Running Head: PREDICTING AFFECTIVE STATES FROM AUTOTUTOR DIALOGUE

Predicting Affective States expressed through an Emote-Aloud Procedure

from AutoTutor's Mixed-Initiative Dialogue

Sidney K. D'Mello, Scotty D. Craig, Jeremiah Sullins, and Arthur C. Graesser

University of Memphis

Contact person:

Sidney D'Mello
209 Dunn Hall
The University of Memphis
Memphis, TN 38152
Phone: 901 678-1690
Fax: 901 678-1336
sdmello@memphis.edu

Abstract

This paper investigates how frequent conversation patterns from a mixed-initiative dialogue with an intelligent tutoring system, AutoTutor, can significantly predict users' affective states (e.g. confusion, eureka, frustration). This study adopted an emote-aloud procedure in which participants were recorded as they verbalized their affective states while interacting with AutoTutor. The tutor-tutee interaction was coded on scales of *conversational directness* (the amount of information provided by the tutor to the learner, with a theoretical ordering of assertion > prompt for particular information > hint), *feedback* (positive, neutral, negative), and *content coverage* scores for each student contribution obtained from the tutor's log files. Correlation and regression analyses confirmed the hypothesis that dialogue features could significantly predict the affective states of confusion, eureka, and frustration. Standard classification techniques were used to assess the reliability of the automatic detection of learners' affect from the conversation features. We discuss the prospects of extending AutoTutor into an affect-sensing intelligent tutoring system.

Predicting Affective States expressed through an Emote-Aloud Procedure

from AutoTutor's Mixed-Initiative Dialogue

It is widely acknowledged that cognition, motivation, and emotions are three fundamental components of learning (Snow, Corno, & Jackson, 1996). Emotion has traditionally been viewed as merely a source of motivational energy (Harter, 1981; Miserandino, 1996; Stipek, 1998), but there is an alternative position that emotion is a complex independent factor that merits direct inquiry in its relation to learning and motivation (Ford, 1992; Meyer & Turner, 2002). During the last few years, the link between emotions and learning has received increasing attention in education, psychology, computational linguistics, and artificial intelligence (Breazeal, 2003; Conati, 2002; Craig, Graesser, Sullins, & Gholson, 2004a; Kort, Reilly, & Picard, 2001; Lepper & Woolverton, 2002; Litman & Forbes-Riley, 2004; Lester, Towns, & FitzGerald, 1999; De Vincente & Pain, 2002; Picard 1997; Wang et al., 2005). For example, Kim (2005) conducted a study that demonstrated that interest and self-efficacy of a learner significantly increased when the learner was accompanied by a pedagogical agent acting as a virtual learning companion sensitive to the learner's affect. Linnerenbrink and Pintrich (2002) reported that the posttest scores of physics understanding decreased as a function of negative affect during learning. Craig et al. (2004a) reported that increased levels of boredom were negatively correlated with learning of computer literacy, whereas levels of confusion and the state of flow (being absorbed in the learning process, Csikszentmihalyi, 1990) were positively correlated with learning in an AutoTutor learning environment. AutoTutor is an intelligent tutoring system that uses naturalistic dialogue patterns to tutor students (Graesser, Chipman, Haynes, & Olney, in press; Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, Harter, Kreuz, & TRG, 1999; Graesser, Person, Harter, & TRG, 2001), as will be elaborated later.

Kort, Reilly, and Picard (2001) proposed a comprehensive four-quadrant model that explicitly links learning and affective states. The learning process is separated into two axes, vertical and horizontal, labeled *learning* and *affect* respectively. The learning axis ranges from "constructive learning" at the top, where new information is being integrated into schemas[1], and "un-learning" at the bottom where misconceptions are hopefully identified and removed from schemas. The affect axis ranges from positive affect on the right to negative affect on the left. According to this model, learners move around the circle from a state of ease, to encountering misconceptions and knowledge gaps, to discarding misconceptions and gaps, to new understanding, and then back into a state of equilibrium. This sequence is not rigid and linear, but rather follows more complex dynamic mechanisms.

The students ideally begin in Quadrant I or II. That is, they might be curious and fascinated about a new topic of interest (Quadrant I) or they might be puzzled or bewildered and motivated to reduce confusion (Quadrant II). When they encounter obstacles, they see that their ideas need some improvement. They may move down into the lower half of the diagram (Quadrant III) with negative valence emotions (e.g., chagrin) as they eliminate their misconceptions. As they consolidate their knowledge (what works and what doesn't) they may move through Quadrant IV (through fresh research) to a new idea (and eventually back to Quadrant I). The Kort et al. (2001) model is an imaginative comprehensive theoretical model, but has not yet been supported by empirical data from human learners.

---

[1] Schemas first proposed by Bartlett (1932) refer to a type of conceptual framework used to interpret incoming information. They can be considered to be packages of world knowledge, such as stereotypes, scripts, frames and other categories of generic knowledge.

Much of the current research on the link between emotions (or affective states) and learning has come from the area of user modeling. Most of the work in this field has focused on identifying the user's emotions as they interact with computer systems such as tutoring systems (Fan et al., 2003) or educational games (Conati, 2002; 2004). Conati (2002) has developed a probabilistic system that can reliably track multiple emotions of the learner during interactions with an educational game. Their system relies on dynamic decision networks to assess the affective states of joy, distress, admiration, and reproach. The performance of their system has been measured on the basis of learner self reports (Conati, 2004) and inaccuracies that were identified have been corrected by updating their model (Conati & Mclaren, 2005).

Unfortunately, many of these types of systems only assess intensity, valence (Ball & Breeze, 2000), or a single affective state (Hudlicka & McNeese, 2002). For example, Guhe, Gray, Schoelles, and Ji (2004) have recently reported a system in which the user is monitored in an attempt to detect confusion during interaction with an intelligent tutoring system. The problem with this method is that one affective state is not sufficient to encompass the whole gamut of learning (Conati, 2002). Additional complexities arise from the fact that a person's reaction to the presented material can change depending on their goals, preferences, expectations and knowledge state (Conati, 2002). As a contrast to the single affective state approach, the present research is directed towards a set of affective states (confusion, frustration, boredom, etc.) that are believed to accompany processes involving the learning of deep level conceptual information.

Our research investigates the detection of affective states that arise during interactive dialogue in natural language. The use of dialogue to detect affect in learning environments is a reasonable information source to initially explore, as opposed to technical bodily sensors,

because dialogue information is abundant in virtually all conversations and inexpensive to collect. Some earlier work investigating dialogue and emotions has been conducted on the program ITSPOKE (Litman et al., 2004; Litman & Silliman, 2004). ITSPOKE integrates a spoken language component into the Why2-Atlas tutoring system (VanLehn et al., 2002). The spoken student dialogue turns were analyzed on the basis of lexical and acoustic features, with codings of either negative, neutral, or positive affect. Their results show that the student's affect (positive, neutral, negative) can be accurately detected by combining acoustic-prosodic and lexical features. However, when used individually, the lexical features outperform the acoustic-prosodic features (Litman & Forbes-Riley, 2004).

It is important to understand the mechanisms of dialogue and cognition before one can dissect the links with emotions. For example, the collaborative theory of communication (Schober & Clark, 1989) stipulates that it is important to understand the participants' roles in conversation as well as the grounding criterion. The grounding criterion is the mutual belief that the addressees have understood what the speaker meant to a criterion sufficient for current purposes (Clark & Shaefer, 1989). This notion of a grounding criterion could also point to possible links between dialogue and affective states in the learning process. While the participants are working toward a grounding criterion, the addressees are in a state of attempting to understand the content. When this fails, there is a state of cognitive disequilibrium (Graesser, Lu, Olde, Cooper-Pye, & Whitten, in press; Otero & Graesser, 2001), which produces confusion (Craig et al., 2004a). The grounding criterion is often restored in a state of understanding, which is occasionally preceded by an abrupt transition to eureka (the "ah hah" experience). However, if the learners fail to reach a grounding criterion, then they should eventually give up and

disengage, which could be displayed as boredom or frustration. Frustration could also occur

when the speaker moves ahead before the addressee has reached understanding.

