High Accuracy Detection of Collaboration From Log Data and Superficial Speech Features

Sree Aurovindh Viswanathan, Arizona State University, sviswa10@asu.edu
Kurt VanLehn, Arizona State University, kurt.vanlehn@asu.edu

Abstract: Effective collaborative behavior between students is neither spontaneous nor continuous. A system that can measure collaboration in real-time may be useful. For instance, it could alert an instructor that a group needs attention. We tested whether superficial measures of speech and user interactions of students would suffice for measuring collaboration. As pairs of students solved complex math problems on tablets, their speech and tablet gestures were recorded. These data and multi-camera videos were used by humans to code episodes as collaborative vs. various kinds of non-collaboration. Using just the speech and tablet log data, several detectors were machine learned. The best had an overall accuracy of 96% (Kappa=0.92), which is higher than earlier attempts to use speech and log data for detecting collaboration. The improved accuracy appears to be due both to our analytic methods and to the particular mathematical task, which involves moving objects.

Keywords: collaboration detection, cooperation detection, learning analytics

Introduction

Many projects have worked on the challenge of automating the analysis of interaction among group members. These antecedents will be briefly reviewed by defining two dimensions, *purpose* and *input*, then describing the few systems whose position along these two dimension match the position of the project reported here. The two dimensions are excerpted from several similar multi-dimensional reviews (Diziol & Rummel, 2010; Magnisalis, Demetriadis, & Karakostas, 2011; Rummel, Walker, & Aleven, 2016; Soller, Martinez, Jermann, & Muehlenbrock, 2005; VanLehn, 2016). When a large number of projects could be cited as illustrations of a dimension, only those published most recently will be cited.

The first dimension concerns the purpose or function of the collaboration measure. That is, what does the system do with the output of the collaboration detector? This dimension has the following categories: Clustering, Classification, Mirroring, Meta-cognitive, Guiding, Orchestration and Restructuring. Our project fits into the *Classification* category. Projects in this category (e.g., Anaya & Boticario, 2011; Chounta & Avouris, 2012, 2014; Dragon, Floryan, Woolf, & Murray, 2010; Gweon, Agarawal, Raj, & Rose, 2011; Gweon, Jain, McDonough, Raj, & Rose, 2013; Martinez-Maldonado, Dimitriadis, Martinez-Mones, Kay, & Yacef, 2013; Martinez-Maldonado, Kay, & Yacef, 2013a; Martinez-Maldonado, Wallace, Kay, & Yacef, 2011; Martinez-Maldonado, Yacef, & Kay, 2013) used human judges to code group interactions into a variety of collaboration categories, then used supervised machine learning methods to induce classifiers (also called detectors) whose accuracy was measured and reported to researchers.

The second dimension classifies prior work by input to the detector. All projects had students working on a shared workspace, so they included the users' interactions (log data) as one input. Most projects also analyzed some form of communication among group members. The communication input can be classified as:

- Group members communicated in a formal language (Tedesco, 2003).
- Group members used a small set of buttons to express agreement/disagreement (de los Angeles Constantino-Gonzalez, Suthers, & de los Santos, 2003).
- Group members communicated by typing natural language and classifying their contribution using a menu of sentence openers or speech acts. Some systems ignored the text and used *only* the students' classifications of their text (e.g., Bravo, Redondo, Verdejo, & Ortega, 2008; Soller, Wiebe, & Lesgold, 2002)
- Group members communicated via typing (chat), with or without sentence openers. The text was analyzed by human "wizards" (Tsovaltzi et al., 2010), keywords (e.g., Adamson, Dyke, Jang, & Rose, 2014; Dragon et al., 2010; Martinez-Maldonado, Yacef, & Kay, 2015) or machine-learned text classifiers (e.g., McLaren, Scheuer, & Miksatko, 2010; Walker, Rummel, & Koedinger, 2014).
- Group members conversed in unconstrained speech, recorded by individual microphones (Bachour, Kaplan, & Dillenbourg, 2010; Gweon et al., 2011; Gweon et al., 2013; Martinez-Maldonado,

Dimitriadis, et al., 2013; Martinez-Maldonado, Kay, et al., 2013a; Martinez-Maldonado et al., 2011; Martinez-Maldonado et al., 2015; Roman, Mastrogiacomo, Mlotkowski, Kaplan, & Dillenbourg, 2012).

