Discerning Affect in Student Discussions

Jihie Kim, Erin Shaw, Saul Wyner, Taehwan Kim, and Jia Li
(jihie@isi.edu)
University of Southern California/ Information Sciences Institute
Marina del Rey, CA 90292 USA

Abstract

Students' emotions and attitudes are discernible in messages posted to online question and answer boards. Understanding student sentiment may help instructors identify students with potential course issues, optimize help-seeking, and potentially improve student achievement, as well as identify both positive and negative actions by instructors and provide them with valuable feedback. Towards this end, we present a set of context-independent emotion acts that were used by students in a university-level computer science course to express certainty and uncertainty, frustration, and politeness in an online Q&A board and develop viable classification approaches. To explore the potential of sentiment-based profiling, we present a heuristic-driven analysis of thread resolution and detail future research.

Keywords: student online discussions, discourse analysis.

Introduction

Online discussion boards are widely used in higher education, extending the availability of instructors, assistants, and materials to students beyond the traditional classroom. Students use discussion forums to collaborate, exchange information, and seek answers to problems from their instructors and classmates. Discussion board use is generally associated with improved academic performance and greater student satisfaction (Kumrow, 2005; Newman and Schwager, 1995).

Previous work on analyzing student discussions has been based on rhetorical speech acts, course topics, and problem tasks (Kim et al., 2007; McLaren et al., 2007). Classification systems for these features enable researchers to automatically identify student problems. Similarly, understanding student affect may help instructors identify students with potential course issues, optimize help-seeking, and potentially improve student achievement. In addition, by examining the results of different instructor-student interactions in terms of affect, instructors could potentially receive valuable feedback about their online interactions.

In this paper we present a set of dialogue features, or *emotion acts* (EAs), analogous to Speech Acts (Searle, 1969), that characterize student sentiment with respect to 1) frustration and tension, 2) high and low certainty (confidence) and 3) politeness. These sentiments were exhibited in student discourse within a question and answer (Q&A) board in an undergraduate Computer Science course. A discussion corpus consisting of 1,030 student posts was manually tagged with the emotion acts.

We describe the first stages of the development of practical classification for emotion acts and explore the potential of sentiment-based student profiling. Specifically, in this paper, we do the following:

1. Identify categories of affect exhibited in an online student discussion in an undergraduate CS course.
2. Examine the frequency of affect in the corpus by gender, role and types of participants.
3. Examine the influence of affect in instructor messages on student responses (discerned by affect).
4. Examine the correlation between affect and type of thread (resolved or unresolved).
5. Illustrate of how emotion acts can be used in assessing and predicting student discussion outcome.
6. Describe our approach to and initial results of automatically classifying three categories of affect.

Identifying Categories of Affect

It is extremely difficult to devise a category of affect labels given the gradations and subtlety of the way feelings and emotions are expressed in language. It is not surprising then that there is no general agreement on how to label affective content and that instead there exist a number of different labeling schemes for different domains (Ordelman and Heylen, 2005). However, previous work suggests that at least some affective content can be identified and selected for, independent of context. For example, acknowledgements are recognizable by the presence of common *politeness* phrases (Brown & Levinson, 1987), and may be used to indicate resolution in Q&A discussion; and *certainty* categorization was shown to assist in distinguishing between editorial and news writing (Rubin et al., 2006), and may be used to distinguish questions and answers by the presence and absence of student confidence.

Identifying a set of categories was an iterative process, and there were three criteria for selection: a) category examples had to be well represented in the corpus, b) researchers had to agree on the categories, and c) categories had to be relevant to student learning. Selection was originally motivated by the desire to identify students' self-efficacy and attitudes, although these categories were too abstract to be practical. We examined discourse that indicated confidence, interest and mastery, and also, urgency, understanding and technicality. Our final categories were *high and low certainty (confidence)*, *tension/frustration*, and *politeness*.

