Text-Based Affect Detection in Intelligent Tutors

Sidney D’Mello and Arthur Graesser
University of Memphis, USA

ABSTRACT
Affect-sensitive Intelligent Tutoring Systems are an exciting new educational technology that aspire to heighten motivation and enhance learning gains in interventions that are dynamically adaptive to learners’ affective and cognitive states. Although state of the art affect detection systems rely on behavioral and physiological measures for affect detection, we show that a textual analysis of the tutorial discourse provides important cues into learners’ affective states. This chapter surveys the existing literature on text-based affect sensing and focuses on how learners’ affective states (boredom, flow/engagement, confusion, and frustration) can be automatically predicted by variations in the cohesiveness of tutorial dialogues during interactions with AutoTutor, an intelligent tutoring system with conversational dialogues. We discuss the generalizability of our findings to other domains and tutoring systems, the possibility of constructing real-time cohesion-based affect detectors, and implications for text-based affect detection for the next generation affect-sensitive learning environments.

INTRODUCTION
Attempts to acquire a deep level understanding of conceptual information requires cognitive activities such as a systematic exploration of the problem space, generating self-explanations, making bridging inferences, asking diagnostic questions, causal reasoning, and critical thinking. These effortful deep reasoning and problem solving activities often lead to episodes of failure and the learner experiences a host of affective responses (Mandler, 1984; Stein & Levine, 1991). Negative emotions are experienced when expectations are not met, failure is imminent, and important goals are blocked. For example, confusion occurs when learners face obstacles to goals, contradictions, incongruities, anomalies, uncertainty, and salient contrasts (Festinger, 1957; Graesser & Olde, 2003; Piaget, 1952). Unresolved confusion can lead to irritation, frustration, anger, and sometimes even rage.

On the other hand, a learner may experience a host of positive emotions when misconceptions are confronted, challenges are uncovered, insights are unveiled, and complex concepts are mastered. Students who are actively engaged in the learning session may have a flow-like experience when they are so engrossed in the material that time and fatigue disappears (Csikszentmihalyi, 1990). They may also experience other positive emotions such as delight, excitement, and even one of those rare eureka (i.e. “a ha”) moments. Simply put, emotions are systematically affected by the knowledge and goals of the learner, as well as vice versa (Dweck, 2002; Mandler, 1984; Stein & Levine, 1991). Cognitive activities such as causal reasoning, deliberation, goal appraisal, and planning processes operate continually throughout the experience of emotion.

Given this inextricable link between emotions and learning, it is reasonable to hypothesize that computerized learning environments that are sensitive to the affective and cognitive states of a learner would positively influence learning. This is particularly true when deep learning is accompanied by confusion, frustration, anxiety, boredom, delight, flow, surprise and other affective experiences (D’Mello, Graesser, & Picard, 2007; Graesser, Chipman, King, McDaniel, & D’Mello, 2007; Lepper & Woolverton, 2002; Picard, 1997). An affect-sensitive learning environment would incorporate assessments of the
students’ cognitive, affective, and motivational states into its pedagogical strategies to keep students engaged, boost self-confidence, heighten interest, and presumably maximize learning. For example, if the learner is frustrated, the Intelligent Tutoring System (ITS) might generate hints to advance the learner in constructing knowledge and make supportive empathetic comments to enhance motivation. If the learner is bored, the tutor might present more engaging or challenging problems for the learner to work on.

Affect-sensitive ITSs need to detect learners’ emotions before they can be responsive to their emotions. Affect detection does not need to be perfect, but must have a reasonable degree of accuracy. The affect detection problem is quite challenging because emotions are notoriously noisy, murky, and have fuzzy boundaries. This is compounded by the fact that there are individual differences and contextual influences in experience and expression. Despite these challenges, the aspiring affect-sensitive ITSs leverage recent advances in affective computing (Picard, 1997) to automatically detect learners’ affective states. The early systems primarily monitored physiological measures, such as skin conductance and heart rate, or bodily measures, such as facial expressions, posture, and gestures, and paralinguistic features of speech such as loudness, pitch, and formant frequencies (Pantic & Rothkrantz, 2003; Zeng, Pantic, Roisman, & Huang, 2009).

However, there is an alternative to using physiological, bodily, and vocal measures for affect detection: the monitoring of textual features from tutorial dialogues. There are several advantages to utilizing textual features as an independent channel for affect detection. First, textual features are abundant and inexpensive to collect in ITSs that support natural language dialogues, such as AutoTutor and the Why-Atlas conceptual physics tutor (Graesser, Lu et al., 2004; VanLehn et al., 2002). Second, the textual features that are derived from tutorial dialogues are contextually constrained in a fashion that provides cues regarding the urgency, attitudes, and social dynamics between the student and tutor.

Can a textual analysis of tutorial dialogue be diagnostic of the affective states of the learner? There is some evidence to suggest that this is a distinct possibility. Recent advances in computational psycholinguistics and discourse have demonstrated that textual features can predict complex psychological phenomenon such as personality, deception, knowledge levels, and even physical and mental health outcomes (Campbell & Pennebaker, 2003; Hancock, Curry, Goorha, & Woodworth, 2008; Mairesse, Walker, Mehl, & Moore, 2007; Williams & D’Mello, in press). Therefore, it is plausible to expect that a textual analysis of tutorial dialogues provide insights into learners’ affective states. This possibility is explored in the present chapter.

