annotating student emotional states in spoken tutoring dialogues. In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*.

D. J. Litman and K. Forbes-Riley. 2004b. Predicting student emotions in computer-human tutoring dialogues. In *Proc. Assoc. Computational Linguistics (ACL)*.

D. Litman and S. Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL)*.

D. J. Litman, C. P. Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proc. Intelligent Tutoring Systems Conference*.

J. D. Moore, K. Porayska-Pomsta, S. Varges, and C. Zinn. 2004. Generating tutorial feedback with affect. In *Proc. of the 17th International Florida Artificial Intelligence Research Sociey Conference*.

S. Narayanan. 2002. Towards modeling user behavior in human-machine interaction: Effect of errors and emotions. In *Proc. ISLE Workshop on Dialogue Tagging for Multi-modal Human Computer Interaction*.

N. Reithinger, R. Engel, M. Kipp, and M. Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proc. International Conf. on Spoken Language Processing (ICSLP)*.

J. Rickel, N. B. Lesh, C. Rich, C. L. Sidner, and A. Gertner. 2001. Building a bridge between intelligent tutoring and collaborative dialogue systems. In *Proc. of the International Conference on Artificial Intelligence in Education (AI-ED)*, pages 592–594.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

F. Shah, M. Evens, J. Michael, and A. Rovick. 2002. Classifying student initiatives and tutor responses in human-human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1).

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, M. Meteer, and C. Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics 26:3*.

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 Conference*.