Tension (kappa: 0.74)	Examples
Instructor Judgments: Possible student issues with class attendance, judgment or choices	If you really want to do this; I stated in class on at least 2 occasions
Student Judgments: Possible student issues with questioner or target	Result of this sucks; Wow… That was..
Frustration (kappa: 0.92)	**Examples**
Repetitious Actions, Continual Actions: Descriptions of continuous actions without real progress	A lot (15+ times); Never seems to end; High rate of redundancy; Another can of worms
Large Quantities: Descriptions of overwhelming amounts of work and other material	Zillions of references; Super-huge; Simply gargantuan; Monstrous, super-verbose
Difficulty/Impassability, Material Denigration: Statements of explicit difficulty in either solution or understanding of issues, as well as frustration about the material itself	Serious disk quota problems; Severe annoyances; A pain to fix; Makes it really hard
Self-Denigration/Lack of Confidence: Declarations of a personal belief in a lack of ability on the part of the poster	I have spent FAR too long; …I'm stumped; Longer than they should have
High Certainty (kappa: 0.80)	**Examples**
Specificity of Question/Answer: Specific phrasing that concisely explains through examples and pre-conditions	The only way; I found the answer; It only appears
Ease of Understanding/Completeness: Emphasis of the simplicity or completeness of a solution or question	The trick is; Just wait till; Will be simple; All you need to do is
Necessity: Specifically stating that the presented solution is required, or in the case of a question, its importance	Must be able to; Vitally important task; Must have something; You will
Logical Presentation: A method of presenting a proposition, solution, or question that makes it a logical proposal	I assume that; Granted,; Likewise,; On the other hand,
Low Certainty (kappa: 0.95)	**Examples**
Vagueness in Question/Answer: Statements that imply only general or surface understanding of the material at hand by stating personal understandings over factual presentation	What is wrong?; If I understand; Seems to me; Read it somewhere
Lack of Understanding: Statements that clearly state a lack of understanding; differs from other Speech Acts as it implies a continuing lack, rather than an individual issue	I am still confused; Not sure if I understand; I follow most; I'm not sure
Optional Nature: Statements indicating a not strongly recommended or vital issue, solution, or question	Should be compiled from the network directory; You might try; …maybe I'll try making; What is wrong?
Weakened Presentation: Phrases that weaken or justify logical proposal statements	Correct me if I am wrong; Apparently; I am guessing that is the way; As far I know
Politeness categories	**Examples**
Positive (kappa: 0.99): Language strategies used according to formal cultural rules to avoid losing face. Commonly identified as typical polite speech	Thanks; Okay thanks; Good luck with your project
Negative (kappa: 0.99): Dealing with a face-threatening act, by lightening the request or response into a less pressing, informal status.	I was wondering if; Thought I'd throw this out there; Get this cleared up early; Just a head's up,
Bald on record (kappa: 0.84): Dealing with a face-threatening situation by ignoring or emphasizing the consequences of the threat	I question the; don't bzero anything; Change it to this; Do we?
Off record (kappa: 0.82): Attempting to change the request or response into a non-face-threatening statement, i.e., by generalizing a query to a rather than asking for direct help	Has anyone else had this problem; What would do; Asking for answers directly is way easier

Table 1. Categories of affect – description and examples.

The final categories had high agreement among the research team, and thus had potential for use in an automatic classification system.

Annotation Methodology

Annotating affect involved identifying those speech fragments that reliably indicated an identified emotion act in a repeatable fashion throughout the corpus of student discussion board posts. This was complicated by the highly irregular nature of the message content, which was characterized by frequent misspellings and grammar and syntactical errors, stemming from common parlance, simple carelessness, and Computer Science student subculture language use. This necessitated a high level of selectivity and repeatability in all annotations, as well as reliance on specific patterns of distinct phrases and grammar from within the corpus rather than whole statements.

Table 1 lists and describes the final EA categories. A dataset of 1,030 messages in 210 threads from an Operating Systems course was analyzed. Several iterations were performed until we minimized ambiguity among the categories and finalized clear EA definitions. For inter-annotator agreement, we compared two annotators' data on 322 messages in 30 discussion threads. For the current categories, the kappa values are

greater than 0.7. Some Politeness EAs show very high agreement ratios since the annotators consider them very clear and there are only small numbers of cases.

The labeling process consisted of EA classification as well as the marking up of contextual information within the message content. The markup included information about the type of response (question/query or answer/statement) and the role of the author (student, instructor, or TA).

Affect Frequency by Type of Participants

The final frequency distribution of emotion acts for messages posted by different participants within the dataset is shown in Table 2. Of interest are the high occurrences of low certainty and the relatively high frequency of frustration. Female students seemed to present less frustration than male students. Also, females present more positive politeness in their messages. As expected, the instructor's messages show high confidence. Among politeness categories, the instructor presents more bold-on-record politeness (BOR) than students.

