AUTOMATIC CONTENT ANALYSIS IN COLLABORATIVE INQUIRY-BASED LEARNING

Catalina Espinoza[1], Joni Lämsä[2], Roberto Araya[1], Raija Hämäläinen[2], Abelino Jiménez[1], Raúl Gormaz[1], Jouni Viiri[3]

[1]Centre for Advanced Research in Education (CIAE) – University of Chile, Santiago, Chile
[2] University of Jyväskylä, Department of Education, Jyväskylä, Finland
[3] University of Jyväskylä, Department of Teacher Education, Jyväskylä, Finland

In the field of science education, content analysis is a popular way to analyse collaborative inquiry-based learning (CIBL) processes. However, content analysis is time-consuming when conducted by humans. In this paper, we introduce an automatic content analysis method to identify the different inquiry-based learning (IBL) phases from authentic student face-to-face discussions. We illustrate the potential of automatic content analysis by comparing the results of manual content analysis (conducted by humans) and automatic content analysis (conducted by computers). Both the manual and automatic content analyses were based on manual transcriptions of 11 groups' CIBL processes. Two researchers performed the manual content analysis, in which each utterance of the groups' discussions was coded to an IBL phase. First, an algorithm was trained with some of the manually coded utterances to prepare the automatic content analysis. Second, the researchers tested the ability of the algorithm to automatically code the utterances that were not used in the training. The algorithm was a linear support vector machine (SVM) classifier. Since the input of the SVM must be a numerical vector of constant size, we used a topic model to build a feature vector representation for each utterance. The correspondence of the manual and automatic content analyses was 52.9%. The precision of the classifier varied from 49% to 68%, depending on the IBL phase. We discuss issues to consider in the future when improving automatic content analysis methods. We also highlight the potential benefits of automatic content analysis from the viewpoint of science teachers and science education researchers.

Keywords: Collaborative learning; Inquiry-oriented learning; Quantitative methods

INTRODUCTION

In the field of science education, content analysis is a popular method for analysing collaborative inquiry-based learning (CIBL) processes (e.g., Lämsä et al., 2018, 2020; Wang et al., 2014). Typically, researchers have pre-defined codes, such as the phases of inquiry-based learning (IBL), on which the content analysis is based. Identification of different phases of IBL is needed when designing scaffolds for CIBL processes. For instance, previous research has shown that students need support, especially in the first phases of IBL (Lämsä et al., 2018; Wang et al., 2014). However, it is known that optimal scaffolding is context-specific, so there is a need to study CIBL processes in various contexts. Due to the vast human resources that content analysis of CIBL processes requires, in this study, we illustrate the potential of automatic content analysis for analysing CIBL processes. In the following, we briefly present how automatic content analysis has been applied in learning sciences so far.

Automatic content analysis in learning sciences

A major challenge in building a practical tool that supports teachers in the classroom is the need to analyse students' discussions in each group. Hence, we need novel methods to process and analyse massive amounts of data. This is not only time-consuming when performed by humans, but also requires the analysis of parallel speech produced in each group of students. Moreover, the analysis needs to be done in real time in order to be useful to the teacher. Furthermore, it is not straightforward to produce a robust diagnosis, as trained human raters must agree on coding. Specifically, in order to provide a robust analysis, it is necessary to have at least two raters and ensure sufficient inter-rater and intra-rater agreement.

So far, studies have investigated the potential for automatic content analysis in developing a better understanding of computer-mediated student–student interaction as well as teacher–student interaction. In the context of student–student interaction, Rosé et al. (2008) presented an overview of work on automatic analysis of computer-supported collaborative learning (CSCL) in which the authors stated, 'Our specific goal has been to extend and apply current text classification technology to CSCL, exploring which classification techniques are most effective for improving the performance on different types of coding dimensions used in the CSCL community.' They found that some discourse processes in educational psychology could be automatically detected in text messages at a level of agreement comparable to that between human coders. Dowell et al. (2018) also developed a group communication analysis by combining automated computational linguistic techniques with analyses of the sequential interactions of online group communications to detect emergent roles in group interactions.