Our project falls into the last category, where group members' speech was analyzed. Two other projects also developed classifiers with speech input, so they will be described in more detail.

Although our work used machine-learned classifiers based on low level features of speech similar to Gweon et al. (2011; 2013), it differs in several ways. First, the Gweon et al classifiers used only the students' speech while ours used their actions as well. This allowed us to compare the accuracy attained from actions alone, speech alone and both actions and speech together. Second, whereas collaboration was the focal code in both projects, the two project chose different non-collaboration codes. This choice may impact accuracy, so we measured the accuracy of classifiers trained with different combinations of non-collaboration codes.

Our project is similar to one done by Martinez-Maldonado et al. (2015) in using both speech and actions. Their analysis of the participants' speech used a silence detector to convert the speech into a binary feature (present vs. absent). Machine-learned detectors induced from log data and the silence/talk data stream were only moderately accurate (Martinez-Maldonado, Dimitriadis, et al., 2013; Martinez-Maldonado et al., 2011). Increasing the length of the segments to 60 seconds or 90 seconds did not have much impact on accuracy (Martinez-Maldonado et al., 2011). The group also used differential sequence mining to find sequences of speech, silence and action that would reliably split groups into high and low collaboration (Martinez-Maldonado, Dimitriadis, et al., 2013; Martinez-Maldonado, Kay, et al., 2013a; Martinez-Maldonado, Yacef, et al., 2013). However, they did not convert their findings into a collaboration detector and measure its accuracy.

Generalizing across these two projects, it appears that a simple binary analysis of the speech signal into "talk" vs. "silence" may suffice for achieving moderate accuracy in collaboration detection. This was a Martinez-Maldonado et al. finding in several studies, and it is consistent with the strength of such features in the Gweon et al. detectors. However, we do not know how accuracy would change with the addition of other acoustic or prosodic features. Moreover, when speech is analyzed with this enriched feature set, we do not know how accuracy of a single modality, either speech or actions, compares with accuracy when using both modalities. Our project has addressed these questions, as well as studying classification accuracy in a different task domain from those studied earlier.

Methodology

System setup

This section describes the hardware and software setup that was used. Although students worked in pairs and sat beside each other at a table, they each had their own tablet. The tablet had active digitizer technology and a stylus that allowed students to write legibly on the 10-inch touchscreen. Students wore headset microphones.

The software used by participants is called FACT, an acronym for the Formative Assessment using Computational Technology (Cheema, VanLehn, Burkhart, Pead, & Schoenfeld, 2016). The FACT user interface mimics the original materials, which were a large poster (about 24" by 36") and paper cards (about 1" by 2") that are moved and eventually pasted onto the poster. Each group of students had one electronic poster. Group members could scroll and zoom independently, and they could edit different parts of the poster at the same time.

Participants and procedure

The study enrolled 28 paid participants. They were a mix of graduate and undergraduate students from Arizona State University. They were run one pair at a time in a lab. After the experimenter briefly described the problem, the students mostly worked head-down on their own tablets, but occasionally would huddle over one of them. They generally spoke without looking at each other. They worked until the problem was solved, which took 30 to 40 minutes. The same set-up and procedure have been used in approximately 15 classroom studies, but without the microphones.

The problem to be solved

Students were given a table with 3 columns and 9 rows, and a set of 27 cards to put into the 27 cells of the table. There were three types of cards: Graph cards, table cards and story cards. The graph cards were already positioned in the left column of the table. Students positioned the story cards in the middle column and the table cards in the right column. All cards in a row should describe the same process, which involves Tom making a short journey. (See Fig 1 for one such row).