We also looked at the presence of emotion acts among high and low frequency contributors. Figure 1 shows the distribution of different emotion acts for seven groups of contributors. As can be predicted from the distribution in Table 2, confidence and polite acts dominate. For the students who post many messages, the number of other emotion acts increases, especially confidence, but also frustration and negative politeness.

Influence of Instructor Affect on Students

The course instructor participated in discussions in many ways; he provided answers directly, gave alternative perspectives, supported student ideas, and elaborated on student answers. It is useful to analyze the influence of the instructor on student dialogue.

In Table 3, we consider what happens when an instructor exhibits emotion. It appears that students tend to express more emotion themselves (certainty, frustration, negative politeness) after an instructor shows emotion. Students appear to express high certainty when they respond to an instructor's high certainty. Similarly,

student frustration and low certainty can follow the instructor's expression of low certainty.

While these results show many interesting possible relationships between expressed emotion acts and topic success, the clear and immediate indication shows that emotion acts can show distinctions between different types of posts and threads, which prove their potential usefulness as a profiling mechanism.

Table 2. Distribution of Emotion Acts among participants.

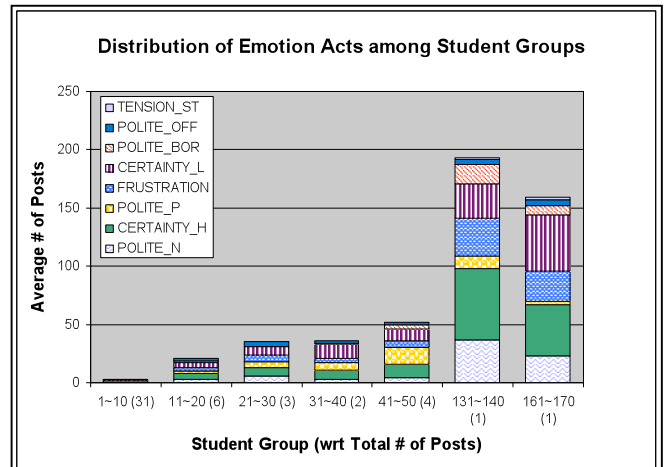| Emotion Act	Percent Emotion Acts found in messages:			
	Total (N=1179)	Male Student (N=782)	Female Student (N=62)	Instruct-or (N=300)
Tension	2%	1%	0%	6%
Frustration	14%	19%	9%	2%
Certainty_High	32%	31%	36%	35%
Certainty_Low	20%	26%	27%	4%
Politeness_Pos	13%	15%	55%	0%
Politeness_Neg	13%	18%	3%	3%
Politeness_OFF	5%	6%	11%	0%
Politeness_BOR	10%	8%	11%	16%

Figure 1: Distribution of Emotion Acts in infrequent and frequent discussion board contributors.

Table 3: Students' EAs following an instructor EAs.

Following Student EAs / Instructor EAs	#	Tension	Frustration	High_Cert	Low_Cert	Polite_Pos	Polite_Neg	Polite_OFF	Polite_BOR
		0	31	60	44	22	35	6	12
Tension	19	0	2	4	6	1	1	1	2
Frustration	7	0	2	2	1	1	2	0	0
High_Certainty	107	0	16 (15%)	33 (31%)	23 (21%)	5	20(19%)	3	9
Low_Certainty	12	0	5 (42%)	3 (25%)	4 (33%)	2	4 (33%)	0	0
Politeness_Pos	1	0	1	0	1	0	1	0	0
Politeness_Neg	11	0	1	2	1	12	2	1	0
Politeness_OFF	0	0	0	0	0	0	0	0	0
Politeness_BOR	49	0	4	16 (33%)	8	1	5	1	1