BACKGROUND

Extended discourse-based affect detection has not been previously automated in learning environments. One exception is ITSpoke, a speech-enabled ITS that has successfully fused paralinguistic features with relatively shallow lexical features for affect detection (Litman & Forbes-Riley, 2006). Our survey of the literature will therefore focus on general approaches to text-based affect detection and will not exclusively focus on learning environments.

In general, there are four major methods used to analyze the affective content of text samples. Some of the first attempts identified a small number of dimensions that underlie expressions of affect (Osgood, May, & Miron, 1975; Samsonovich & Ascoli, 2006). This research was pioneered by Osgood and colleagues who analyzed how people in different cultures rated the similarity of various emotional words. Dimensional reduction analyses on the similarity matrices converged upon evaluation (i.e., good or bad) and potency (i.e., strong or weak), and activity (i.e., active or passive) as the critical dimensions. These dimensions of affective expressions are qualitatively similar to valence and arousal, which are considered to be the fundamental dimensions of affective experience (Barrett, Mesquita, Ochsner, & Gross, 2007; Russell, 2003).

The second strand of research involves a lexical analysis of the text in order to identify words that are predictive of the affective states of writers or speakers (Bestgen, 1994; Cohn, Mehl, & Pennebaker, 2004; Hancock, Landrigan, & Silver, 2007; Kahn, Tobin, Massey, & Anderson, 2007; Pennebaker, Mehl, & Niederhoffer, 2003; Shields et al., 2005). Several of these approaches rely on the Linguistic Inquiry and
Word Count (LIWC\textsuperscript{1}) (Pennebaker, Francis, & Booth, 2001), a validated computer tool that analyses bodies of text using dictionary-based categorization. LIWC based affect-detection methods attempt to identify particular words that are expected to reveal the affective content in the text (Cohn et al., 2004; Hancock et al., 2007; Kahn et al., 2007). For example, “crying,” and “grief” are words that map on to the sad category, while “love” and “nice” are words that belong to the positive emotion category. In addition to LIWC, researchers have developed lexical databases that provide affective information for common words. Two examples of this are WordNet-Affect, an extension of WordNet for affective content (Strapparava & Valitutti, 2004) and the Affective Norm for English Words (Bradley & Lang, 1999).

The third set of text-based affect detection systems go a step beyond simple word matching by performing a semantic analysis of the text. For example Gill and colleagues (Gill, French, Gergle, & Oberlander, 2008) analyzed 200 blogs and reported that texts judged by humans as expressing fear and joy were semantically similar to emotional concept words (e.g. phobia, terror, for fear and delight, bliss for joy). They used latent semantic analysis (LSA) (Landauer, McNamara, Dennis, & Kintsch, 2007) and the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) to automatically compute the semantic similarity between the texts and emotion keywords (e.g., fear, joy, anger). Although this method of semantically aligning text to emotional concept words showed some promise for fear and joy texts, it failed for texts conveying six other emotions, such as anger, disgust, and sadness. So it is an open question whether semantic alignment of texts to emotional concept terms is a useful method for monitoring affective states.

The fourth approach to textual affect sensing involves systems that construct affective models from large corpora of world knowledge and applying these models to identify the affective tone in texts (Akkaya, Wiebe, & Mihalcea, 2009; Breck, Choi, & Cardie, 2007; Liu, Lieberman, & Selker, 2003; Pang & Lee, 2008; Shaikh, Prendinger, & Ishizuka, 2008). For example, the word “accident” is typically associated with an undesirable event. Hence, the presence of “accident” will increase the assigned negative valence of the sentence “I was late to work because of an accident on the freeway.” This approach is sometimes called sentiment analysis or opinion extraction, or subjectivity analysis because it focuses on valence of a textual sample (i.e., positive or negative; bad or good), rather than assigning the text to a particular emotion category (e.g., angry, sad). Sentiment and opinion analysis is gaining traction in the computational linguistics and is extensively discussed in a recent review (Pang & Lee, 2008).

Issues, Controversies, Problems

As evident from the discussion above, the last decade has witnessed a surge of research activity aimed at automatically identifying the emotional tone of spoken and written text. Most of the systems have been successfully applied to targeted classes of texts with notable affect, such as movie reviews, product reviews, blogs, and email messages (Gill et al., 2008; Liu et al., 2003; Shaikh et al., 2008). Despite the methodological differences, one commonality that emerges is that the systems operate under the assumption that affective content is explicitly and literally articulated in the text (e.g. “I have some bad news”, “This movie is a real drag”). Although this may be a valid assumption for obviously emotion-rich corpora such as blogs and movie reviews, where people are directly expressing opinions, this may not generalize to learners’ responses to computer tutors.