The present study reports some data on the first-scale project to integrate affect-sensing

capabilities into an intelligent tutoring system with tutorial dialogue, namely AutoTutor

(D'Mello et al., 2005; Graesser et al., 1999; 2001; in press). There were three goals to the present

study. The first was to identify affective states that occur frequently during learning. The

affective states of interest were anger, boredom, confusion, contempt, curious, disgust, eureka,

and frustration. The second goal was to correlate the conversation features from AutoTutor's

dialogue with affective states expressed by the learners. An alternative refinement was to apply

multiple regression techniques that assessed which of the affective states can be predicted from

the conversation features. The third goal was to apply various classification algorithms towards

the automatic detection of the learners affect from the dialogue patterns manifested in

AutoTutor's log files.

<div align="center">AutoTutor's Mixed Initiative Dialogue</div>

The Tutoring Research Group (TRG) at the University of Memphis developed

AutoTutor, a fully automated computer tutor that simulates human tutors and holds

conversations with students in natural language (Graesser et al., 1999; 2001; in press). AutoTutor

attempts to comprehend the students' natural language contributions and then respond to the

students' verbal input with adaptive dialogue moves similar to human tutors. The design of

AutoTutor was inspired by explanation-based constructivist theories of learning (Aleven &

Koedinger, 2002; VanLehn, Jones, & Chi, 1992) and by previous empirical research that has

documented the collaborative constructive activities that routinely occur during human tutoring

(Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser & Person, 1994; Moore, 1995).

AutoTutor helps students learn by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialogue while the learner constructs an answer.

AutoTutor provides *feedback* on what the student types in (positive, neutral, or negative feedback), *pumps* the student for more information ("What else?"), *prompts* the student to fill in missing words, gives *hints*, fills in missing information with *assertions,* identifies and corrects *misconceptions* and erroneous ideas, *answers* the student's questions, and *summarizes* topics. A full answer to a question is eventually constructed during this dialogue, which normally takes between 30 and 200 turns between the student and tutor for one particular problem or main question (approximately the same number of turns as with human tutors).

AutoTutor's knowledge about the topic being tutored (computer literacy in this study) is represented by Latent Semantic Analysis (LSA) (Foltz, 1996; Foltz, Britt, & Perfetti, 1996; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) and a curriculum script on the material. LSA is a statistical technique that measures the conceptual similarity of two text sources. LSA computes a geometric cosine (ranging from 0 to 1) that represents the conceptual similarity between the two text sources. In AutoTutor, LSA is used to assess the quality of student responses and to monitor other informative parameters, such as topic coverage and student ability level. Student response quality is measured by comparing each of the student's verbal contributions with two classes of content stored in the curriculum script: one that contains potential good answers to the topic being discussed (called *expectations*) and one that contains anticipated bad answers (called *misconceptions*). The higher of the two geometric cosines (i.e., the conceptual match between the student input and expectations versus the conceptual match between the same input and misconceptions) is considered the best conceptual match and

thereafter determines how AutoTutor responds to the student with short feedback (positive,

negative, or neutral). For the domain of computer literacy, we have found our application of LSA

to be sufficiently accurate in evaluating the quality of learner responses (Graesser et al., 2000;

Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999).

A session with AutoTutor is comprised of a set of subtopics (main questions) that cover

specific areas of the main topics (hardware, internet, and operating systems). Each subtopic has

an associated set of expectations, potential dialogue moves to elicit expectations (e.g., hints,

prompts, assertions), misconceptions, corrections of misconceptions, and other slots in the

curriculum script that need not be addressed here. The expectations are ideally covered by a

series of turns in AutoTutor's conversation with the student in an attempt to construct an answer

to the current main question (subtopic). When an acceptable answer with the appropriate details

is gleaned from the student's responses (usually after 30 – 200 turns), AutoTutor moves on to the

next subtopic. At the end of each student turn, AutoTutor maintains a log file that captures the

student's response, a variety of assessments of the response, the feedback provided, and the

tutor's next move. Temporal information, such as the student's reaction time and response time,

is also maintained.

As an example, the exchange below is one of the shorter tutorial dialogues between a

college student and AutoTutor. This student was an above average student in the computer

literacy class.

01   AUTOTUTOR: When you turn on the computer, how is the operating system first

        activated and loaded into RAM? [this is the *main question*]

02   STUDENT: When the computer is turned on a file is automatically booted up.

03   AUTOTUTOR: Anything else?   [this is a *pump* for more information]

04   STUDENT: This file contains the operating system.

05   AUTOTUTOR: What about storage?   [this is a *hint*]

06   STUDENT: The operating system is then read into RAM and activated.

07   AUTOTUTOR: Right. The CPU executes instructions in ROM, which helps the

operating system to start.   [after giving *positive feedback*, there is an *assertion*]

08   STUDENT: After being read from ROM the OS is activated.

09   AUTOTUTOR: Right. Let's review what we've just gone over. The operating system

must first be stored on the hard disk. When you turn on the computer, the CPU executes

instructions stored in ROM. These instructions help boot the operating system. The

operating system is then loaded from the hard disk into RAM. [after *positive feedback*

comes a *summary*]

Table 1 provides an overview of information channels of the interaction history that are

available in AutoTutor's log files. Information channels that are not relevant to this study have

been ignored. The information can be divided into five categories: session information, response

information, LSA assessments, the dialogue advancer, and the tutor's feedback.

INSERT TABLE 1 ABOUT HERE

Session Information

The session information can be viewed as a combination of global and local temporal

markers that span the period of interaction. The *subtopic* number indicates the number of main

questions answered. It provides a global measure of sequential position within the entire tutorial

session. For example, for a one-hour session covering three subtopics, the third subtopic would

indicate that the student is approximately in the 40-60 minute time span[2]. The *turn,* on the other hand, provides a local measure of the number of student contributions in the current question (subtopic). One would expect fatigue or boredom with high subtopic numbers or possibly frustration from being stuck in a single subtopic manifested through high turn numbers.

Response Information

Since AutoTutor relies on LSA for the majority of its assessments of the student's responses to a question, we only consider the verbosity of the response in this section. The verbosity is measured by the number of words and characters in the student's response. Perhaps, short responses reflect frustration or confusion except in particular situations in which the tutor prompts the student for one- or two-word answers. Long responses may be indicative of a deeper grasp of concepts, possibly reflecting the experience of a state of flow (Csikszentmihalyi, 1990). An alternative hypothesis would be that long answers could indicate confusion as the student muddles through his or her thoughts while trying to arrive at the solution.

Latent Semantic Analysis (LSA) Assessments

AutoTutor relies on LSA as its primary computation of the quality of student responses in student turns. The local assessments for a given turn measure the student's response for that turn on the basis of its similarity to good answers (expectations) and bad answers (misconceptions and bugs). The Local Good Score is the highest match score between the content of student turn N and the set of expectations representing good answers. The Local Bad Score is the highest match to the set of bad answers. A high Local Good Score reflects progress in answering the main question, whereas a high Local Bad Score reflects resonance with misconceptions. The

---

[2] It should be noted that while it would be more accurate to use real time information from the log file rather than the subtopic number, the latter scheme is used due to inaccuracies in the timing recorded in the logs.

Delta Local Good Score and the Delta Local Bad Score measure changes in the Local Good

Score and the Local Bad Score, respectively. Therefore, a large Delta Local Good Score could be

caused by one of those rare *eureka* experiences.

The four Global parameters perform the same assessments as the Local parameters with

the exception that the text used for the LSA match is an aggregation of all of the student's turns

(1 through N) for a given subtopic. With this scheme, a student's past responses to a subtopic are

considered in AutoTutor's assessment of the student's current response.

Dialogue Advancer

At the end of each student turn, AutoTutor incorporates the various LSA assessments

when choosing its next pedagogically appropriate dialogue move. The dialogue move chosen can

be ordered on a scale on the basis of the amount of information AutoTutor supplies to the

learner. The ordering is pump < hint < prompt < correction < assertion < summary, with a pump

conveying the minimum amount of information (on the part of AutoTutor) and a summary

conveying the most amount of explicit information. Within the context of the emote-aloud study,

one might expect confusion to heighten after the occurrence of hints and prompts (when the

student is expected to think, often to no avail) and to diminish in the presence of assertions and

summaries (when the student can simply receive information from AutoTutor rather passively).

Similar predictions can be made for various other affective states.