Text analysis also allows the study of the effect of questions posed by the teacher. This is a subject of great importance in teaching practice and in the preparation of teachers. More than a century ago, Stevens (1912) emphasised the realisation of questions as a fundamental component in teacher training. Araya et al. (2018) analysed written responses to open-ended questions of students from various elementary school classes and built an automatic predictor of the length of the answers of each student based on the presence of keywords in the teacher's questions. Donnelly et al. (2017) used an automatic speech recognition (ASR) algorithm to automatically detect teachers' questions, and Caballero et al. (2017) used an ASR algorithm to automatically provide teachers with a visualisation of the structure of concepts present in their discourse in science classrooms. Araya et al. (2012) built an automatic classifier of estimations of teacher practices using ratings of a subset of transcriptions made by trained raters. The agreements between the automatic classifiers and the corresponding raters that were computed for transcriptions from an independent subset were better than the agreements between the human raters. More recently, Kelly et al. (2018) compared human coding and semi-automated computer coding of the authenticity of teacher questions. They concluded that the correlations were sufficiently high to provide a valuable complement to human coding in research efforts.

These examples show that automatic content analysis has been applied to the study of computer-mediated communication between students and teachers in CSCL classrooms. CSCL research could benefit from the opportunities afforded by technological and methodological development. So far, the results of automatic content analysis have been encouraging. In this paper, we apply automatic content analysis to authentic student face-to-face interactions taking place in computer-supported settings in order to identify the different IBL phases in students' discussions. We address the research question (RQ): How similar are the results of the proposed automatic and manual content analyses in a CIBL context?

METHODS

Our study was conducted in introductory university physics courses on thermodynamics at a Finnish university. The participants were divided into groups of five students at the beginning of the course. Here, we focus on face-to-face discussions as the groups solved problems collaboratively in a technology-enhanced learning environment with shared laptop computers. Eleven groups screen-captured and audio-recorded their group-work sessions. First, we manually transcribed these sessions as they solved an inquiry problem (on average, 180 utterances per group). The inquiry problem was a study of how the displacement of an atom in a two-dimensional gas depends on time. The groups had a Python programme that calculated the displacement of an atom with different values of the number of collisions and then plotted the atom's path. Based on the output of the Python programme, the groups inferred the relationship between the displacement and time.

Second, we conducted theory-driven content analysis (Neuendorf, 2002) in which two researchers coded the transcriptions based on the IBL framework presented by Pedaste et al. (2015), i.e., each utterance was coded to one of the IBL phases (orientation, conceptualisation, investigation, conclusion, and discussion). In the orientation phase, the students became familiar with the given assignment, its main variables, and technological resources (in this case, the Python programme). In the conceptualisation phase, the students identified the dependent and independent variables of the problems and proposed research questions or hypotheses. In the investigation phase, the students planned the data collection and collected, analysed, and interpreted the data. In the conclusion phase, the students drew conclusions and offered solutions to the research questions or hypotheses. In the discussion phase, the students could communicate and reflect on the process at the end of the inquiry or in relation to an IBL phase. The inter-rater agreement was 67.7%, and any disagreements were discussed and resolved.

After this manual content analysis, we performed automatic content analysis to identify the IBL phase from a given utterance. All the analysed transcriptions were written in Finnish. We approached this task as a text classification problem—finding a characterisation of utterances (vector representation) and building an automatic classifier. Figure 1 shows the four stages of the automatic content analysis that we conducted, describing the process for each stage and the corresponding input and output.

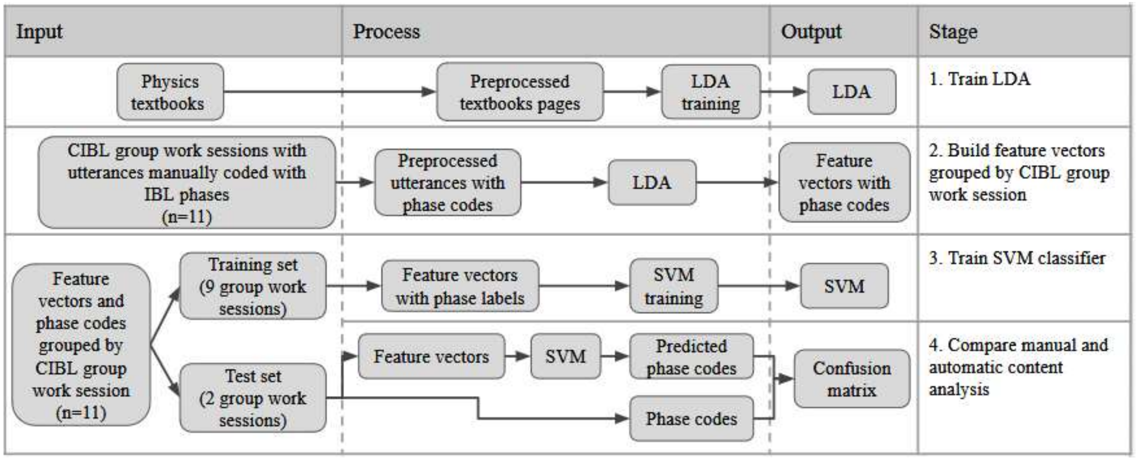

Figure 1. Stages for training and assessing the automatic content analysis.