In fact, when it comes to learning environments there is some evidence to the contrary. For example, an examination of 1637 student responses generated from a tutorial session with AutoTutor, an ITS with conversational dialogues (Graesser, Lu et al., 2004), yielded only a handful of utterances with explicit affective expressions (< 1%). The frequency is extremely low even though an in-depth analysis of videos of the tutorial sessions yielded approximately 3,000 affective experiences (Graesser et al., 2006). Students were clearly experiencing affective states while interacting with AutoTutor, but their typed responses rarely conveyed affective content explicitly. Instead, their responses mainly consisted of domain specific answers to the tutor’s questions (e.g., “RAM is short-term memory”) even when they are in the midst of rich affective experiences.

\textsuperscript{1} Please see Chapter X (this volume) for more information on LIWC.
In addition to models that relied on a direct expression of affective content, even more sophisticated semantic entailment methods failed in the tutorial domain (D’Mello & Graesser, in review). Hence, there is some doubt as to whether the existing text-based affect detection mechanisms are applicable to tutorial discourse, at least when learners are not making explicit affective expressions.

**Solutions and Recommendations**

A more systematic and subtle textual analysis of tutorial dialogues might be necessary to uncover subtle cues that might be diagnostic of learners’ affective states. This chapter describes a recent study that attempted to test this hypothesis. We analyzed cohesion relationships in naturally occurring tutoring dialogues with AutoTutor (D’Mello, Dowell, & Graesser, 2009). Cohesion, a textual construct, is a measurable characteristic of text that is signaled by relationships between textual constituents (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, McCarthy, & Graesser, in press). Cohesion is related to coherence, a psychological construct that is a characteristic of the text together with the reader’s mental representation of the substantive ideas expressed in the text (Graesser, McNamara et al., 2004). We hypothesize that variations in the cohesiveness of the tutorial dialogues will be predictive of the learners’ affective experiences. More specifically, we expect that a breakdown in cohesion to be predictive of affective states such as confusion and frustration, whereas strong cohesive relationships might be indicative of engagement/flow.

Cohesion relationships in tutorial dialogues were automatically computed with Coh-Metrix, a validated computational tool that provides over 100 measures of various types of cohesion, including co-reference, referential, causal, spatial, temporal, and structural cohesion (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser & McNamara, in press; McNamara et al., in press). Coh-Metrix also has hundreds of measures of linguistic complexity (e.g., syntactic complexity), characteristics of words, and readability scores. We begin with a brief description of Coh-Metrix\(^2\) with an emphasis on a subset of the cohesive indices that we expected to be diagnostic of learners’ emotional states.

**BRIEF INTRODUCTION TO COH-METRIX**

A detailed description is beyond the scope of this chapter, so we focus on brief description of the cohesion indices that were utilized (see Table 1). These include measures of co-reference cohesion, pronoun referential cohesion, causal cohesion, semantic cohesion, connectives and some measures of readability and word usage. A majority of these measures (i.e. the measures that were explicitly intended to assess cohesion differences) have been validated in their ability to discriminate between low and high cohesion texts obtained from 12 published studies (McNamara et al., in press).

**Co-reference cohesion**

This is an important type of cohesion that occurs when a noun, pronoun, or noun-phrase refers to another constituent in the text. For example, consider the following two sentences: (1) *Bob* decided to *clean* his carpets. (2) *So Bob* went into the store to purchase a vacuum *cleaner*. In this example the word *Bob* in the first sentence is a co-referent to the word *Bob* in the second sentence. This is an example of noun overlap. Co-reference cohesion can also be measured by morphological stem overlap. The word *cleaner* in sentence 2 shares the same morphological stem (i.e. clean) to the word *clean* in sentence 1, although one is a noun and the other a verb. Coh-Metrix computes the proportion of sentences with noun and stem overlap across window sizes of 1 and 2 sentences (C1 – C4 in Table 1).

**Pronoun referential cohesion**

Pronoun referential cohesion (C5) occurs when a pronoun in sentence N has at least one referent in sentence N-1 that is a suitable match to the pronoun in N. A suitable match occurs if the pronoun agrees with referent in gender and number. For example, consider the following sentences: (1) The user's

---

\(^2\) Please see Chapter X (this volume) for more information on Coh-Metrix
computer was extremely slow. (2) So he closed some files. The pronoun he in sentence 2 refers to the possessive noun user in sentence 1. Binding pronouns to previously defined entities in the text play a significant role in grounding the discourse. Unreferenced pronouns have a negative effect on the cohesiveness and thereby readability of the text. Pronoun resolution is a difficult and open computational linguistics problem (Jurafsky & Martin, 2008), so Coh-Metrix measures pronoun referential cohesion by computing the proportion of pronouns in the current sentence that have at least one grounded referent in a previous sentence.