Tutor Feedback

AutoTutor's short feedback (positive, neutral, negative) is manifested in its verbal

content, intonation, and a host of other non-verbal conversational cues. Table 1 shows examples

of AutoTutor's responses, characterized by the type of feedback being provided. One could

predict the occurrence of particular emotions as a result of the type of feedback provided. For

example, repeated negative feedback could cause frustration in a motivated student and boredom in a student lacking motivation.

<div align="center">Emote-Aloud Study Methodology</div>

Our emote-aloud procedure is a modification of the think-aloud procedure (Ericsson & Simon, 1993; van Someren, Barnard, & Sandberg, 1994). During a think aloud procedure, participants talk about their thought process while working on tasks that require deeper levels of thought, such as solving problems (Ericsson & Simon, 1993), comprehending text (Trabasso & Magliano, 1996), reading poetry (Eva-Wood, 2004), reading English literature (Earthman, 1992) or solving break down scenarios of electronic equipment (Graesser et al., in press). Our emote-aloud procedure works in a similar way. Participants are asked to state the affective states they are feeling while working on a task, in this case computer literacy with AutoTutor. This method allows for on-line identification of emotions while working on a task with minimal task interference.

It should be noted that think aloud studies and this current emote-aloud study collect data from a small number of participants because of the labor intensive nature of the data collection and analysis (e.g., transcription of protocols, segmenting and identifying meaningful units, and scoring interjudge reliability). For example, Newell and Simon's (1972) pioneering work on problem solving had less than a handful of participants contributing think aloud data. Chi et al.'s (1989) classical work on self-explanation similarly had a small sample of participants. Moreover, although not discussed here, this study collects and analyzes the facial actions of the participants during an emote-aloud episode, an equally labor-intensive process. The number of participants

can be small in the studies using the action unit coding[3], yet still yield rich and reliable data (Ekman, 2003).

Participants

The participants used in this study consisted of seven undergraduates. They were selected from the department of psychology subject pool at The University of Memphis. Two participants were discarded from the current analysis due to an extremely small number of emotional expressions. While the number of participants in this study was small, this methodology still yielded some reliable and statistically significant findings. However, this does raise the important empirical point that not all participants are amenable to the emote-aloud procedure.

Materials

Electronic materials. Participants interacted with a computer program called AutoTutor on topics in computer literacy. AutoTutor asked questions about computer hardware. The questions were deep-level (such as why, how, what-if) and required about a paragraph of information to answer correctly. AutoTutor holds a mixed-initiative dialogue[4] to assist the students in answering each question, as discussed in the previous section.

Knowledge Tests. The test consisted of 4-alternative multiple-choice questions on computer hardware. There were two versions of the test, with 24 items per version. These two

---

[3] The Facial Action Coding System (Ekman & Friesen, 1978) allows for "basic emotions" to be identified by coding specific facial behaviors based on the muscles that produce them.

[4] Mixed-initiative interaction allows for the direction and control of the interaction to be shifted between participants. Allen (1999) discusses different levels of mixed-initiative interaction. Louwerse, Graesser, Olney, and the Tutoring Research Group (2002) discusses several of the conversation skills required for a mixed-initiative interaction to be realized in AutoTutor.

tests were counterbalanced across participants to serve as either a pretest of domain knowledge

or a posttest to compute learning gains (i.e., posttest minus pretest). The two versions of the tests

have produced equivalent means in past research (Craig, et al., 2004a).

Procedure

When participants arrived in the lab, they were given an informed consent followed by a

24-item pretest on computer hardware. The participants subsequently interacted with AutoTutor

for approximately an hour and a half, during which they engaged in an emote-aloud activity.

During the interaction with AutoTutor, the participants were video recorded. They were asked to

make verbal reports when they experienced an affective state. Participants were given a list of

eight affective states along with definitions. The list of affective states consisted of anger,

boredom, confusion, contempt, curious, disgust, eureka, and frustration. Participants were also

encouraged to express any affective state not included in the provided list as well as instances in

which they experienced multiple affective states.

The affective states were functionally defined for the participants. Anger was defined as a

strong feeling of displeasure and usually of antagonism. Boredom was defined as the state of

being weary and restless through lack of interest. Confusion was defined as a failure to

differentiate similar or related ideas. Contempt was defined as the act of despising, a lack of

respect or reverence for something. Curious was defined as an active desire to learn or to know.

Disgust was defined as marked aversion aroused by something highly distasteful. Eureka was

defined as a feeling used to express triumph on a discovery. Frustration was defined as making

vain or ineffectual efforts however vigorous; a deep chronic sense or state of insecurity and

dissatisfaction arising from unresolved problems or unfulfilled needs. All definitions were taken

from Merriam-Webster online (2003). In addition to providing participants with the list of

affective states, the experimenters discussed the meanings of each affective state with the participants. The experimenters also answered any questions related to the definitions of these affective states.

After the 90 minute session ended, a 24-item posttest on computer hardware was administered, followed by the debriefing. The two versions of the 24-item test were counterbalanced across the participants.

Data Treatment

Affective State Identification. During the emote-aloud procedure, participants were video recorded while they self identified their affective states. These verbal reports of affective states were then identified in the videos. We scored an affective state as being in one of the 8 categories if the emotion terms were expressed verbatim. No obvious ambiguities were encountered in the participants' articulation of the targeted emotions, so the participants' self reports of affect states were taken to be valid, and no additional coding process was initiated.

Data Cleaning. As part of the experimental procedure, participants were encouraged to verbalize any affective state they experienced including those that were not included in the list of the eight specified emotions. This caused reports of two additional affective states, happiness (n=4) and upset (n=1). Due to their low frequencies and their not being among the original list of affective responses, these two states were eliminated from the analyses presented below. Participants were also encouraged to report multiple affective states if so experienced. Four instances were noted in which multiple emotions were expressed by the participants. These four multiple emotes were expanded into eight records by treating each response individually. The frequencies of the verbal reports made by the 7 participants for the eight listed affective states are presented in Table 2.

INSERT TABLE 2 ABOUT HERE

Due to a low frequency of observations, anger (n = 17), contempt (n=8), curious (n=1), and disgust (n=5) were not included in the subsequent analyses. This data cleaning procedure yielded reliable data for boredom, confusion, eureka, and frustration. Additionally, two participants were discarded due to a lack of expressed emotions (n = 6, n = 9). Therefore, an original database of 215 emote-alouds was reduced to 170 "emotes" that were verbal expressions of an emotion or affective state. Although eureka was relatively well reported, we suspect that this response functionally signified happiness or delight from giving a correct answer rather than a deep eureka experience. True eureka experiences are likely to be more rare than our data suggest. In a previous study by Craig et al. (2004a), who had judges observe learners during interactions with AutoTutor, there was only one eureka experience identified in 10 hours of tutoring. Therefore, although we suspect that eureka emotes may in reality be an expression of positive affect (joy, delight, happiness), we included all utterances of eureka in order to apply a consistent principled methodology.

Data Selection. After eliminating the participants and the emotions mentioned above, the AutoTutor log files were mined to obtain information from the various dialogue channels described in Table 1. More specifically, the turn that immediately preceded or accompanied the emote-aloud was selected as the representative turn for that emote. If any of the 23 dialogue parameters for such a turn were missing, that turn and its associated emote-aloud observation were discarded from the analysis. This resulted in a further reduction in the database from 170 to 144 records, with 35 instances of boredom, 44 of confusion, 27 of eureka, and 38 of frustration.

Results

The data for the information channels (See Table 1) were mined from AutoTutor's log files. The data were selected from the turn that occurred immediately preceding or during an emote-aloud utterance. These channels' data were then correlated with the affective state expressed by the participant. Preliminary analyses revealed significant correlations for the affective states of confusion, eureka, and frustration, whereas no significant correlations with boredom were found. The significant correlations were discovered between the affective states and the dialogue advancer, feedback, and LSA cosine channels. This suggests that the emotions experienced during interactions with AutoTutor are based on dialogue assessments on these three dimensions. The phase of the tutoring session (subtopic number and turn number) and the verbosity of student contributions were never statistically significant, so they were dropped from subsequent analyses.