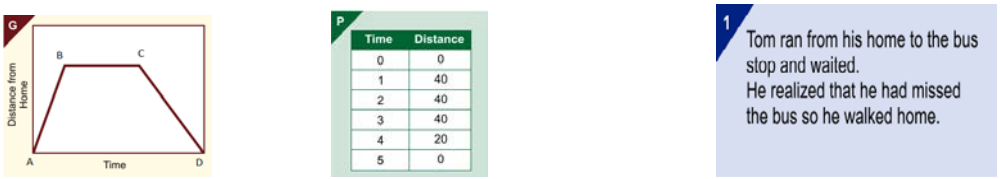

Figure 1. A row in the card matching task describing a story.

In building the row of Figure 1, students should make the following inferences about each points: a) *Point A on the graph card*: Tom is at his home, because the point is at zero on the vertical axis b) *Line segment A-B*: Tom moved away from home rapidly, because the segment has a steep slope c) *Line segment B-C*: Tom waited at the bus stop, because the segment is flat. d) *Line Segment C- D:* Tom returned home, because the segment has a negative slope. e) *Point D:* Tom reached home, because the point is on the x-axis.

The cards are designed around commonly observed misconceptions. For instance, students often view the graphs as a cross-section, so they often match the card in Figure 1 to the following story "*Opposite to Tom's home is a hill. Tom climbed slowly up the hill, walked across the top and rank quickly towards the down the other side.*" Our subjects found this task rather difficult, but all were able to solve it in less than 90 minutes.

Raw data collection

The recording setup combined several different input streams.

- Unidirectional headset microphones were used to capture each user's speech.
- The tablet screen content was streamed to a desktop computer using an HDMI cable.
- Log data were collected at the FACT server.
- Web cameras, one per student, recorded the student's head and shoulders. The video data were streamed to desktop computer.

The desktop computer showed all four videos on its screen: two tablet screen videos and two head and shoulder videos. It also received the two audio streams. Figure 2 shows a snapshot of the desktop computer's screen. In order to synch all 6 streams, the desktop screen was saved as a single video. Thus, all the data sources except the log data were synched as they were recorded.

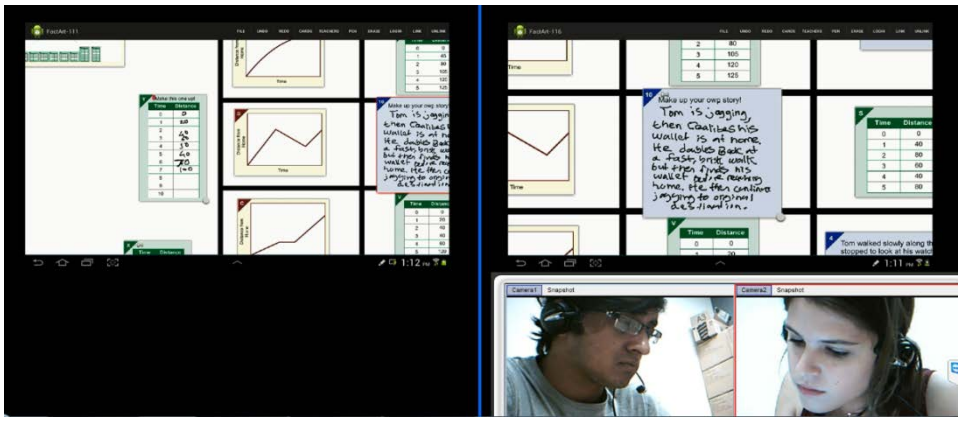

Figure 2. Snapshot of desktop recording student's action along with gestures.