Table 1. Cohesion features derived from Coh-Metrix

<table>
<thead>
<tr>
<th>Cohesion Features</th>
<th>Connectives</th>
<th>Measures of Readability and Verbosity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C5. Pronoun referential cohesion</strong></td>
<td>C14. Incidence of all connectives</td>
<td>C20. Negations</td>
</tr>
<tr>
<td><strong>Causal Cohesion</strong></td>
<td>C15. Negative temporal connectives</td>
<td>C21. Mean hypernym of nouns</td>
</tr>
<tr>
<td>C7. Causal verbs</td>
<td>C17. Positive additive connectives</td>
<td>C23. Content words</td>
</tr>
<tr>
<td><strong>Semantic Cohesion</strong></td>
<td>C18. Negative additive connectives</td>
<td>C24. Flesch Reading Ease Score</td>
</tr>
<tr>
<td>C8. LSA sentence to sentence adjacent mean</td>
<td>C19. Conditional connectives</td>
<td>C25. Maximum sentence length</td>
</tr>
<tr>
<td>C9. LSA sentence to sentence adjacent stdev</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C10. LSA paragraph to paragraph mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C11. LSA paragraph to paragraph stdev</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C12. Given information in each sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C13. Type token ratio for content words</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Causal cohesion**

Causal cohesion occurs when actions and events in the text are connected by causal connectives and other linking word particles (Graesser, McNamara et al., 2004; Zwaan & Radvansky, 1998). Events and actions have main verbs that are designated as intentional or causal (e.g. *kill, impact*), as indicated by categories in the WordNet lexicon (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Causal particles connect these events and actions with connectives, adverbs, and other word categories that link ideas (e.g., *because, consequently, hence*). Coh-Metrix provides measures on the incidence of causal verb categories (occurrences per 1000 words) (C6). But the most significant measure of causal cohesion is the *causal ratio* that specifies the ratio of causal particles to events and actions (C7). The causal ratio of a text is directly proportional to its cohesiveness and is specified as: \( \text{Causal Ratio} = \frac{\text{Causal Particles}}{\text{Causal Verbs} + 1} \). A high causal ratio indicates that there are many connectives and other particles that stitch together the explicit actions and events in the text.

**Semantic cohesion**

Semantic cohesion is measured as the conceptual similarity across text constituents. LSA is used as the primary computational tool for measuring semantic cohesion. For example, a paragraph in which adjacent
sentences have higher LSA scores (i.e. higher semantic similarity) is expected to be more cohesive than a paragraph in which the meaning widely fluctuates across sentences (low LSA scores for adjacent sentences). Coh-Metrix computes several measures of semantic cohesion at varying window sizes. The window can be a sentence, a paragraph, or a text. A subset of these measures were selected for the current analyses (C8-C11). Both mean LSA scores and standard deviations in the scores are computed. A semantic cohesion gap is expected to occur for adjacent sentences with low LSA scores and also when there is a high standard deviation (because of an occasional adjacency with a very low LSA score).

Given information (C12) computes the extent to which an incoming sentence is redundant with (i.e., LSA overlap) all of the previous sentences in the dialogue history for a particular problem. Finally, the type-token ratio for content words (C13) was included as a measure of lexical diversity (see Graesser et al., 2004 for a description of this measure).

Connectives
Connectives are words and phrases that play a significant role in signifying cohesion and coherence relationships by explicitly linking ideas expressed in a text (Graesser, McNamara et al., 2004; Louwerse, 2001; McNamara et al., in press). Coh-Metrix provides incidence scores on several types of connectives in a text. These include temporal (before, when), additive (also), and causal (because); these categories have both negative valences (however, in contrast) and positive valences (therefore, in addition) (C14-C19) (Louwerse, 2001).

Shallow Measures of Readability and Verbosity
Coh-Metrix provides these measures in addition to the primary cohesion indices (C20-C25). Table 1 lists the measures selected in the present analyses. There was a measure of the incidence of negations in the text (C20). There were measures of the degree of abstraction of nouns and verbs in the text (C21, C22), as computed as the hypernym index in WordNet (Miller et al., 1990). The mean frequency of content words (e.g. nouns, lexical verbs, adverbs, adjectives) was also included (C23). The frequency of content words provides a useful measure of the amount of substantive content in the text (Graesser et al., 2004). The last two measures were reading ease and verbosity. These included the Flesch Kincaid Reading Ease (Klare, 1974) (C24) and the length of the longest sentence in the text (C25).

TUTORIAL DIALOGUE CORPUS ANNOTATED FOR LEARNER AFFECT
In order to investigate relationships between the cohesion indices and learner affect, one needs a corpus of tutorial dialogues (to extract cohesion indices) and measures of learner affect at various points in the dialogues. We used an existing tutorial dialogue corpus that was collected in a study where 28 undergraduate students from a mid-south university in the U.S. interacted with AutoTutor. The tutorial sessions were for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, Internet, or operating systems (Graesser et al., 2006). AutoTutor is a validated ITS that teaches students conceptual physics, computer literacy, and critical thinking via a mixed-initiative dialogue in natural language (Graesser, Lu et al., 2004; VanLehn et al., 2007). During the tutoring session, we recorded a video of the participant’s face, their posture pressure patterns (not elaborated here), and a video of their computer screen for offline analyses. The student-tutor conversational dialogues were submitted to Coh-Metrix in order to extract the cohesion indices, whereas the videos of the learners’ faces and computer screens were used to infer their affective states.