Table 3 presents descriptive statistics on the percentages of observations with the four particular emotion categories as a function of the significant dialogue channels and subchannels. For example, if boredom and the various dialogue advancer categories are considered, 12.5% of the boredom emotes occur synchronously with hints, 27.5% with prompts, 12.5% with corrections, 45% with assertions, and 2.5% with summaries. The last column is a percentage calculated across all four emotions. As an example, consider the feedback scale. According to Table 3, 41.7% of the feedback provided by AutoTutor across all emotions was negative feedback, 7.7% neutral negative feedback, 13.1% neutral feedback, 2.4% neutral positive feedback, and 35.1% of the feedback provided was positive. Data for the LSA channel was restricted to the Local Good Score. This is because the eight measures of LSA scores (see Table 1) that evaluate the quality of the student's verbal contributions in turn N were all highly

intercorrelated. Therefore, we needed to select one of the eight measures to represent the student's answer quality. The best measure turned out to be the Local Good score, so we adopted this measure in Table 3 and all of the subsequent analyses, with the variable label *student answer quality*.

A number of conclusions can be gleaned from Table 3. We found, for example, that more emotions were elicited after hints, prompts, and assertions than the other dialogue advancer categories. These results should be interpreted with caution, however, because hints, prompts, and assertions were more frequently generated by AutoTutor than other dialogue moves such as summaries, pumps, and corrections. Regarding feedback, we found that more emote-alouds occurred after AutoTutor provided negative feedback than positive feedback. Perhaps AutoTutor was providing more negative feedback because, with the exception of eureka, the affective states analyzed were negative. Additionally, the lower local good cosine scores via LSA assessments, coupled with the generally low pretest scores (not shown here), indicate that the participants involved in this study were typically low domain knowledge students. The LSA scores reported in Table 3 are an average of the various Local Good Score observations (see Table 1) that occurred with each emote-aloud utterance. AutoTutor's LSA threshold for considering an expectation covered was .70 whereas the mean cosine in our data analyses was .53.

INSERT TABLE 3 ABOUT HERE

The dialogue advancer categories capture the method in which AutoTutor attempts to have expectations covered during the dialogue. For some learners, all AutoTutor needs to do is pump or give a few hints and the learner can fill in all of the expectations correctly. For other learners, however, it is AutoTutor who has to deliver information through assertions because students lack knowledge or are not forthright in supplying information. When AutoTutor tries to

get a single expectation E covered (e.g., The *hard disk is a storage medium*), it first gives a pump

(*What else?*), then a hint (*What about the hard disk?*), then a prompt for specific information

(i.e., an important word, *The hard disk is a medium of what?*), and then simply asserts the

information (*The hard disk is a medium for storage*). More specifically, after an initial pump

(*what else?*), there are two cycles of hint-prompt-assertion; AutoTutor exits these cycles as soon

as the student articulates expectation E. A summary is provided by AutoTutor after all of the

expectations are covered. A correction occurs after a student expresses a misconception or bad

answer.

Therefore, as mentioned earlier, these dialogue moves can be aligned on an ordinal scale

of *directness* of AutoTutor supplying information. A pump is the least direct while a summary is

the most direct, using the scale specified in Table 4. A scale ranging from -1 (negative feedback)

to 1 (positive feedback) was constructed for the five *feedback* sub-channels in Table 1.

Therefore, the dialogue categories were converted to two quantitative scales that measured

directness and feedback, as defined in Table 4. It should be noted that the directness scale in

Table 4 deletes pumps because emotions were never associated with this dialogue move.

INSERT TABLE 4 ABOUT HERE

Our analysis proceeded by first computing a set of 12 correlations which would be

generated in a matrix that crosses the four emotions (boredom, eureka, confusion, frustration)

with the three quantitative scales of the conversation features (tutor directness, tutor feedback,

student answer quality). A multiple regression procedure was then used to predict the affective

states expressed by the participants as a function of the directness, feedback, and answer quality.

Correlations between Conversation Features and Affective States

Point biserial correlations in the 3 by 4 matrix are presented in Table 5. Each of the conversation features predicted one or more of the affective states, but there were considerable variations among the affective states. None of the conversation features correlated with boredom, but there were correlations with the other affective states. Tutor directness showed a significant negative correlation with confusion. Tutor feedback showed a significant negative correlation with confusion and frustration but a strong positive correlation with eureka. Student answer quality had a significant positive correlation with eureka. These correlations were all in expected directions. One drawback in this statistical analysis is that the 144 observations that aggregated data from the 5 subjects and 4 emotions ignores potential correlations among the conversational features. These statistical problems are remedied in the multiple regression analyses reported next.

<div align="center">INSERT TABLE 5 ABOUT HERE</div>

Predicting Affective States from Conversation Features

Four multiple regression analyses were performed, one for each of the four affective states, with the three conversation features as predictors. In order to partial out variability among participants, we conducted the multiple regression analyses in two steps: participant variables and conversation features. In step 1, the predictors included the participants' pretest scores and dummy coded variables to differentiate participants. In step 2, the group of predictors was the three conversation features (tutor directness, tutor feedback, and student answer quality. Step 1 was entered first, with the residual variance passed onto step 2. In this fashion, we could partial out any of the variability associated with the participants' characteristics and determine the unique variance that could be ascribed to particular conversation features.

Significant overall relationships were found for eureka, confusion, and frustration, but not for boredom. In the case of boredom, there was a statistically significant effect of step 1 individual differences ($p < .05$ being adopted in this and all subsequent statistical tests). However, the introduction of the conversation features in step 2 was nonsignificant, F (3,136) = 2.07, R^2_{adj} = .261, with nonsignificant $\beta-$ weights of .101, .167, and -.018 for tutor directness, tutor feedback, and student answer quality, respectively. In essence, none of our conversational measures predicted boredom.

Eureka. This multiple regression analysis resulted in a significant step 1 and also step 2, F (3,136) = 20.36, R^2_{adj} = .343. The step 2 conversation features accounted for an increment of .27 of the variance over and above step 1, with β– weights of .134, .422, and .176 for tutor directness, tutor feedback, and student answer quality, respectively. The tutor feedback feature was statistically significant, but not the tutor directness and student answer quality. Student answer quality had a significant correlation in Table 5, but not a significant β– weight because answer quality was correlated with pretest scores, which had a significant β– weight (β = .107) in step 1 of the multiple regression analysis. The fact that the pretest score was a significant positive predictor of eureka implies that students endowed with superior domain knowledge produce better answers and exhibit more eureka states. As could be expected, the positive feedback provided by AutoTutor was a strong positive predictor of eureka.

Confusion. This multiple regression analysis resulted in a significant step 1 and also step 2, F (3,136) = 6.49, R^2_{adj} = .144. The step 2 conversation features accounted for an increment of .101 of the variance over and above step 1, with β– weights of -.388, -.216, and -.165 for tutor directness, tutor feedback, and student answer quality, respectively. The tutor directness and tutor feedback feature had statistically significant β– weights, but not the student answer quality.

The pretest score in step 1 was a significant positive predictor of confusion ($\beta = .132$),

suggesting that it was the more knowledgeable students who had deeper thoughts and

consequently were more prone to confusion. However, as the feedback provided by AutoTutor

leaned towards the negative direction, the learner experienced more instances of confusion.

Directness of AutoTutor showed a negative relationship with confusion, so it was the tutor hints

and prompts that were affiliated with confusion rather than the tutor's assertions and summaries.

Confusion is manifested much more often when the learner has to work and think. While the

results suggest that higher knowledge students experience more confusion, an alternative

explanation would be that they may be more comfortable with expressing confusion. On the

other hand, lower knowledge students who may be more prone to feeling very confused are

perhaps self conscious and intentionally neglect to report confusion. Further investigations with

more detailed protocol analyses would be required to decide between one of these two

hypotheses.

Frustration. This multiple regression analysis resulted in a nonsignificant step 1 but a

significant step 2, $F (3,136) = 9.01$, $R^2_{adj} = .133$. The step 2 conversation features accounted for

an increment of .133 of the variance over and above step 1, with $\beta-$ weights of .154, -.373, and

.008 for tutor directness, tutor feedback, and student answer quality, respectively. The tutor

feedback feature had a statistically significant $\beta-$ weight, but not tutor directness and student

answer quality. The fact that AutoTutor's negative feedback was the only predictor affiliated

with frustration is consistent with the correlations in Table 5 and intuitively plausible.