Subset for Analysis	(N)	Emotion Act Percentage (%) From Each Group									
		Certainty Level				Frus-tration	Tension	Politeness			
		High	Low	Med	N/A			Bald-On-Record	Positive	Negative	Off-Record
All Posts	(1030)	49.03	13.50	3.79	33.69	24.27	2.72	19.71	17.38	24.66	6.89
All Resolved	(916)	50.11	12.99	3.17	33.73	23.47	2.84	19.43	17.36	23.91	6.33
All Unresolved	(114)	40.35	17.54	8.77	33.33	30.70	1.75	21.93	17.54	30.70	11.40
Resolved Answers	(645)	56.12	10.23	2.95	30.70	18.60	3.72	22.79	11.16	20.31	2.02
Unresolved Answers	(79)	43.04	10.13	8.86	37.97	25.32	1.27	29.11	13.92	32.91	6.33
Resolved Questions	(271)	35.79	19.56	3.69	40.96	35.06	0.74	11.44	32.10	32.47	16.61
Unresolved Questions	(35)	34.29	34.29	8.57	22.86	42.86	2.86	5.71	25.71	25.71	22.86
Resolved Inst. Ans.	(233)	62.66	2.58	0.43	34.33	5.58	8.58	35.19	3.00	7.30	0.43
Unresolved Inst. Ans.	(25)	48.00	0.00	8.00	44.00	0.00	4.00	44.00	8.00	16.00	0.00
Resolved First Posts	(180)	33.89	17.78	2.22	46.11	33.33	0.56	13.89	37.22	32.78	20.00
Unresolved First Posts	(30)	43.33	16.67	3.33	36.67	50.00	0.00	23.33	36.67	40.00	23.33
Resolved Final Posts	(180)	45.56	16.11	4.44	33.89	17.22	2.22	17.22	26.11	24.44	5.00
Unresolved Final Posts	(30)	26.67	13.33	3.33	56.67	20.00	0.00	30.00	6.67	13.33	6.67

Table 5. The distribution of relevant emotion acts in resolved vs. unresolved threads.

Affect Patterns in Threads

While sentiment-based discussion analysis is a significant development, emotion acts represent only the lowest level of potential analysis of student message content. With consistent and functional emotion acts, posters, posts, and entire threads can be analyzed in terms of repeatable EA profiles. As a proof of concept, we wished to develop an independent heuristic to classify threads with a hypothetically robust emotional distinction, and examine the resulting EA profile for such a distinction.

We chose the concept of resolved and unresolved discussion threads, where resolved threads contain a final solution or demonstrable ratification of issues, as well as a beneficial discussion, and open threads are those for which initial questions are not satisfied or which have unresolved issues. The ultimate goal is to identify patterns of affective states that help to discern students who may require further assistance, and topics that may require further clarification. Towards this goal, we experimented with several classification measures based upon observed trends in annotated threads. To fulfill the need for a conclusion, we focused on threads that concluded with an answer, or an acknowledgement of thanks for a provided solution. To ensure a basic level of back-and-forth pedagogic discourse we included only the subset of threads that also contained equal numbers of or more answer/statement posts. The generated results by these criteria were examined by the annotators and found to closely conform to their intrinsic impressions of "resolved" threads. Those threads that were not considered resolved were classified as an "unresolved". This revealed 180 resolved, and 30 unresolved threads.

After this classification, both resolved and unresolved threads were further broken-down into relevant subsets for EA analysis. The analysis was based upon a simple presence test for specific EAs, and the percentage of posts within the subset that contained that emotion act. Certainty, however, as the most common emotion act, was instead calculated as a level, defined by containing over 75% of a specific type of either high or low certainty emotion acts. If the ratio was less than 75%, it was designated as medium certainty. While rudimentary, this examines the potential for more rigorous profiling, by revealing any obvious difference among threads.

The results show a clear distinction between resolved and unresolved threads. Distinctions were noted when there existed at 10% or above difference from resolved vs. unresolved versions of the chosen subset.

Within the certainty measures, high certainty is shown to strongly influence the resolution of a thread with respect to answers, while having little effect on questions. However, in initial posts, high certainty seems to counter-indicate resolution. In contrast, low certainty seems to have minimal effect, except in the case of questions, in which it is strongly represented in unresolved questions. A lack of certainty (both high and low) also strongly differentiates resolved and unresolved questions and initial posts, while it shows the inverse in final posts.

In terms of frustration and politeness, frustration is unsurprisingly well-represented in unresolved posts, though most notably in initial posts. Bald-on-record politeness shows strongly in unresolved instructor answers, original posts, and final posts. Positive politeness is seen greatly in resolved questions and final posts, while negative politeness is greater in resolved final posts. Off-record politeness shows little effect overall.