Computing Cohesion Indices from AutoTutor's Dialogue
AutoTutor’s dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers (e.g. When you turn on the computer, how is the operating system first activated and loaded into RAM?). These questions require the learner to construct approximately 3-7 sentences in an ideal answer and to exhibit reasoning in natural language. However, when students are
asked these challenging questions, their initial answers are typically only 1 or 2 sentences in length. Therefore, AutoTutor engages the student in a mixed-initiative dialogue that draws out more of what the student knows and assists the student in the construction of an improved answer. AutoTutor provides feedback on what the student types in (e.g. “good job”, “not quite”), pumps the student for more information (e.g. “What else”), prompts the student to fill in missing words (e.g. “What about X”), gives hints (e.g. “What about X”), fills in missing information with assertions (e.g. “X is Y”), identifies and corrects misconceptions and erroneous ideas, answers the student’s questions, and summarizes topics. Figure 1 shows an excerpt of the dialogues from an actual tutorial session along with the AutoTutor interface.

The tutorial sessions yielded 28 student-tutor dialogue transcripts that encompassed 164 problems and 1637 student and tutor turns. Two sets of dialogues were obtained from the transcripts. The first set, called student dialogues, were obtained by only considering the student turns in each transcript. For example, for the student-tutor exchange presented in Figure 1B, the student dialogue would include items 2, 4, and 6. The second set, called tutor dialogues, consisted of the tutor’s responses and would include items 1, 3, 5, and 7 for the sample dialogue in Figure 1B.

![AutoTutor Interface and Sample Tutorial Dialogue](image)

Figure 1. (A) AutoTutor interface and (B) Sample dialogue from a tutorial session

The purpose of dividing the transcripts into separate student and tutor dialogues is to assess the impact of each dialogue type in predicting learners’ affective states. The student dialogues are representative of the effects that their affective states have on the responses they provide, while the tutor dialogues provide insights into potential causes of the affective states. For example, incessant repetition by the tutor might cause the student to experience boredom. On the other hand, long responses by the student might be a textual manifestation of heightened engagement. By considering each dialogue type separately, we can assess the relative impact of students’ versus tutor’s responses on the student’s affective states. We are also in a position to ascertain whether there is a dialogue type × affective state interaction. That is, are particular states best predicted from student dialogues while others are best predicted from tutor dialogues.

The student and tutor dialogues for each problem were independently submitted to Coh-Metrix 2.0 and 25 predictors were obtained for each dialogue type (50 in all). An aggregated score for each predictor was derived for each subject by averaging across problems.
Annotations Tutorial Sessions for Affective States

Similar to a video-based cued-recall procedure (Rosenberg & Ekman, 1994) the judgments for a learner’s tutoring session proceeded by playing a video of the face along with the screen capture video of interactions with AutoTutor on a dual-monitor computer system. The screen capture included the tutor’s synthesized speech, printed text, students’ responses, dialogue history, and images, thereby providing the context of the tutorial interaction.

Judges were instructed to make judgments on what affective states were present in each 20-second interval at which the video automatically stopped. The affective states were boredom, flow (engagement), confusion, frustration, delight, surprise, and neutral. These states were the prominent emotions in previous studies with AutoTutor and other learning environments (Baker, D’Mello, Rodrigo, & Graesser, in review; Craig, Graesser, Sullins, & Gholson, 2004; D’Mello, Craig, Sullins, & Graesser, 2006; Graesser et al., 2006). Judges were also instructed to indicate any affective states that were present in between the 20-second stops along with the time of the observation.

Four sets of emotion judgments were made for the observed affective states of each participant’s AutoTutor session. First, for the self judgments, the learner watched his or her own session with AutoTutor immediately after having interacted with the tutor. Second, for the peer judgments, each learner came back a week later to watch and judge another learner’s session. Finally, there were two trained judges who judged all the sessions separately. The trained judges were undergraduate research assistants who were trained extensively on AutoTutor’s dialogue characteristics and how to detect facial action units according to Ekman’s Facial Action Coding System (FACS) (Ekman & Friesen, 1978).

The affect judgment procedure yielded 2,967 self judgments, 3,012 peer judgments, and 2,995 and 3,093 judgments for the two trained judges. We evaluated the reliability by which the affective states were rated by the four judges. For each judge, we computed a vector of proportion scores for each participant, which corresponded to the proportion of observations in each of the cognitive-affective states, with values adding to 1.00 for all of the states experienced by a participant. Interjudge reliability scores were computed from these proportion scores, using Cronbach’s alpha. The reliability scores across the four judges were moderate to high among the six cognitive-affect states, with values of .680, .788, .876, .595, .533, and .659 for boredom, confusion, delight, engagement/flow, frustration, and surprise, respectively.

It should be noted that our use of multiple judges is justified by the fact that there is no clear gold standard to declare what the learner’s cognitive-affective states truly are (Graesser et al., 2006). Is it the self, the untrained peer, the trained judges, or physiological instrumentation? A neutral, but defensible position is to independently consider ratings of the different judges, thereby allowing us to examine patterns that generalize above and beyond individual differences among the judges. This strategy was adopted in the analyses described in this chapter.