Classifying Affective States from Conversation Features

The multiple regression analyses presented above produced significant models for the

affective states of confusion, eureka, and frustration. In order to address the larger goal of

extending AutoTutor into an affect-sensitive intelligent tutoring system, the need for real time

automatic affect detection becomes paramount. Therefore, we present some preliminary results

on automatic detection of the learner's affect from the various conversation features manifested

through an interaction with AutoTutor. Specifically, we apply several standard classification

techniques in an attempt to detect the learner's affect from AutoTutor dialogue. The major

advantages of using dialogue for emotion detection lie in its simplicity and cost effectiveness

when compared to most sensors that monitor bodily expressions.

The multiple regression analyses failed to converge on a significant model for boredom,

so this affective state was eliminated from the analyses. This caused a further reduction of the

data set of 144 records to 109 records. Due to the relatively small size of this data set, the

classification results presented below are intended to serve as merely a proof of concept that

conversation features from an interaction with AutoTutor can be used for the automatic detection

of certain affective states that frequent the learning experience.

The Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005)

was used to comparatively evaluate the performance of various standard classification techniques

in an endeavor to detect affect from dialogue. The data set consisted of 44 samples of confusion,

28 of eureka, and 38 of frustration yielding a base rate (chance) of 34.5%. The classification

algorithms tested were a Naïve Bayesian classifier (John & Langley, 1995), a multilayer

perceptron (neural network using back propagation for training), a nearest neighbor classifier

(Aha & Kibler, 1991), C4.5 decision trees (Quinlan, 1993), and an additive logistic boosting

classifier (Friedman, Hastie, & Tibshirani, 1998) with a decision stump as the base learner. The

selection of these algorithms was influenced by the desire to compare accuracies in detecting

affect across a variety of unique classification schemes, such as Bayesian classification (Naïve

Bayes classifier), neural networks (multilayer perceptron), lazy classifiers (nearest neighbor),

decision tree classifiers (C4.5 decision tress), and meta classification schemes (additive logistic

boosting). Table 6 shows the overall classification results using k-fold cross-validation[5] (k = 10)

for the various classifiers when evaluated on the data consisting of the three conversation

features for the affective states of confusion, eureka, and frustration.

INSERT TABLE 6 ABOUT HERE

The various classification algorithms were moderately successful in detecting affect with

accuracies ranging from 52.3% to 62.4%. The additive logistic regression technique with a

decision stump as the base learner provided the highest accuracy with an 80.8% improvement

over the baseline of 34.5%. The additive logistic regression technique yielded a kappa value of

.43, which is comparable to interrater reliability scores achieved by actual human coders. For

example, Litman and Forbes-Riley (2004) report kappa scores of .4 in detecting positive,

negative, and neutral affect. Similarly, Ang et al. (2002) report a kappa score of .47 in human

judgments of frustration and annoyance in human-computer dialogue. Shafran, Riley, and Mohri

(2003) report kappa scores ranging from .32 to .42 in coding affect.

While the classification accuracies and kappa scores for the various classification

algorithms are useful in obtaining an overview of the reliability of detecting affect from

conversation features, they do not provide any insight on class level accuracies. Table 7 lists the

---

[5] In k-fold cross-validation the data set (N) is divided into k subsets of approximately equal size (N/k). The classifier is trained on (k-1) of the subsets and evaluated on the remaining subset. Accuracy statistics are measured. The process is repeated k times. The overall accuracy is the average of the k training iterations. Goutte (1977) has shown k-fold cross-validation to be superior than other techniques for small data sets.

precision, recall, and F-measure scores as metrics for assessing class level accuracy for the three affective states.

INSERT TABLE 7 ABOUT HERE

Precision (specificity) and recall (sensitivity) are standard metrics for assessing the discriminability of a given class. The precision for class C is the proportion of samples that truly belong to class C among all the samples that were classified as class C. Table 7 indicates that the precision for eureka and frustration were highly similar while the precision for confusion was somewhat lower. The recall score (sensitivity or true positive rate) provides a measure of the accuracy of the learning scheme in detecting a particular class. Table 7 shows that the recall of the eureka state is much higher than that of confusion and frustration. Finally, the F-measure provides a single metric of performance by combining the precision and recall. We see that the F-measure for confusion and frustration are quantitative similar and lower than that of eureka. These results support the claim that all three affective states can be automatically detected with reasonable accuracy through the additive logistic boosting algorithm, with eureka being easier to detect than confusion and frustration. The correlations between the affective states and the three conversation features (see Table 5) reveal that while both confusion and frustration were negatively correlated with positive tutor feedback, whereas eureka had a positive correlation. This outcome would explain some of the difficulty the classifier would face in discriminating between confusion and eureka. Additionally, eureka was positively correlated with the student answer quality feature. This would somewhat alleviate the uncertainties between eureka and the other two affective states.

Discussion

The emote-aloud methodology in this study has proven to be useful for classifying emotions while college students learn with AutoTutor. This procedure allowed us to identify the points during learning where affective events were occurring. From this, we were able to obtain participants' interaction patterns from AutoTutor's log files. These interaction patterns were used to predict the affective states expressed by the participants. They were also used for the automatic detection of particular affective states by the application of a variety of standard classification techniques.

We acknowledge that the emote-aloud methodology has some potential limitations. The frequency with which affective states were reported could be one potential pitfall with this methodology. Four affect states were removed from the analysis due to floor effects with reporting. These were anger, disgust, contempt, and curious. Three possible explanations can be offered for this phenomenon. First, perhaps the reporting was so low because these affective states do not occur often during learning. Past research by Kort et al. (2001), for example, has suggested that the basic emotions investigated by Ekman (2003), i.e., anger, fear, sadness, enjoyment, disgust, and surprise, do not occur frequently during learning. Other researchers have challenged the adequacy of basing a complete theory of emotions on these "basic" emotions (Rozin & Cohen, 2003). Three out of the four emotions on our list of low frequency emotions during learning can be considered to be basic emotions; these include anger, disgust (Ekman, 1973; Izzard, 1971) and contempt (Izzard, 1971). Perhaps these emotions in fact occur but participants underreport these states. A closer look at these affective states reveals that three of the four states are negative states so perhaps participants are less likely to report such negative emotions during learning. However, this would not explain why confusion, frustration, and

boredom were reported frequently, with only eureka being a positive emotion frequently expressed. Another reason why students tend to report low frequencies of some affective states could be attributed to the nature of the learning environment that the intelligent tutoring system (AutoTutor in this case) imposes on the learner. Perhaps this might account for the low reporting frequencies of curiosity. Curiosity might not have been experienced because students had no choice of tutoring topics in our experimental environment. If participants had been given a choice of topics, they might have picked one more relevant to their interests and displayed more curiosity. Research by Lepper and Woolverton (2002) has proposed that curiosity and engagement are systematically related to the learner's freedom of choices.

It appears that there are significant relationships between the conversation features and affective states experienced during learning. Our current analyses show that the directness in which speech acts are expressed by the tutor and the type of feedback given can significantly predict the learners' affective states. Confusion is affiliated with indirect tutor dialogue moves (hints and prompts rather than assertions and summaries), with negative tutor feedback, and with higher student knowledge. Eureka is affiliated with positive tutor feedback and higher student knowledge whereas frustration is affiliated with negative tutor feedback. It should be noted that, since negative tutor feedback is a predictor of both frustration and confusion, the other conversation features (dialogue advancer and student answer quality) would be required to differentiate between these two affective states. The overall predictability of the relations between these three affective states and the conversation features can be considered to be a measure of validity of the emote-aloud methodology, but alternative methods are needed to address the limitations discussed above.

Our results on the links between affect and conversation are compatible with the

grounding criterion hypothesis (Clark & Shaefer, 1989; Schober & Clark, 1989). According to

this hypothesis, confusion should be observed while learners are attempting to understand the

material but are having trouble reaching a grounding criterion for understanding. As this

hypothesis predicts, the present study revealed that confusion occurs when the tutor expresses

indirect statements and negative feedback. The less direct statements and the negative feedback

would make it harder for the learners to believe their understanding meets the grounding

criterion. The patterns for the affective states of eureka and frustration are also compatible with

the grounding criterion hypothesis. The hypothesis would predict that, as seen in our data, eureka

should occur when the grounding criterion is met and frustration when it is not met.