Coding categories

When a collaboration monitoring system is used the classroom, it should probably help a teacher make a binary decision—whether to visit a group or not. For this, only two categories are essential: successful collaboration or not. Although Martinez-Maldonado and other prior researchers used a coding scheme developed by Meier, Spada and Rummel (2007), it requires thresholding in order to convert its scores into the binary collaboration/non-collaboration classification. We prefer to leave that decision to the coder. Nonetheless, we are interested in deeper level of granularity, so we defined four codes instead of just two:

- *Collaboration*. The interaction between the pair was considered collaboration when they both worked on placing the same card (i.e., they had the same immediate goal) and they often built on and extended

each other's reasoning. In a few situations, they disagreed with their partner's reasoning instead of extending it, and the pair engaged in argumentative co-construction until they reached agreement. This definition of collaboration includes attributes of joint problem solving such as common ground, knowledge convergence, co-construction, transactivity, scaffolding contributions and have a shared conception of the problem. The following is an example

> Student A: It is S [a table card] because all other tables are ending at zero
>
> Student B: No. This cannot be right. The distance can never be decreasing. In that [card] the distance is decreasing with time. In S
>
> Student A: 40...80...60...40...80... [Reading the table card] He is never going back.
>
> Student B: This one is for [the story card where] he has forgotten his watch
>
> Student A: Oh! Okay hmm...

- *Asymmetric Contribution*: The interaction between a pair was considered asymmetric contribution when they were working on placing the same card but one student did most of the work. That is, one person led the conversation and the other person added at most a few reasoning statements. We define two different levels of asymmetric contribution:
  - *Asymmetric Contribution (Low)*: The interaction between a dyad was characterized as "low" when no reasoning statements or exchanges occurred, but the human coder could tell from the videos that both students were attending to the same card. The following is an example:

    > Student A: For this card…
    >
    > Student B: Yes Tom is… [B moves the card to the solution grid] Yeah done.
    >
    > Student A : [Head Nod and both moves to the next problem]

  - *Asymmetric Contribution (High)*: The interaction between a dyad was characterized as "high" when the active person expressed reasoning and the passive person accepted the reasoning without contributing additional reasoning. The following is an example:

    > Student A: Probably, the first one will be 20 40 40 [Reading a table card] and it goes to zero. So that table…Because the slope while going up is little longer than the slope while coming down. So T [a table card] goes with E [a graph card]
    >
    > Student B: Yeah.. Yeah…

  This definition of asymmetric contribution shares a few characteristics of joint problem solving sessions such as common ground and establishing a shared conception of a problem. However, it lacks other properties such as transactivity, scaffolding contributions or argumentative co-construction.
- *Cooperation*. The interaction between a dyad was considered cooperation when subjects have different immediate goals, that is, they were working on placing different cards. Although there was usually little or no conversation between the pair, sometimes one student idly chattered about the problem and the other student did not respond back. Since students worked on different immediate goals, cooperation episodes do not have any attributes of joint problem solving.

This coding scheme is similar to ones used by other projects. In addition to a Collaboration code, almost all have noted that students sometimes work independently (our Cooperation code) and that sometimes one student is passive while the other does most of the work (our Asymmetric Contribution code). The distinction between High and Low Asymmetric Contribution is included because episodes where one person is explaining their actions while the other appears to listen (coded as High Asymmetric Contribution) may be consider a form of collaboration. In several studies where asymmetric verbal collaboration was pointed out to participants, the low-verbal participants rejected it as a meaningful measure of their participation (DiMicco, Pandolfo, & Bender, 2004; Roman et al., 2012), which suggests that they considered listening intently to be a form of collaboration.

Analysis methods

Data cleaning and segmentation

This section briefly describes our data cleaning, synchronization and segmentation techniques. The unidimensional mics meant that the audio streams obtained from tablets corresponded to single speakers' voice. The background noise was removed by using Audacity. Log data were recorded separately from the audio and video, so all data streams were synched post-hoc manually at the millisecond level.