RESULTS AND DISCUSSION

The major goal of the analyses was to find a set of predictors that are most diagnostic of learners’ affective states. When averaged across all judges the proportions were as follows: neutral (.312), confusion (.239), flow (.161), boredom (.167), frustration (.072), delight (.032), surprise (.017). We focused on constructing models that predict boredom, flow, confusion, and frustration because these were the major affective states that learners experienced.

Selecting Predictors that Generalize Across Judges

The fact that multiple judges were used to infer the affective states of the learner causes some complications in the construction of the regression models. Although it is possible to construct models on the basis of each judge’s affect ratings, the utility of such models is unclear. For example, consider a situation where feature X is the most diagnostic of emotion E, when E is measured by the self. But feature Y predicts E when the measurement of E is provided by the peer, and Z predicts E when E is measured by one of the trained judges. The fact that different features (X, Y, Z) predict the same emotion when
measured by different judges suggests that the relationship between these features and the affective state $E$ are dependent on the affect judge providing the judgment for $E$. But if feature $M$ correlates with $E$ irrespective of whether $E$ is measured by the learner (i.e. self judgments), a peer, or trained judges, then the relationship between $M$ and $E$ transcends any individual judge bias.

We reduced our set of predictors to cohesion features that significantly correlated with at least two of the four ratings of the affect judges. The requirement that predictors have to significantly correlate with the affective states ensures that only diagnostic predictors are considered for further inspection. By ensuring that the predictors need to correlate with affect scores of at least half of the judges ensures that, to some extent, they generalize across judges. This procedure narrowed the landscape of potential predictors to two predictors each for boredom, confusion, and flow. There were three potential predictors for frustration. Therefore, the large feature set of 50 predictors (25 for each dialogue type) was effectively reduced to nine potential predictors.

**Multiple Regression Models**

Our analyses proceeded by constructing separate multiple regression analyses for each emotion. The dependent variable for each multiple regression analysis was the proportional occurrence of the emotion averaged across the four judges. The independent variables were two or three of the diagnostic predictors of each emotion (see above). Stepwise regression methods (Hocking, 1976), which incorporate a combination of forward and backward variable entry techniques, were used to isolate individual predictors or combinations of predictors that yielded the most robust models. The stepwise procedure was used due to the exploratory nature of the present analyses.

Significant ($p < .01$) one-predictor models were discovered for the four affective states. $R^2$ adj. values were .246, .223, .330, and .258 for boredom, confusion, flow, and frustration, respectively. Hence, it appears that the cohesion predictors explained 26% of the variance averaged across the four affective states. This is consistent with a large effect ($f^2 = .35$) for statistical power of .8 (Cohen, 1992), and supports the hypothesis that it is possible to predict the learners’ affective states by monitoring cohesion in tutorial dialogues. We now turn our focus to the coefficients of the regression models.

**Boredom.** It appears that bored students tend to use a significant amount of negations ($\beta = .525^3$). These included a general use of negations (e.g. no, never) as well as specific expressions such as “I don’t care” or incessantly repeating “I don’t know”. Table 2 list excerpts from participants with high and low boredom levels. Note that the highly bored student makes no attempt to answer the tutor’s question and responds with frozen expressions instead of domain related contributions.

**Confusion.** Confusion is marked by a breakdown in pronoun referential cohesion by the tutor ($\beta = -.505$). Recall, that this predictor measures the proportion of pronouns that have a grounded reference. Reading ease and comprehension are directly proportional to the proportion of grounded pronouns (Jurafsky & Martin, 2008). Hence, it is no surprise that tutorial dialogues that have a higher proportion of ungrounded pronouns are linked to heightened confusion.

**Engagement/Flow.** It appears that causally cohesive responses (i.e. causal ratio) accompany the experience of heightened engagement or flow ($\beta = .598$). The ability of learners to produce such responses indicates that they are able to construct a causal network linking causal events and objects (Zwaan & Radvansky, 1998). Such a network is essential for learning at deeper levels of comprehension and engaged learners use this mental representation to produce causally cohesive responses to the tutor’s questions.

Table 3 presents sample dialogues from learners with low and high engagement levels. Note that the unengaged learner provides a few key words and phrases instead of well articulated cohesive responses.

---

3 A $\beta$ coefficient of .525 indicates that a 1 sigma (standard deviation) increase in the incidence of negations leads to a .525 sigma increase in the proportion of boredom.
In contrast, responses by the highly engaged learner are relatively well formed sentences with particularly high causal cohesion.

Table 2. Tutorial dialogue excerpts for students with low and high boredom levels

<table>
<thead>
<tr>
<th>Boredom Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong>: Pr[Boredom] = .384</td>
</tr>
</tbody>
</table>

| [S] | I have no idea. | [T] | What operations do utility programs perform? |
| [T] | Could you elaborate on that a little? | [S] | Saving to a disk |
| [S] | I can't | [T] | Well. What role do utility programs play in copying files? |
| [T] | Okay. How can the Intranet link to the Internet? | [S] | It sends them to the cpu |
| [S] | I haven't the slightest idea. | [T] | Wrong. Try this. Utility programs are used to copy files |
| [T] | No that is only partly right. See if you know the answer to this. With a gateway connection you can connect from an intranet to the? | [S] | Utility programs are used to copy files |
| [T] | That is correct! the Internet. Let's keep going. Why is an intranet more secure than the Internet? | [S] | I don't know |