If the grounding criterion hypothesis holds in future replication, then it would give

indications of how to help structure a dialogue to generate the best interaction between affective

states and learning. For example, according to the grounding criterion hypothesis, the process of

attempting to reach a criterion, but failing, will cause confusion. The learner is in a state of

cognitive disequilibrium (Graesser, & Olde, 2003; Graesser et al., in press), with the hope that

more thoughtful cognitive activity will increase learning (Craig et al. 2004a; Crutcher & Healy,

1989; McNamara & Healy, 1995). It is important to point out that the learner must have a

sufficient amount of world knowledge and metacognitive ability in order to detect problems of

grounding and the state of cognitive disequilibrium; indeed, it is the better learners who know

what they do not know, who experience cognitive disequilibrium, and who manifest confusion

(Azevedo & Crowley, 2004; Chi et al., 1989; Craig et al., 2004; Graesser & Person, 1994;

Miyake & Norman, 1979). After a period of time, confusion might decrease with (a) successful

thoughts and constructive activities, resulting in more positive feedback and eureka or (b) more

direct answers from the tutor. This is where the AutoTutor system can help establish or restore the knowledge needed to satisfy the grounding criterion. If the state of confusion is held too long, then the learner potentially slips into frustration, as the grounding hypothesis predicts, with negative feedback elevating this affective state of frustration.

The multiple regression analyses resulted in significant predictions for confusion, eureka, and frustration. The two-step regression allowed us to statistically partial out variance attributable to individual differences before assessing the impact of the conversation features on emotions. After partialling out individual differences, we found that tutor directness, tutor feedback, and student answer quality were able to explain about 17% of the variance of learner affect states. Our data mining of the conversation log files was therefore successful in predicting emotions.

The results suggest that boredom is extremely hard to detect from the dialogue patterns obtained while interacting with AutoTutor. One could offer two plausible alternatives that address the apparent failure in detecting boredom from the conversation features. The easiest explanation would be that boredom is simply not manifested through AutoTutor's conversation features. Perhaps more sophisticated sensors may be needed to detect boredom. Some work in this area has been conducted at the Affective Computing Lab at MIT (Kapoor, Mota, & Picard, 2001; Kort et al, 2001; Mota & Picard, 2003; Picard, 1997). They have successfully developed a system that uses body posture patterns to automatically detect interest levels in children while they learn with a computer system (Mota & Picard, 2003). By using a neural network for real time classification of nine static postures (leaning back, sitting upright, etc) motivated by Bull (1987), they achieved an overall accuracy of 87.6% when validated on new subjects. Their system also recognized interest (high interest, low interest, and taking a break) by analyzing

posture sequences with an overall accuracy of 82.3% on known subjects and 76.5% on new subjects.

While the use of sensors that monitor gross body language or other bodily sensors may be necessary for detecting boredom, perhaps a more rigorous investigation of the methodology of the analyses presented here may be in order. One of the known limitations of the data analyses presented in this paper is that each verbal report of an emotion was analyzed only in the context of the immediately preceding turns of the student and tutor. Perhaps boredom detection requires a broader scope of contextual information, including patterns of conversation that evolve over a series of turns leading up to an emotional experience. Exclusion of this larger snapshot of context preceding an emotion utterance could be the second reason for the inability in detecting of a more subtle affective state like boredom. Future efforts will be directed towards the analysis of conversation features across a larger temporal resolution and number of turns.

Although the classification results were evaluated on a relatively small sample, they serve as a validation that it would be worthwhile to pursue sophisticated classification techniques. Such techniques include biologically motivated classifiers as well as traditional classification methods. We plan on developing classifiers that are based on the dynamic behaviors of neural populations involved in the olfaction processes of rabbits (Kozma & Freeman, 2001). Classifiers based on dynamical systems with chaotic activity observed in brains have been experimentally validated as powerful pattern classifiers for difficult classification problems, particularly in situations in which the data set is not linearly separable (Kozma & Freeman, 2001, 2002).

The larger project of extending AutoTutor into an affect-sensitive intelligent tutoring system involves the emersion of AutoTutor and the learner into an *affective loop*. In AutoTutor and other integrated learning environments, this involves the *detection* of the learners' affective

states relevant to learning, the *selection* of appropriate tutor actions that maximize learning while influencing the learner's affect, and the *synthesis* of emotional expressions by the tutor as it attempts to engage the learner in a more human like, naturalistic manner. We now direct the focus of this discussion to ways in which the detection and selection phases of the affective loop may be effectively incorporated in AutoTutor.

Affect detection in AutoTutor would involve the development of appropriate classification systems. In addition to the conversation features and the learners' pretest knowledge, AutoTutor will be endowed with sensors that gauge facial expressions, speech intonation contours, and posture parameters. The reported classification results based on conversation features, indicated that the additive logistic boosting classifier was more successful in detecting eureka than detecting confusion and frustration. Perhaps detection accuracies could be increased by combining the conversation features with more sophisticated sensing technologies. We are in the initial phase of an investigation that attempts to isolate a subset of facial action units (Ekman & Friesen, 1978) that are routinely observed in congruence with particular emotions associated with learning. In particular, D'Mello et al. (2005) reported action units 1, 2, and 14 (outer brow raise, inner brow raise, and dimpler respectively) were primarily associated with frustration, with a strong link between action units 1 and 2 occurring together. Confusion displayed associations with action units 4, 7, and 12 (brow lowerer, lid tightener, and lip corner puller respectively), and also a link of action unit 7 triggering action unit 4. Boredom showed an association with 43 or eye closure. While boredom did not display any explicit links with the other action units, it did show several weaker trends between eye blinks and various mouth movements, such as mouth opening, mouth closing and jaw dropping, perhaps indicative of a yawn (Craig et al., 2004b). Efforts towards affect detection from posture will be based on

the Body Scoring System (Bull, 1987) and interest detection in children (Mota & Picard, 2003).

Detection of affective states from speech would involve the combination of acoustic and vocal

prosodic features, as has been substantiated in previous emotion detection research (Ang et al.,

2002; Fernandez & Picard, 2005; Forbes-Reilly & Litman, 2004; Litman & Forbes-Reilly, 2004).

If the affect detection methods described above prove to be effective in identifying the

learner's emotions, we can direct our focus to the second stage of the affective loop. During this

phase, AutoTutor would select dialogue moves that adapt to the learner's affective states in

addition to cognitive states. This adaptation would increase the bandwidth of communication and

allow AutoTutor to respond at a more sophisticated metacognitive level.

The increased metacognitive responsivity could help AutoTutor and other ITS systems

encourage active learning by the participant. For example, this added metacognitive channel

could allow the system to determine when the learner needs help and when the tutor needs to

provide extra opportunities for learning. This might occur, for example, when the learner is

simply attempting to "game" the system (Aleven, & Koedinger, 2002; Aleven, Stahl, Schworm,

Fischer, & Wallace, 2003), i.e., the learner attempts to get through the system by repeated asking

for help to get answers. There could be many possible responses to the different affective states

of the learner and the context of the interaction. It is still an open question as to what the optimal

pedagogical strategies should be. The following are examples of possible strategies that address

the presence of boredom, frustration, and confusion in the learner. If the learner is bored, a state

that has been negatively correlated with learning (Craig et al., 2004a), then AutoTutor should

engage the learner in a task that increases interest and cognitive arousal, such as a simulation,

options of choice, a challenge, or a seductive embedded game. AutoTutor might implement

dialogue strategies that will have a high likelihood of inducing confusion or eureka to get the

student to reengage with the material. The affective state of frustration could also be remedied with dialogue strategies by using more direct feedback, assertions, and corrections of misconceptions. Confusion presents a key opportunity for the ITS to encourage learning. Since confusion has been positively correlated with learning (Craig et al., 2004a), it is not in itself a state to avoid during the learning process. It might be good to allow the student to stay in a state of confusion for awhile. The system could handle this in two ways. Highly engaged learners or learners with a history of successful learning might be allowed to work out their own confusion in a discovery learning environment (Bruner, 1961; Vavik, 1993) that requires self-regulated cognitive activities (Azevedo & Crowley, 2004). A second method would systematically scaffold the student out of the confused state. This method might work better for learners with lower domain knowledge and lower ability to self-regulate their learning activities.

The broader scope of the research is to fortify future learners with enhanced dynamic reasoning, automated cognitive assessment, and intelligent handling of emotions. A learning environment that monitors learner emotions is expected to be more motivating and personally relevant to the learner.

References

Aha, D., & Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning, 6,* 37-66.

Aleven V. & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer based Cognitive Tutor. *Cognitive Science, 26*, 147-179.

