Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States

Sidney D’MELLO a,1, Nia DOWELL b, and Art GRAESSER b

a Department of Computer Science, University of Memphis, USA
b Department of Psychology, University of Memphis, USA

Abstract. We explored the possibility of predicting learners’ affective states (boredom, flow/engagement, confusion, and frustration) by monitoring variations in the cohesiveness of tutorial dialogues during interactions with AutoTutor, an intelligent tutoring system with conversational dialogues. Multiple measures of cohesion (e.g., pronouns, connectives, semantic overlap, causal cohesion, coreference) were automatically computed using the Coh-Metrix facility for analyzing discourse and language characteristics of text. Cohesion measures in multiple regression models predicted the proportional occurrence of each affective state, yielding medium to large effect sizes. The incidence of negations, pronoun referential cohesion, causal cohesion, and coreference cohesion were the most diagnostic predictors of the affective states. We discuss the generalizability of our findings to other domains and tutoring systems, as well as the possibility of constructing real-time, cohesion-based affect detectors.

Keywords. affect, affective states, emotion, discourse, cohesion, coreference, Coh-Metrix, intelligent tutoring systems, AutoTutor

Introduction

Affect-sensitive learning environments have emerged as a result of the recent research exploring connections between affect and deep learning [1-3]. In addition to adapting pedagogical strategies to the learners’ cognitive states, affect-sensitive intelligent tutoring systems (ITSs) are designed to adapt to learners’ affective states. These ITS’s incorporate state-of-the-art affect-sensing devices to recognize the learners’ affective states and tailor their pedagogical and motivational strategies to be responsive to these states [4-6].

The aspiring affect-sensitive ITSs leverage recent advances in affective computing [7] to detect the learners’ affective states. The early systems primarily monitored physiological measures, such as skin conductance and heart rate, or bodily measures, such as facial actions, acoustic-prosodic features of speech, and gross body language. However, there is an alternative to using physiological and bodily measures for affect detection, namely 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. Second, textual features derived from tutorial

1 Corresponding Author. 209 Dunn Hall, Computer Science Department, University of Memphis, Memphis, TN 38152, USA. E-Mail: sdmello@memphis.edu
dialogues are contextually constrained in a fashion that provides cues regarding the  
social dynamics of the student and tutor. It is plausible to expect a textual analysis of  
tutorial dialogues to provide insights into learners’ affective states. Recent advances in  
computational psycholinguistics has convincingly demonstrated that textual features  
can predict complex phenomenon such as personality, deception, and even physical and  
mental health outcomes [8-10].

One ITS has successfully fused lexical and acoustic-prosodic information for  
affect detection [11], but extended discourse-based affect detection has not been  
previously automated in learning environments. A number of research groups have  
proposed domain-independent, text-based, affect detectors. These systems operate by  
constructing affective models from large corpora of world knowledge. The models are  
used to predict the affective tone of segments of text such as movie reviews, product  
reviews, blogs, and email messages [12-14]. However, these 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, it is unclear whether learners’ responses  
to computer tutors resonate with affect rich content. In fact, there is some evidence to  
the contrary. An examination of 1637 student responses generated from a tutorial  
session with AutoTutor, an ITS with conversational dialogues [15], yielded only a  
handful of utterances with explicit affective expressions (< 1%). But an in-depth  
analysis of videos of the tutorial sessions yielded approximately 3000 affective  
experiences [2]. Although students are experiencing affective states while interacting  
with AutoTutor, their typed responses do not necessarily convey affective content  
explicitly. Instead their responses mainly consist of domain specific answers to the  
tutor’s questions even when they are in the midst of rich affective experiences.  
Therefore, a more systematic textual analysis of tutorial dialogues might be necessary  
to uncover subtle cues that might be diagnostic of a learners’ affective states.

This hypothesis was investigated in the current paper by analyzing cohesion  
relationships in naturally occurring tutoring dialogues with AutoTutor. Cohesion, a  
textual construct, is a measurable characteristic of text that is signaled by relationships  
between textual constituents [16, 17]. 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 [17]. Therefore, we  
hypothesize that variations in the cohesiveness of the tutorial dialogues will be  
predictive of the learners’ affective experiences. We expect a breakdown in cohesion to  
be predictive of affective states such as confusion and frustration, while strong  
cohesive relationships might be indicative of engagement.