Table 3. Tutorial dialogue excerpts for students with low and high engagement levels

<table>
<thead>
<tr>
<th>Engagement/Flow Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong>: Pr[Flow] = .407</td>
</tr>
</tbody>
</table>

| [T] | Why is it important to consider the characteristics of the CPU? | [T] | What operations do utility programs perform? |
| [S] | It also does the calculations for the computer | [S] | Copying files |
| [T] | Yeah. Why shouldn't you purchase a computer that runs Microsoft Windows if you want to run Macintosh programs? | [T] | Kind of. What role do utility programs play in copying files? |
| [S] | Because the CPU's would get crossed up and create a virus | [S] | Don't know |
| [T] | No. Sorry that isn't quite the answer. What about this? If you have a computer with Microsoft Windows you may not be able to run? | [T] | Okay See if you know the answer to this. Utility programs are used to copy |
| [S] | Cpu | [T] | Alright Let's try this. Some utility programs help the user copy files between? |
| [S] | By speeding up any functions that you engage in okay Try this. The newer the program the faster the? | [T] | Rom and Ram |
Frustration. Frustration is predicted by a lack of noun co-reference cohesion in the student dialogues ($\beta = -.536$). These results suggest that students in states of heightened frustration construct responses that have cohesive gaps. This might be because they are devoting a significant amount of their cognitive resources in experiencing and regulating their frustration rather than on constructing meaningful responses to the tutor’s questions.

In the example dialogue segments of low and high frustrated learners presented in Table 4, there is no cohesiveness in the responses of the highly frustrated learner across turns. Each turn provides a fragmented piece of information that has little or no relationship with the previous turn. In contrast, the learner that is not frustrated is incrementally providing a cohesive, albeit poor answer, across turns.

Table 4. Tutorial dialogue excerpts for students with low and high frustration levels

<table>
<thead>
<tr>
<th>Frustration Level</th>
<th>High: Pr[Frustration] = .139</th>
<th>Low: Pr[Frustration] = .005</th>
</tr>
</thead>
<tbody>
<tr>
<td>[T]</td>
<td>How is video-conferencing and teleconferencing delivered?</td>
<td>[T] On what type of computer would you need an operating system that is stored permanently in ROM?</td>
</tr>
<tr>
<td>[S]</td>
<td>It is delivered through streaming data from one camera to the computer of the receiving party</td>
<td>[S] A personal computer as well as public computers would need an operating system that is stored permanently in ROM</td>
</tr>
<tr>
<td>[T]</td>
<td>No. I think you can get this. You can communicate with several people at once using?</td>
<td>[T] Umm. I bet you know the answer to this. An operating system that is stored entirely on ROM is useful for?</td>
</tr>
<tr>
<td>[S]</td>
<td>Multiple programs</td>
<td>[S] An operating system that is stored entirely on ROM is useful for compact disks and floppy disks.</td>
</tr>
<tr>
<td>[T]</td>
<td>Nah that's not the full answer. Using teleconferencing. How do you have on line conversations?</td>
<td>[T] Moving on. Some computers store their operating systems in ROM. Where is the operating system on special purpose computers?</td>
</tr>
<tr>
<td>[S]</td>
<td>I message people through an instant message program</td>
<td>[S] The operating system on special purpose computers is stored in the ROM.</td>
</tr>
</tbody>
</table>

Generalizing Across Domains

One potential concern with the multiple regression models is that some of the effects might be linked to our particular tutorial domain, namely computer literacy with AutoTutor. This is an important concern that would adversely impact the generalizability of our results. We partially addressed this question by assessing whether topic differences in computer literacy (i.e. hardware, internet, operating systems) affected the regression models. We conducted a follow-up analysis by constructing two-step multiple regression models for each emotion. Step 1 predictors consisted of dummy coded variables for the three computer literacy sub-topics, whereas Step 2 predictors were the diagnostic predictors of each affective state (see Table 1).

The results indicated that none of the Step 1 models were statistically significant while the Step 2 models were all statistically significant and the significant predictors listed in Table 1 were also significant in the Step 2 models. Therefore, we can conclude that our set of predictors are diagnostic of the learners’ affective states above and beyond differences in computer literacy subtopics, and might generalize to other domains as well. However, these predictors will have to be validated on corpora from other domains and tutoring systems before we can be assured of their generalizability.

FUTURE DIRECTIONS

This chapter explored the possibility of detecting learners’ affective states by monitoring the cohesiveness of student and tutor dialogues. Although learners’ do not directly express their affective states to the tutor, our results indicate that these states can be monitored by analyzing various measures of cohesion. This
serves as an important concern. The next step of this research is to implement real-time cohesion-based affect detectors. The current set of regression models were constructed at the subject level as the primary goal of the analyses was to explore the possibility of deriving a set of diagnostic predictors that generalized across affect judges. However, these models can be extended to predict affective experiences as they occur by analyzing incremental windows of student and tutor dialogues that are generated as the tutoring session progresses.