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R.M. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research*, *73*(2), 277-320.

Allen, J. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems*. Vol. 14(5), pp.14-16.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Azevedo, R., & Cromley, J.G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia. *Journal of Educational Psychology, 96*, 523-535.

Ball, G., & Breeze, J. (2000). Emotion and personality in a conversational agent. In J. Cassel, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 189-219). Boston: The MIT Press.

Bartlett, F.C. (1932) *Remembering: An Experimental and Social Study*. Cambridge: Cambridge University Press.

Breazeal, C. (2003). *Designing sociable robots* .Cambridge: MIT Press.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31* (1), 21-32.

Bull E. P. (1987) *Posture and gesture*. New York, Pergamon Press.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145-182.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Clark, H. H. & Shaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259-294.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Conati C. (2002). Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence, 16*, 555-575.

Conati C. (2004). How to evaluate models of user affect?. *Proceedings of ADS 04, Tutorial and Research Workshop on Affective Dialogue Systems*. Kloster Irsee, Germany, June 2004. p. 288-300.

Conati, C., & Mclaren, H. (2005). Data-driven Refinement of a Probabilistic Model of User Affect. *Proceedings of UM2005 User Modeling: Proceedings of the Tenth International Conference.* Edimburgh, UK, July 26-30.

Craig, S.D., Graesser, A. C.,Sullins, J., & Gholson, B. (2004a). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media, 29,* 241-250.

Craig, S. D., D'Mello, S. K., Gholson, B., Witherspoon, A., Sullins, J., & Graesser A.C. (2004b). Emotions during learning: The first steps toward an affect sensitive intelligent tutoring system.In J. Nall, & R. Robson (Eds.). *Proceedings of E-learn 2004: World conference on E-learning in corporate, Government, Healthcare, & Higher Education*

(pp.284-288). Norfolk, VA: Association for the Advancement of Computing in Education.

Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 669-675.

Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience. Harper-Row: NY.

De Vicente, A., & Pain, H. (2002). Informing the detection of students' motivational state : An empirical study. In S.A. Cerri, G. Gouarderes, and F. Paraguacu (Eds.), *Intelligent tutoring systems 2002*. Berlin, Germany: Springer.

D'Mello, S. K., Craig, S. D., Gholson, B., Franklin, S., Picard, R.,& Graesser, A. C. (2005). Integrating affect sensors in an intelligent tutoring system. In Affective Interactions: The Computer in the Affective Loop Workshop at *2005 International conference on Intelligent User Interfaces* (pp. 7-13) New York: AMC Press

Earthman, E. A. (1992). Creating the virtual work: Reader's processes in understanding literary texts. *Research in the Teaching of English, 26*, 351-384.

Ekman, P. (1973). Cross-cultural studies of facial expression. In P. Ekman (Ed.), *Darwin and facial expression: A century of research in* review(pp. 169-222). New York: Academic Press.

Ekman, P. (2003). *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Henry Holt and company, LLC.

Ekman, P, & Friesen, W. V. (1978). *The facial action coding system: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. (Rev ed.) Cambridge, MA: The MIT Press.

Eva-Wood, A. L. (2004). How think-and feel-aloud instruction influences poetry readers. *Discourse Processes, 38*, 173-192.

Fan, C., Sarrafzadeh, A., Overmyer, S., Hosseini, H. G., Biglari-Abhari, M., & Bigdeli, A. (2003). A fuzzy approach to facial expression analysis in intelligent tutoring systems. In Antonio Méndez-Vilas and J.A.Mesa González(Eds.) *Advances in Technology-based Education: Towards a Knowledge-based Society Vol 3.* (pp. 1933-1937). Badajoz, Spain: Junta De Extremadura.

Fernandez, R. & Picard, R. W. (2005). Classical and Novel Discriminant Features for Affect Recognition from Speech. *Interspeech 2005 - Eurospeech 9th European Conference on Speech Communication and Technology*. September 4-8, Lisbon Portugal.

Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers, 28,* 197-202.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 110-115). Mahwah, NJ: Erlbaum.

Forbes-Riley, K., & Litman, D. (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Boston, MA.

Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs.* London: Sage.

Friedman, J., Hastie T., & Tibshirani, R. (1998). *Additive logistic regression: A statistical view of boosting*. Stanford University.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation*, *9*, 1211-

1215.

Graesser, A.C., Chipman, P., Haynes, B.C., & Olney, A. (in press). AutoTutor: An intelligent

tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*.

Graesser, A.C., Lu, S., Olde, B.A., Cooper-Pye, E., & Whitten, S. (in press). Question asking and

eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices

when the devices break down. *Memory and Cognition*.

Graesser, A. C. & Olde, B. (2003). How does one know whether a person understands a device?

The quality of the questions the person asks when the device breaks down. *Journal of*

Educational Psychology, 95, 524–536.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational*

Research Journal, 31, 104-137.

Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group (2001). Teaching tactics

and dialogue in AutoTutor. *International Journal of Artificial Intelligence in Education,*

12, 257-279.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N. K., &

Tutoring Research Group (2000). Using latent semantic analysis to evaluate the

contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 129-148.

Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research

Group (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems*

Research, 1, 35-51.

Guhe, M., Gray, W. D., Schoelles, M. J., & Ji, Q. (2004, July). *Towards an affective cognitive*

architecture. Poster session presented at the Cognitive Science Conference, Chicago, IL.

Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the
    classroom: Motivation and informational components. *Developmental Psychology, 17*,
    300-312.

Hudlicka, E., & McNeese, D. (2002). Assessment of user affective and belief states for interface
    adaptation: Application to an Air Force pilot task. *User Modeling and User-Adapted
    Interaction, 12*(1), 1-47.

Izard, C. E. (1971). *The face of emotion.* New York: Appleton-Century-Crofts.

John, G. H., Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers.
    Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 338-
    345.

Kapoor, A., Mota, S., & Picard, R. (2001). Toward a learning companion that recognizes affect.
    In *Proceedings from Emotional and Intelligent II, the tangled knot of social cognition,
    AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Kim, Y. (2005). Empathetic Virtual Peers Enhanced Learner Interest and Self-Efficacy.
    Workshop on Motivation and Affect in Educational Software at the *12th International
    Conference on Artificial Intelligentce in Education*, Amsterdam, The Netherlands.

Kort, B., Reilly, R., & Picard, R. (2001). An affective model of interplay between emotions and
    learning: Reengineering educational pedagogy—building a learning companion. In T.
    Okamoto, R. Hartley, Kinshuk, & J. P. Klus (Eds.), *Proceedings IEEE International
    Conference on Advanced Learning Technology: Issues, Achievements and Challenges*
    (pp. 43-48). Madison, Wisconsin: IEEE Computer Society.

Kozma, R., & Freeman, W. J. (2001). Chaotic resonance: Methods and applications for robust

 classification of noisy and variable patterns. *International Journal of Bifurcation &*

 Chaos, 11, 1607-1629.

Kozma, R., & Freeman, W. J. (2002). Classification of EEG patterns using Nonlinear

 Neurodynamics and identifying chaotic phase transitions. *Neurocomputing, 44-46*, 1107-

 1112.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic

 analysis theory of acquisition, induction, and representation of knowledge. *Psychological*

 Review, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis.

 Discourse Processes, 25, 259-284.

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study

 of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of*

 psychological factors on education (pp. 135-158). Orlando, FL: Academic Press.

Lester, J. C., Towns, S.G., FitzGerald, P.J. (1999). Achieving affective impact: visual emotive

 communication in lifelike pedagogical agents. *The International Journal of Artificial*

 Intelligence in Education, 10(3-4), 278-291.

Linnenbrink, E. A., & Pintrich, P. R. (2002). The role of motivational beliefs in conceptual

 change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in*

 theory and practice (pp. 115-135). Dordretch, The Netherlands: Kluwer Academic

 Publishers.

Litman, D. J., & Forbes-Riley, K. (2004). Predicting student emotions in computer-human

 tutoring dialogues. In *Proceedings of the 42nd annual meeting of the association for*

computational linguistics (pp. 352-359). East Stroudsburg, PA: Association for

Computational Linguistics.

Litman, D. J., Rose, C. P., Forbes-Riley, K., VanLehn, K., Bhemhe, D., Silliman, S. (2004).