1. Methods

1.1. Participants, Materials, and Procedure

The participants were 28 undergraduate students from a mid-south university in the U.S.  
who participated for extra course credit. Participants interacted with AutoTutor for 32  
minutes on one of three randomly assigned topics in computer literacy: hardware,  
Internet, or operating systems. During the tutoring session, a video of the participant’s
face, their posture pressure patterns (not elaborated here), and a video of their computer screen were recorded for offline analyses.

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, pumps the student for more information, 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.

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) [18]. The judgments for a learner’s tutoring session proceeded by playing a video of the face along with a screen capture video of interactions with AutoTutor. 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 [2, 19]. Participants were also instructed to indicate any affective states that were present in between the 20-second stops along with the time of the observation. The affect judgment procedure yielded 2967 self judgments, 3012 peer judgments, and 2995 and 3093 judgments for the two trained judges.

### 1.2. Extracting Transcripts of Tutorial Dialog

Transcripts of the tutorial dialogue between the student and the tutor were extracted for each problem that was collaboratively solved during the tutorial session. Two sets of dialogues were obtained from the transcripts. One focused on the student turns while the other focused on the tutor turns. These two dialogue viewpoints presumably offer different perspectives on the learners’ affective experiences and expressions. The student dialogue turns reflect how their affective states influence their responses, whereas the tutor dialogue turns reveal events that trigger or maintain the student’s affective states. For example, verbose but cohesive responses by the student might be a textual manifestation of heightened engagement. On the other hand, incessant repetition by the tutor might cause the student to experience boredom. Simply put, the students’ responses are hypothesized to be consequents of their affective states, while the tutor’s responses are potential antecedents of student affect.
2. Measures of Cohesion

The Coh-Metrix program was used to compute multiple measures of cohesion from the dialogues [17]. Coh-Metrix is a validated computational tool that provides over 100 measures of various types of cohesion. The indices of cohesion used in the study are briefly described below.

Co-reference cohesion 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. Here, 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 adjacent sentences with noun and stem overlap across window sizes of 2, and 3 sentences.

Pronoun referential cohesion 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 [16]. A suitable match occurs if the pronoun agrees with referent in gender and number. For example, consider the following sentences: (1) The user's 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. Referencing 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 [20], so Coh-Metrix uses an approximate algorithm to compute pronoun referential cohesion.

Causal cohesion occurs when constituents in the text that signal events and actions are connected by causal particles [17, 21]. 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 [22]. 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 these verb categories and causal particles (occurrences per 1000 words). But the most significant measure of causal cohesion is the causal ratio that specifies the ratio of causal particles to events and actions. The causal ratio of a text is directly proportional to its cohesiveness [16].

Semantic cohesion is measured as the conceptual similarity between adjacent text constituents. LSA [23] is used as the primary computational tool for measuring semantic cohesion. LSA is a statistical technique that measures the conceptual similarity between two texts (e.g., word, sentence, turn, paragraph, document). Adjacent sentences that have higher LSA cosine scores (i.e. higher semantic similarity) are more cohesive than adjacent sentences that have low LSA scores. Coh-Metrix computes several measures of semantic cohesion at varying window sizes. The current analyses considered mean and standard deviations of LSA scores for adjacent sentences in a student or tutor turn, and across adjacent turns.

Connectives are textual constituents that play a significant role in signifying cohesion and coherence relationships by explicitly linking ideas expressed in a text [16, 17, 24]. Coh-Metrix provides incidence scores on several types of connectives in a text. These include temporal (before, when), additive (also), causal (because); these
categories have both negative valences (however, in contrast) and positive valences (therefore, in addition) [24].

Shallow Measures of Readability and Verbosity. These measures are provided in addition to the primary cohesion indices. The verbosity measures include the number of: words in a sentence, sentences in a turn, turns in a problem, syllabus per word, and many others. There are also measures of readability including the Flesch Kincaid Grade Level and the Flesch Kincaid Reading Ease [25].