We are also in the process of combining a cohesion-based analysis of tutorial dialogues with a simple keyword matching approach such as LIWC (Pennebaker et al., 2001). There is the important question of assessing whether there are any advantages to the deeper cohesion-based analysis or if shallower methods will suffice. More interestingly, there is the question of whether a fusion of features from Coh-Metrix and LIWC will resonate with superadditive or redundant effects. When there are superadditive effects, the combination of features explains more variance than what would be expected from a simple additive effect. Redundancy would occur when models constructed from a combination of Coh-Metrix and LIWC features result in negligible improvements over models that consider these feature sets individually.

CONCLUSIONS

The broader implications of this research venture into the area of affective computing and the design of a version of AutoTutor that is sensitive to learners’ cognitive and affective states. For example, if the learner is frustrated, the tutor can give hints to advance the learner in constructing knowledge or can make supportive empathetic comments to enhance motivation. If the learner is bored, the tutor needs to present more engaging or challenging problems for the learner to work on. Whereas when the learner is in a state of flow, the tutor would probably want to lay low and stay out the learner’s way (Csikszentmihalyi, 1990).

Pedagogical interventions when confusion is detected are interesting because confusion is believed to play an important role in learning and has a large correlation with learning gains (Craig et al., 2004; Graesser et al., 2007). When the learner is confused, there might be a variety of paths for the tutor to pursue. The tutor might want to keep the learner confused in order to force the learner to stop and think and try to resolve the source of the confusion. Alternatively, after some period of time waiting for the learner to progress, the tutor might give indirect hints to nudge the learner into more productive trajectories of thought.

We have recently implemented some of these strategies in a new version of AutoTutor that detects and responds to learners’ boredom, confusion, and frustration by monitoring facial features, gross body language, and conversational cues (D’Mello, Picard, & Graesser, 2007; D’Mello, Craig, Fike, & Graesser, 2009). The sensed cognitive-affective states are used to select AutoTutor’s pedagogical and motivational dialogue moves in a manner that optimally coordinates learners’ cognitive and affective states in order to heighten engagement and enhance learning gains.

Although initial evaluations of the affect-sensitive AutoTutor are promising, one disadvantage of some of these sensors used for affect detection is that they require expensive customized hardware and software, such as the Body Pressure Measurement System (Tekscan, 1997) and automated facial feature tracking systems. This raises some scalability concerns for those who want to extend this program of research beyond the lab and into the classroom. It is in the applied context of the classroom where text-based affect detectors, such as the cohesion-based system described in this chapter, have a unique advantage over behavioral and physiological sensors. Text-based affect sensing is advantageous because it is cost-effective, requires no specialized hardware, is computationally efficient, and is available to any ITS with conversational dialogues. Hence, although text-based affect detectors currently relinquish the center stage to the more popular behavioral and physiological approaches, they are expected to play a more significant role in next-generation affect detection systems, particularly when efficiently, cost-effectiveness, and scalability are important concerns. Whether fully automated cohesion-based affect
detectors can compliment or even replace existing systems that monitor physiological and bodily measures awaits future research and empirical testing.

ACKNOWLEDGEMENTS

The research on was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918, ALT-0834847, DRK-12-0918409), the Institute of Education Sciences (R305G020018, R305H050169, R305B070349, R305A080589, R305A080594), and the Department of Defense Counter Intelligence Field Activity (H9C104-07-0014). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD.

Requests for reprints should be sent to Sidney D’Mello, Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152; sdmello@memphis.edu.

REFERENCES


D’Mello, S., & Graesser, A. (in review). Language and discourse are powerful signals of learners’ cognitive-affective states during tutoring.


Williams, C., & D’Mello, S. (in press). Predicting student knowledge levels from domain-independent function and content words In J. Kay & V. Aleven (Eds.), *Proceedings of 10th International Conference on Intelligent Tutoring Systems*. Berlin / Heidelberg: Springer.


**ADDITIONAL READING SECTION**


**KEY TERMS & DEFINITIONS**

Affect-sensitive interface: A computer or robotic system that detects and responds to human users’ affective states (e.g., frustration, anger, surprise) to assist in the facilitation of some task of relevance to the human user.

AutoTutor: An intelligent tutoring system with conversational dialogues.

Boredom: State of being weary or restless through lack of interest.

Cohesion: Cohesion, a textual construct, is a measurable characteristic of text that is signaled by relationships between textual constituents.

Coherence: Coherence is a psychological construct that is a characteristic of the text together with the reader’s mental representation of the substantive ideas expressed in the text.

Emotional Intelligence: Skill or ability to perceive, use, understand, and regulate emotions in oneself and in others.

Confusion: State of having a noticeable lack of understanding.

Engagement/flow: State of interest that results from involvement in an activity.

Frustration: Dissatisfaction or annoyance from being stuck when goals are blocked.

Intelligent Tutoring System: Artificially intelligent computer system that tutors human students by providing customized explanations, direct instruction, and feedback.
Learning-centered affective states: Affective states such as boredom and frustration that occur in diverse learning contexts. These can be distinguished from the “basic emotions” (e.g., anger, sadness) which have little relevance to learning.