Spoken versus typed human and computer dialogue tutoring. In *Proceedings of the*

seventh international conference on intelligent tutoring systems (pp. 368-379). Berlin:

Springer Verlag.

Litman, D. J., & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system.

In *Proceedings of the human language technology conference: 3rd meeting of the North*

American chapter of the association of computational linguistics (pp. 52-54). Edmonton,

Canada: ACL.

Louwerse, M. M., Graesser, A. C., Olney, A. & the Tutoring Research Group. (2002). Good

computational manners: Mixed-initiative dialog in conversational agents. In C. Miller,

Etiquette for Human-Computer Work. *Papers from the 2002 Fall Symposium, Technical*

Report FS-02-02 (pp. 71-76).

McNamara D. S. & Healy, A. F. (1995). A procedural explanation of the generation effect: The

use of an operand retrieval strategy for multiplication and addition problems. *Journal of*

Memory and Language, 34, 399-416.

Merriam-Webster, Incorporated. (n.d.) *Merriam-Webster Online*. Retrieved July 7, 2003, from

http://www.m-w.com

Meyer, D. K., & Turner, J. C. (2002). Discovering emotion in classroom motivation research.

Educational Psychologist, *37*(2), 107-114.

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology, 88*, 203-214.

Miyake, N. & Norman, D.A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior, 18*, 357-364.

Moore, J. D. (1995). *Participating in explanatory dialogues*. Cambridge: MIT Press.

Mota, S. & Picard, R. W. (2003). "Automated Posture Analysis for Detecting Learner's Interest Level." *Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI*, June, 2003.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Otero, J., & Graesser, A.C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction, 19*, 143-175.

Picard, R. W. (1997). *Affective computing*. Cambridge: MIT Press.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Robson C. (1993). *Real word research: A resource for social scientist and practitioner researchers*. Oxford: Blackwell.

Rozin, P. & Cohen, A. B. (2003). High Frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion, 3,* 68-75.

Schober, M. F. & Clark, H.H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21,* 211-232.

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

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Snow, R., Corno, L., & Jackson, D. (1996). Individual differences in affective and cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243-310). New York: Macmillan.

Stipek, D. (1998). *Motivation to Learn: From Theory to Practice 3rd edition*. Boston: Allyn and Bacon.

Trabasso, T., & Magliano, J. (1996). Conscious understanding during comprehension. *Discourse Processes, 21*, 225-286.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self- explanation effect. *Journal of the Learning Sciences, 2*, 1-60.

VanLehn, K., Jordan, P., Rosé, C. P., Bhembe, D., Bottner, M., Gaydos, A., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S.A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring* (pp.158-167). Berlin: Springer – Verlag.

Vavik, L. (1993). Facilitating discovery learning in computer-based simulation learning environments. In R.D. Tennyson & A.E. Baron (Eds.), *Automating instructional design: Computer-based development and delivery tools* (pp. 403-449). Berlin, Germany: Springer-Verlag.

Wang, N., Johnson, W.L., Mayer, R., Rizzzo, P., Shaw, E., & Collins, H. (2005). The politeness effect: Pedagogical agents and learning gains. In Looi, C., McCalla, G., Bredeweg, B., &

Breuker, J. (Eds.), *Artificial intelligence in education* (pp. 686—693). Amsterdam: IOS Press.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. *International Journal of Artificial Intelligence in Education*, 535-542. Amsterdam: IOS Press.

Witten, I. H. & Frank E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Acknowledgments

Table 1.

Description of the information mined from AutoTutor's log files at the end of each student turn.

Channel	Subchannel	Description
Session	Subtopic Number	The current subtopic (question) in this session
Information	Turn Number	The number of the conversation turn within a subtopic
Response	Number of words	The number of words in the student's turn
Information	Number of characters	The number of characters in the student's turn
	Local Good Score	Similarity of content of student's turn to an expectation
	Delta Local Good Score	The change in the Local Good Score
	Global Good Score	Similarity of the history of student turns to expectations
Latent	Delta Global Good Score	The change in the Global Good Score
Semantic	Local Bad Score	Similarity of content of student's turn to a bad answer
Analysis	Delta Local Bad Score	The change in the Local Bad Score
Assessments	Global Bad Score	Similarity of the history of student turns to bad answers
	Delta Global Bad Score	The change in the Global Bad Score
	Pump	Minimal information provided. e.g. "What else"
	Hint	Provides a hint to the student to fill in proposition
Dialogue	Prompt	Prompts student to fill in a missing content word
Advancer	Correction	Corrects the student's misconception
	Assertion	Asserts information about an expectation
	Summary	Provides a summary of the answer

	Positive	Provides feedback terms such as: "good job", "correct"
Tutor	Neutral Positive	Provides feedback terms such as: "yeah", "hmm right"
Feedback	Neutral	Provides feedback terms such as: "uh huh", "alright"
	Neutral Negative	Provides feedback terms such as: "possibly", "kind of"
	Negative	Provides feedback terms such as: "wrong", "no"

Table 2.

Frequencies of emote-aloud's reported by each participant across the various affective states.

Participant	Affective States							
	Anger	Boredom	Confusion	Contempt	Curious	Disgust	Eureka	Frustration
PN3	0	15	9	0	1	0	2	14
PN6	0	1	1	0	0	0	3	4
PN8	1	6	4	0	0	1	1	4
PN9	11	2	33	0	0	0	23	20
PN11	0	2	0	1	0	0	0	3
PN14	1	10	4	5	0	3	0	5
PN16	4	7	3	2	0	1	2	6
Total	17	43	54	8	1	5	31	56

Table 3.

Percentages of observations with emotions as a function of dialogue channels.

Dialogue Channel	Sub Channel	Percentage (%)				
		Boredom	*Confusion*	*Eureka*	*Frustration*	*Overall*
	Pump	0.0	0.0	0.0	0.0	0.0
	Hint	12.5	25.0	17.9	8.3	16.1
	Prompt	27.5	36.5	35.7	29.2	32.1
Dialogue Advancer	Correction	12.5	15.4	0.0	6.3	9.5
	Assertion	45.0	23.1	35.7	54.2	39.3
	Summary	2.5	0.0	10.7	2.1	3.0
	Negative	32.5	42.3	0.0	72.9	41.7
	Neutral Negative	7.5	13.5	0.0	6.3	7.7
Tutor Feedback	Neutral	10.0	23.1	3.6	10.4	13.1
	Neutral Positive	0.0	3.8	7.1	0.0	2.4
	Positive	50.0	17.3	89.3	10.4	35.1
LSA Assessments	Local Good Score	0.57	0.45	0.72	0.48	0.53

Note.   The information for Local Good Score is not computed as a percentage but as an average across each emotion.

Table 4.

Scales for directness and feedback of AutoTutor's dialogue moves

Directness Channels	Feedback Channels	Score
Hint	Negative Feedback	-1.0
Prompt	Neutral Negative Feedback	-0.5
Correction	Neutral Feedback	0.0
Assertion	Neutral Positive Feedback	0.5
Summary	Positive Feedback	1.0

Table 5.

Correlations between the affective states and the tutor directness, tutor feedback, and student

answer quality.

Affective State	Tutor Directness	Tutor Feedback	Student Answer Quality
Boredom	.11	.13	.04
Eureka	.01	.53**	.26**
Confusion	-.26**	-.19*	-.14
Frustration	.16	-.40**	-.12

* significant at p< .05

** significant at p< .01

Table 6.

Comparison of various standard classification techniques to automatically detect learner's affect

from the conversation features.

Algorithm	Classification Accuracy (%)	Kappa Statistic
Naïve Bayes	61.5	0.42
Multilayer Perceptron	54.1	0.32
Nearest Neighbor	52.3	0.27
C4.5 Decision Tree	57.8	0.37
Additiive Logistic Regression	62.4	0.43

Note.   The Kappa statistic measures the proportion of agreement between two raters

with correction for chance. Kappa scores ranging from 0.4 – 0.6 are considered to be fair, 0.6 –

0.75 are good, and scores greater than 0.75 are excellent (Robson, 2003).

Table 7.

Detailed accuracies of the adaptive logistic boosting classifier, segregated by affect category.

Affective States	Accuracy Measures		
	Precision	Recall	F-Measure
Confusion	.591	.591	.591
Eureka	.647	.815	.721
Frustration	.645	.526	.580