Segmentation refers to dividing chronological data into segments or episodes (Chi, 1997). It is necessary whenever assigning a code (e.g., "Collaboration" vs. "Cooperation") to a whole session would be unreliable and perhaps even nonsensical. Although some projects have used overlapping segments (e.g., Gweon, Jain, McDonogh, Raj, & Rose, 2012; Rosé et al., 2008), most projects define segments to partition the whole session. Some projects use constant duration segments, such as 30 seconds (e.g., Martinez-Maldonado, Kay, & Yacef, 2013b; Martinez, Kay, Wallace, & Yacef, 2011). This can harm inter-coder reliability, so aligning segments with naturally occurring discontinuities is often preferred when possible. Our project's goal was to divide the solution of time-distance problem into separate sub-problems. A sub-problem is considered "done", and a segment boundary is placed, whenever a story or table card was placed in a table cell and the participant(s) moved on to another card. If the participants came back later and moved that card to a different cell, the new placement was considered a new segment.

Human coding

Once the segmentation was performed, human annotators classified each segment as either cooperative (P), Low asymmetric contribution (L-A), High asymmetric contribution (H-A) or collaboration (C). In addition to the audio recordings and log data from each participant, which were used as inputs to the machine learner and detector, the human coders viewed a four-video stream (see Figure 2). Thus, the human coders had much richer data than the machine detector. Two human coders tagged a sample of 35% of the segments. Inter-rater agreement was considered acceptable with Cohen's kappa = 0.78. For consistency across the whole dataset, the classifications of one annotator (the first author) were used in subsequent analyses.

Feature extraction and selection

The goal of the feature extraction process was to obtain superficial features from the students' work that could potentially differentiate between collaboration, cooperation and asymmetric contribution. Features were extracted computationally from audio files and log files; video data (Fig. 2) were ignored.

The log data were sequences of timestamped events that included card movements, scrolling and zooming. Feature definitions were specific to the user interface and the task. For instance, a sequence of scrolls with limited time between them were categorized as "search" scrolls whereas sequences with larger inter-event times were seen as "reading" scrolls. In addition, some features referenced the past behavior of the students up to this segment. Examples include a card moved twice by the same person to two different cells, a card moved twice by two different students to two different cells, and the total number of card placements so far by the student.

Audio features were extracted using both Audacity (silence and sound features) as well as the OpenSMILE audio feature toolkit, which represents "the state of the art for affect and paralinguistic recognition" (Eyben et al., 2010). We extracted all the features using the Emobase feature set which consisted of 1582 features and has often been used for non-semantic analyses of speech.

Feature selection was performed because the number of features was greater than the number of observations. Pairwise correlations were performed on features likely to be redundant. Sets of highly correlated features (coefficient > 0.9) were reduced to a single feature chosen arbitrarily from the set. Next, we applied resampling of the attributes in order have uniform distribution across class labels. Finally, we applied an attribute selection algorithm using best first search in Weka in order to reduce the feature set further. This reduced set of features was used for inducing the collaboration detectors.

Findings

The overall goal of the study is to induce a classifier that can distinguish collaboration from several types of non-collaboration. However, there is some ambiguity about how to treat the asymmetric contribution category, so three levels of granularity were defined and classifiers were induced for each.

- *Quaternary*: These classifiers were trained to distinguish all four categories coded by the human annotator. That is, their output was drawn from the set: Collaboration, Asymmetric contribution high, Asymmetric contribution low, and Cooperation.
- *Ternary*: These classifiers lumped together the two Asymmetric contribution categories, so their output was drawn from the set: Collaboration, Asymmetric contribution and Cooperation
- *Binary*: These classifiers lumped together Collaboration with Asymmetric contribution, so their output was drawn from the set: non-Cooperation (a broad definition of collaboration) vs. Cooperation.