The student and tutor dialogues for each problem were independently submitted to Coh-Metrix 2.0 and 25 problem-level predictors were obtained for each dialogue type. An aggregated score for each predictor was derived for each subject by averaging across problems. There were 4 measures of co-reference cohesion, 1 measure of pronoun referential cohesion, 2 measures of causal cohesion, 5 measures of semantic cohesion, 6 measures for connectives, and 7 basic indices. 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 [16].

3. Results and Discussion

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

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 (see[2] for a discussion on interrater reliability). For example, consider a situation where features X, Y, and Z are individually diagnostic of emotion E, when E is measured by the self, peer, and trained judges respectively. Different features (X, Y, Z) that 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 judges. But if feature M correlates with E irrespective of the affect judge, 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.

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 [26], 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 resulted in the selection of one diagnostic predictor for each affective state.

The parameters of the multiple regression models are listed in Table 1. It appears that the cohesion predictors explained 26% of the variance averaged across the four affective states. This is consistent with a medium to large effect for a statistical power of .8 [27], and supports the hypothesis that it is possible to predict the learners’ affective states by monitoring cohesion in tutorial dialogues.

### Table 1. Parameters of multiple regression models.

<table>
<thead>
<tr>
<th></th>
<th>Boredom</th>
<th>Confusion</th>
<th>Flow</th>
<th>Frustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td>9.51</td>
<td>8.20</td>
<td>12.40</td>
<td>9.68</td>
</tr>
<tr>
<td>( df_1, df_2 )</td>
<td>1, 25</td>
<td>1, 24</td>
<td>1, 24</td>
<td>1, 24</td>
</tr>
<tr>
<td>( p )</td>
<td>.002</td>
<td>.009</td>
<td>.001</td>
<td>.005</td>
</tr>
<tr>
<td>Adj. R Sq.</td>
<td>.246</td>
<td>.223</td>
<td>.330</td>
<td>.258</td>
</tr>
</tbody>
</table>

**Standardized Coefficient Weights (\( \beta \))**

- \([S]\) Incidence of negations: .525
- \([T]\) Pronoun referential cohesion: -.505
- \([S]\) Causal Ratio: .598
- \([S]\) Noun overlap adjacent sent.: -.536

Notes. \([S]\) and \([T]\) refer to features derived from student and tutor dialogues respectively. \(^*\)Sent. = Sentences.

Turning our focus to the significant predictors of the regression models, it appears that bored students tend to use a significant amount of negations. 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”.

Confusion is marked by a breakdown in pronoun referential cohesion by the tutor. 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 [20]. Hence, it is no surprise that tutorial dialogues that have a higher proportion of ungrounded pronouns are linked to heightened confusion.

It appears that causally cohesive responses accompany the experience of heightened engagement (or flow). The ability of learners to produce such responses indicates that they are able to construct a causal network linking causal events and objects [21]. 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.

Finally, frustration is predicted by a lack of noun co-reference cohesion in the student dialogues. 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.

### 4. General Discussion

This paper explored the possibility of detecting a 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 suggests that it
takes a more systematic and deeper analysis of dialogues to uncover diagnostic cues of
the learners’ affective states. It is also interesting to note that it takes an analysis of
both the student and the tutor dialogues to obtain a set of predictors that are diagnostic
of the entire set of learning-centered affective states.

One potential concern with the multiple regression models is that some of the
effects might be linked to our particular tutorial domain, i.e. 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 analyses 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, while
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.

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. Whether fully automated cohesion-based affect detectors can compliment
or even replace existing systems that monitor physiological and bodily measures [4]
awaits future research and empirical testing.

Acknowledgments

This research was supported by the National Science Foundation (REC 0106965, ITR
0325428, and REC 0633918), Institute of Education Sciences (R305G020018), and the
opinions, findings and conclusions or recommendations expressed in this paper are
those of the authors and do not necessarily reflect the views of funding agencies.

References

reengineering educational pedagogy-building a learning companion," in Advanced Learning
of Emotions during learning with AutoTutor," in Proceedings of the 28th Annual Conference of the