Automatic Affect Classification

For automatic classification of emotion acts, we followed a similar approach that was previously applied to identify speech acts in student discussions (Kim et al., 2009). We focused on certainty and frustration because they are most relevant to student performance. The annotated discussion threads were first pre-processed: Because student discussions are informal and noisy with respect to grammar, syntax and punctuation, our model fixes common typos, transforms informal words to formal words, and converts apostrophes to their original forms. It replaces some typical words and phrases with fixed keywords; for instance, programming code fragments are replaced with by CODE, and contractions such as "I'm" and "You're" were replaced with "I am" and "You are". The features used include:

F1: Cue phrase and their position in the post
We used n-gram features including unigrams (1 word), bigrams (two word sequence), trigram (three word sequence) and two separate unigrams. We also use position in the post as in the first part, last part or elsewhere. Beginning sentences can have different meanings than those in subsequent sentences. For example, "Thank you" in the beginning sentence position may be an expression of gratitude for previous information, while "thank you" in the last sentence may indicate only politeness.
F2: Message position in the thread: Indicates if the post is the first post, last post or one of the other posts.
F3: The emotion acts of the previous message: EAs in the previous message that the current message is replying to.
F4: Poster class: Defined as either a student or instructor.
F5: Poster change: Checks if the current poster is the same as the previous.
F6: Post length: Categorizes the post as Short (1-5 words), Medium(6-30 words), or Long (>30 words).

Given all combination of features F1-F6, we used Information Gain (Yang and Pedersen, 1997) to prune the feature space and select features. For each Emotion Act, we sorted all features (lexical and non-lexical) by Information Gain and used the top N (=200) features.

We used the Support Vector Machine of Chang and Lin (2001). We did a 5-fold cross validation in the training. RBF (Radial Basis Function) was used as the kernel function. We performed a grid search to get the best parameter (C and gamma) in training and applied them to the test corpus. With the training data of 159 threads and the test data of 52 threads, the initial classification result is shown in Table 4.

The initial results indicate that the EA classification is feasible. Due to the relatively small set size of available manually-annotated training data, the result is not yet at a level where it can be immediately applied in a functional setting. However, we strongly expect these results to improve as more training data becomes available.

Test Data Results			
Emotion Act	Precision	Recall	F-Score
High Certainty	0.68	0.64	0.65
Low Certainty	0.80	0.83	0.81
Frustration	0.73	0.75	0.73

Table 4. Automatic classification test results for certainty and frustration.

Related Work

Our work builds on prior research on spoken dialogue analysis including dialogue acts (Searle 1969; Hirschberg and Litman 1993; Samuel 2000; Graesser et. al., 2001; Kim et al., 2009), rhetoric analysis (Mann and Thomson, 1988), and surface cue words analysis (Hirschberg and Litman 1993; Samuel 2000). There have also been Dialogue Acts modeling approaches for automatic tagging and recognition of conversational speech (Stolcke et al., 2000) and related work in corpus linguistics where machine learning techniques have been used to find conversational patterns in spoken transcripts of dialogue corpus (Shawar and Atwell, 2005). Although spoken dialogue is different from message-based conversation in online discussion boards, they are closely related to our thread analysis work, and we plan to investigate potential use of conversation patterns in spoken dialogue in threaded discussions.

Carvalho and Cohen (2005) present a dependency-network based collective classification method to classify email speech acts. However, estimated speech act labeling between messages is not sufficient for assessing contributor roles or identifying help needed by the participants. We included other features like participant profiles. Also our corpus consists of less informal student discussions rather than messages among project participants, which tend to be more technically coherent.

Requests and commitments of email exchange are analyzed in (Lampert et al., 2008). As in their analysis, we have a higher kappa value for questions than answers, and some sources of ambiguity in human annotations such as different forms of answers also appear in our data. However, student discussions tend to focus on problem solving rather than task request and commitment as in project management applications, and their data show different types of ambiguity due to the different nature of participant interests.

There has also been work on non-traditional, qualitative assessment of instructional discourse (Boyer et al., 2008; Graesser et al., 2005; McLaren et al., 2007), and results have been used to find features for critical thinking and level of understanding. Similar approaches for classifying speech acts were investigated in Ravi and Kim (2007). This work captures features that are relevant to analyzing

noisy student discussion threads and supports a full automatic analysis of student discussions instead of manual generation of thread analysis rules. Earlier work on annotating emotion in dialogue focused on polarity (positive or negative) and intensity (Craggs and Wood, 2004) but is less useful for analyzing student discussions.

Finally, there have been studies of student affective states in tutorial dialogue, including boredom, confusion, surprise and frustration. These were analyzed and captured using dialogue states with linguistic features such as cohesion measures (D'Mello et al., 2009). Our work focuses on 'threaded' discussions, and is potentially useful for analyzing student discussion outcome.