Results from binary classifier

This section reports on the binary classifier, which was trained to discriminate only two categories: Cooperation versus Non-cooperation. We built classifiers for both the audio data alone, the log data alone and both sources of data combined. Random forest yielded the best result for all three data sets. The models were validated using the tenfold cross validation. Contingency tables, accuracy, Cohen's Kappa and F are shown in Table 1.

Table 1: Results from Binary Classifier (NP: non-Cooperation; P: Cooperation)

		Predicted Class (Audio) 93% (κ= 0.85; F=0.95)		Predicted Class (Log) 92% (κ= 0.83; F=0.94)		Predicted Class (Combined) 96% (κ= 0.92; F=0.97)	
		NP	P	NP	P	NP	P
True Class	NP	203	6	199	10	207	2
	P	16	100	15	101	10	106

Results from ternary classifier

The ternary classifier distinguished between Collaboration, Cooperation and Asymmetric contribution, where low and high asymmetric contribution was lumped together into one category. Additive logistic regression performed the best for the audio and combined feature sets, while random forests yielded the best result for features extracted from logs. The models were validated using the tenfold cross validation. Table 2 shows the results.

Table 2: Ternary Classifier (C: Collaboration; A: Asymmetric contribution; P: Cooperation)

		Predicted Class (Audio) 88% (κ= 0.82)			Predicted Class (Log) 85% (κ= 0.78)			Predicted Class (Combined) 86%(κ= 0.79)		
True Class		C	A	P	C	A	P	C	A	P
	C	88	10	1	80	11	8	82	12	5
	A	12	87	8	7	87	13	12	89	6
	P	1	8	110	4	5	110	1	9	109

Results from quaternary classifier

This classifier used the same codes as the human annotators. Random forests performed best for audio and combined feature sets while additive logistic regression performed best for log based feature sets. See Table 3.

Table 3: Quaternary classifier (C: Collaboration; H-A & L-A: High & Low Asymmetric; P: Cooperation)

		Predicted Class (Audio) 85% (κ= 0.79)				Predicted Class (Log) 77% (κ= 0.66)				Pred. Class (Combined) 87% (κ= 0.81)			
		Cc	H-A	L-A	P	C	H-A	L-A	P	C	H-A	L-A	P
True Class	C	64	10	5	3	58	1	3	20	74	7	1	0
	H-A	11	27	4	2	3	19	1	21	9	29	3	3
	L-A	5	2	57	3	2	0	45	20	5	1	55	6
	P	0	2	2	128	1	1	1	129	0	3	4	125

With one exception, accuracies were high for all classifiers. The exception is the quaternary classifier that used log data only, whose accuracy was 77%. This makes sense, because the difference between high and low asymmetric collaboration depends on the amount of conversation by the participants, and hence it would have been difficult for the log based features to detect this. However, this explanation predicts confusion between the H-A and L-A categories, and this did not occur with the Log-only detector. Instead, the confusion was spread evenly about the categories.

Discussion and conclusion

When this project began, we did not think the induced detectors would be accurate because they used only a low-level audio analysis that could not understand what the participants are saying, and low-level log features that could not understand the participants' plans and goals. Against these low expectations, the results were surprisingly good, with accuracies between 85% and 96% (except for the quaternary log data detector, discussed

above). For comparison with prior work, Gweon et al. (2011) found F=0.35 for detecting transactivity, and Martinez-Maldonado et al. (2011) found F=0.68 for detecting collaboration, whereas all our binary detectors had F≥0.94. However, there are some caveats and limitations that should be mentioned.

First, the task used here involved moving objects. Much like the classic collaboration examples of moving furniture or assembling a jigsaw puzzle, when our participants were collaborating, they were moving the same physical object, or at least, one person was moving it and the other was watching and offering comments. On the other hand, when participants were cooperating, they were simultaneously moving different objects. We are not sure how well our method would work when the task does not align sub-problems with object movements. Thus, our next study (which is in progress) uses a task with no moving objects: Two people are sharing responsibility for providing written answers to questions.

A second limitation is that the use of an object-moving task allowed automation of segmentation. Subtask segmentation (i.e., not using segments of constant duration) is usually to be done by a human annotator who can understand the participants' speech, plans and goals (Chi, 1997). We are not sure if subtask-based segmentation can be automated with problems where the subtask boundaries are less salient.

The third limitation was that the audio collection and cleaning used here would not be robust enough for use in classrooms. We are currently working on better methods. Synchronization also needs to be improved, as it currently requires too much human attention.

A fourth limitation is use of log data features that probably do not generalize to other tasks. However, the audio features are task-general, and the audio-only detectors were nearly as accurate as the multi-modal detectors. This suggests that lower-level, task-independent log data features might work nearly as well as the task-specific features.

References

Adamson, D., Dyke, G., Jang, H., & Rose, C. P. (2014). Toward an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence and Education, 24*, 92-124.

Anaya, A. R., & Boticario, J. g. (2011). Application of machine learning techniques to analyze student interactions and improve the collaboration process. *Expert Systems with Applications, 38*, 1171-1181.

Bachour, K., Kaplan, F., & Dillenbourg, P. (2010). An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Transactions on Learning Technologies, 3*, 203-213.

Bravo, C., Redondo, M. A., Verdejo, M. F., & Ortega, M. (2008). A framework for process-solution analysis in collaborative learning environments. *international Journal of Human-Computer Studies, 66*, 812-832.

Cheema, S., VanLehn, K., Burkhart, H., Pead, D., & Schoenfeld, A. H. (2016). *Electronic posters to support formative assessment*. Paper presented at the CHI'16: Extended Abstracts.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences, 6*(3), 271-315.

Chounta, I.-A., & Avouris, N. (2012). Time series analysis of collaborative activities. In V. Herskovic, U. Hoppe, M. Jansen, & J. Ziegler (Eds.), *Collaboration and Technology, vol. 7493* (pp. 145-152). Berlin: Springer.

Chounta, I.-A., & Avouris, N. (2014). It's all about time: Towards the real-time evaluation of collaborative activities *IEEE 14th International Conference on Advanced Learning Technologies* (pp. 383-285): IEEE.

de los Angeles Constantino-Gonzalez, M., Suthers, D. D., & de los Santos, J. G. E. (2003). Coaching web-based collaborative learning based on problem solution differences and participation. *International Journal of Artificial Intelligence in Education, 13*(2), 263-299.

DiMicco, J. M., Pandolfo, A., & Bender, W. (2004). *Influencing group participation with a shared display*. Paper presented at the Computer Supported Collaborative Work, Chicago, IL.

Diziol, D., & Rummel, N. (2010). How to design support for collaborative e-learning: A framework of relevant dimensions. In B. Ertl (Ed.), *E-collaborative knowledge construction: Learning from computer-supported and virtual environments* (pp. 162-179). Hershey, PA: IGI Global.

Dragon, T., Floryan, M., Woolf, B. P., & Murray, T. (2010). *Recognizing dialogue content in student collaborative conversation*. Paper presented at the Intelligent Tutoring Systems, Berlin.

Eyben, F., W, M., #246, llmer, Bj, #246, & Schuller, r. (2010). *Opensmile: the munich versatile and fast open-source audio feature extractor*. Paper presented at the Proceedings of the international conference on Multimedia, Firenze, Italy.

Gweon, G., Agarawal, P., Raj, B., & Rose, C. P. (2011). *The automatic assessment of knowledge integration processes in project teams*. Paper presented at the Computer Supported Collaborative Learning.

Gweon, G., Jain, M., McDonogh, J., Raj, B., & Rose, C. P. (2012). Predicting idea co-construction in speech data using insights from sociolinguistics *Proceedings of the International Conference of the Learning Sciences. International Society of the Learning Sciences: Sydney, Australia*.