Summary and Future Work

As the distinctions between resolved and unresolved threads show that profiling and automatic identification by affect is fully possible, it is important to look forward toward methods and directions of higher-level interpretation. The procedure used in investigating closure is only for broad proof-of-concept, rather than developing specific profile criteria for automatic categorization. As such, future development in profiling will require specific categories, defined by interactions within posts between differing affect in a repeatable manner. This can reveal information about important qualities of posts, threads, and students.

We have described an important first step towards the identification and use of emotion acts for instructional analysis of student discussions: We have identified common acts used by students within a course discussion board, developed a promising classification approach, and have shown that these acts are significant within the corpus through an investigation of resolved/unresolved threads. There are many research avenues to explore. In combination with existing metrics based on rhetorical speech acts, contribution quantity and technical depth, the new measures will assist instructors and researchers in understanding how students learn. This study complements prior work on speech acts and discussion topics (Carvalho & Cohen 2005; Feng et al., 2006; Kim et al. 2007).

References

Boyer, K., Phillips, R., Wallis M., Vouk M., Lester, J., Learner Characteristics and Feedback in Tutorial Dialogue (2008), *ACL workshop on Innovative Use of NLP for Building Educational Applications*.

Brown, P. and Levinson, S.C. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Carvalho, V.R. and Cohen, W.W., (2005). On the collective classification of email speech acts, *Proceedings of the SIGIR Conference*.

Chang, C. and Lin, C. (2001). *LIBSVM: a library for support vector machines*.

Craggs R., and Wood M. (2004). A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue. *Affective Dialogue Systems*, Elsevier, 89-100.

D'Mello, S., Dowell, N., and Graesser (2009). A.: Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States. *Proceedings of the AI in Education Conference*.

Feng, D., Kim, J., Shaw, E., and Hovy, E. (2006), Towards Modeling Threaded Discussions through Ontology-based Analysis, *Proceedings of National Conference on Artificial Intelligence*.

Graesser, A. C., Olney, A., Ventura, M., Jackson, G. T., (2005). "AutoTutor's Coverage of Expectations during Tutorial Dialogue." *Proceedings of the FLAIRS Conference*.

Graesser, A., VanLehn, K., Rosé, C., Jordan, P., Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine*, 22(4).

Hirschberg, J. and Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases, *Computational Linguistics*, 19(3).

Kim, J., Kim, T. and Li, J. (2009). Identifying Student Online Discussions with Unanswered Questions, *Proceedings of the International Conference on Knowledge Capture*.

Kim, J., Shaw, E. Chern, G. and Herbert, R. (2007) Novel tools for assessing student discussions: Modeling threads and participant roles using speech act and course topic analysis, *Proceedings of the AI in Education Conference*.

Kumrow, D. (2005). Student Self-Regulatory Resource Management Strategies and Academic Achievement in a Web-based Hybrid Graduate Nursing Course. *Education and Technology*, Editor(s): T.C. Montgomerie, J.R. Parker, ICET.

Lampert, A., Dale, R., and Paris, C. (2008). The Nature of Requests and Commitments in Email Messages", *AAAI workshop on Enhanced Messaging*.

Mann, W.C. and Thompson, S.A., (1988). Rhetorical structure theory: towards a functional theory of text organization. *Text: An Interdisciplinary Journal for the Study of Text*, 8 (3)

McLaren, B. et al., Using Machine Learning Techniques to Analyze and Support Mediation of Student E−Discussions, *Proceedings of the AI in Education Conference*.

Newman, R. and Schwager, M. (1995). Students' Help Seeking During Problem Solving: Effects of Grade, Goal, and Prior Achievement. *American Educational Research Journal*, v32 n2.

Ordelman, R. and Heylen, D. (2005). *Annotation of Emotions in meetings in the AMI project*.

Ravi, S., Kim, J. (2007) Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers, *Proceedings of the AI in Education Conference*.

Rubin, V., Liddy E., and Kando, N. (2006). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In Computing Attitude and Affect in Text: Theory and Applications.

Samuel, K., (2000) *An Investigation of Dialogue Act Tagging using Transformation-Based Learning*, PhD Thesis.

Searle, J., (1969). Speech Acts. Cambridge: Cambridge Univ. Press.

Shawar, B. A. and Atwell, E. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, V10.

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