Gweon, G., Jain, M., McDonough, J., Raj, B., & Rose, C. P. (2013). Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer-Supported Collaborative Learning, 8*(2), 245-265.

Magnisalis, I., Demetriadis, S., & Karakostas, A. (2011). Adaptive and intelligent systems for collaborative and learning support: A review of the field. *IEEE Transactions on Learning Technologies, 4*(1), 5-20.

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Mones, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *Computer-Supported Collaborative Learning, 8*, 455-485.

Martinez-Maldonado, R., Kay, J., & Yacef, K. (2013a). An automatic approach for mining patterns of collaboration around an interactive tabletop. In K. Yacef (Ed.), *Artificial Intelligence in Education, AIED 2013* (pp. 101-110). Berlin: Springer-Verlag.

Martinez-Maldonado, R., Kay, J., & Yacef, K. (2013b). An automatic approach for mining patterns of collaboration around an interactive tabletop *Artificial Intelligence in Education* (pp. 101-110).

Martinez-Maldonado, R., Wallace, J., Kay, J., & Yacef, K. (2011). Modelling and identifying collaborative situations in a collocated multi-display groupware setting *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 196-204). Berline: Springer.

Martinez-Maldonado, R., Yacef, K., & Kay, J. (2013). *Data mining in the classroom: Discovering groups' strategies at a multi-tableop environment*. Paper presented at the Educational Data Mining.

Martinez-Maldonado, R., Yacef, K., & Kay, J. (2015). TSCL: A conceptual model to inform understanding of collaborative learning processes at interactive tabletops. *international Journal of Human-Computer Studies, 83*, 62-82.

Martinez, R., Kay, J., Wallace, J., & Yacef, K. (2011). Modelling Symmetry of Activity as an Indicator of Collocated Group Collaboration. In J. Konstan, R. Conejo, J. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization* (Vol. 6787, pp. 207-218): Springer Berlin Heidelberg.

McLaren, B., Scheuer, O., & Miksatko, J. (2010). Supporting collaborative learning and e-discussions using Artificial Intelligence techniques. *International Journal of Artificial Intelligence and Education, 20*, 1-46.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *Computer-Supported Collaborative Learning, 2*, 63-86.

Roman, F., Mastrogiacomo, S., Mlotkowski, D., Kaplan, F., & Dillenbourg, P. (2012). *Can a table regulate participation in top level managers' meetings*. Paper presented at the Proceedings of the 17th ACM International Conference on Supporting Group Work

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 3*(3), 237. doi:10.1007/s11412-007-9034-0

Rummel, N., Walker, E., & Aleven, V. (2016). Different futures of adaptive collaborative learning suport. *International Journal of Artificial Intelligence and Education, 26*, 784-795.

Soller, A., Martinez, A., Jermann, P., & Muehlenbrock, M. (2005). From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. *International Journal of Artificial Intelligence and Education, 15*, 261-290.

Soller, A., Wiebe, J., & Lesgold, A. (2002). A machine learning approach ot assessing knowledge sharing during collaborative learning activities. In G. Stahl (Ed.), *Proceedings of Computer Supported Collaborative Learning 2002* (pp. 128-137). Hillsdale, NJ: Erlbaum.

Tedesco, P. A. (2003). MArCo: Building an artificial conflict mediator to support group planning interactions. *International Journal of Artificial Intelligence in Education, 13*(1), 117-155.

Tsovaltzi, D., Rummel, N., McLaren, B., Pinkwart, N., Scheuer, O., Harrer, A., & Braun, I. (2010). Extending a virtual chemistry laboratory with a collaboration script to promote conceptual learning. *International Journal of Technology Enhanced Learning, 2*(1/2), 91-110.

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education, 26*(1), 107-112. doi:10.1007/s40593-015-0056-x

Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence and Education, 24*(1), 33-61.