Stage 1: Train an LDA

In order to build a characterisation of utterances, we trained a latent Dirichlet allocation (LDA) topic model (Blei et al., 2003). Topic models are statistical models that are used to find topics in large document collections in a wide range of applications, e.g., analysing historical documents, understanding scientific publications, or machine translation (Boyd-Graber et al., 2017; Dowell et al. 2018). The term *topic*—in the topic models context—is used to refer to groups of words that usually appear together in a document collection. The assumption behind topic models is that documents have a mixture of topics, and the words in a single document will depend on the topics that comprise the document. Applying a topic model to a large collection of documents allows the identification of topics and the description of documents in terms of the topics present in each document.

Table 1. Example of seven topics found by applying the LDA model to physics textbooks. These topics were common in the CIBL group-work sessions. For each topic, the top four words are shown.

Topic 8	Topic 13	Topic 17	Topic 19	Topic 27	Topic 33	Topic 39
[number]	v	=	so	T	?	[proportional to]
a	R	+	for example	1	a	measurement
b	voltage	R	same	o	how	increase
m	1	o	or	R	b	measure

Topic models need a large collection of documents to find topics that are formed by coherent groups of words. In learning sciences, there are usually learning materials available for different contexts, subjects, and languages. Our assumption was that we could use relevant topics from upper secondary school physics textbooks if we wanted to describe the discussions of undergraduate physics students. Thus, the first stage consisted of using 31 Finnish physics textbooks to train the LDA model. We trained the model to identify 60 topics, which allowed us to represent an utterance as a mixture of 60 topic proportions. Table 1 shows the words most relevant to seven of the topics. We have translated the words from Finnish to English. Topic

19 relates to words used in explanations. Topic 39 relates to words describing measurements. The other topics refer to numbers, units, name of variables and symbols.

Stage 2: Build feature vectors grouped by CIBL group-work session

In the second stage, we built a feature vector, or an enhanced representation of each utterance in the CIBL group-work sessions. The feature vector was used to train the automatic classifier in the next stage. Each utterance of the CIBL sessions was pre-processed (removing stopwords, symbols, and infrequent words). The LDA model was then used to obtain, for each utterance, a vector of 60 topic proportions. The output of this stage was a 182-dimensional feature vector representing each utterance: 60 components of the topic proportions for the utterance, 60 for the previous utterance, and 60 for the following utterance. The two additional dimensions corresponded to the number of words and the relative position of the utterance in the group-work session. It is important to note that each feature vector was related to the corresponding human-coded phase.

For the third and fourth stages, we grouped the feature vectors into two sets: a training set with the feature vectors from nine group-work sessions and a test set with the feature vectors from two group-work sessions. The training set would be the input for the third stage and the test set would be the input for the fourth stage.

Stage 3: Train the SVM classifier

In the third stage, we trained a linear support vector machine (SVM) classifier (Burges, 1998). SVMs are one of the most popular automatic classifiers (e.g., Araya et al., 2012; Rosé et al., 2008). The training of the SVM consisted of adjusting a function to optimise the number of utterances in the training set that were automatically coded to be the same as the manual code. By the end of Stage 3, a workflow had been created that allowed automatic coding of an utterance with an IBL phase. The following are examples of utterances coded with the SVM classifier:

> Yes, it was something like 20, approximately [the number of collisions $N = 300$ in the Python programme]. Then it was a little bit more than 30. Let's now try when it [*N*] is 500. I'll try a couple of times: 11, 15, 25, 21, 15 … [manual coding: investigation; automatic coding: investigation]
>
> ...
>
> Would someone else like to tap [run the Python programme]? [manual coding: discussion; automatic coding: orientation]

Stage 4: Compare the manual and automatic content analyses

In the fourth stage, we used the test set to compare the manual coding with the phase automatically assigned by the SVM (see the previous examples). The output of the fourth stage was a confusion matrix, i.e., a 5×5 matrix that summarises the number of utterances that were manually coded as the phase indicated by the row and automatically coded as the phase indicated by the column.

As the dataset was small, the results were highly dependent on the test set selected. To get a more robust estimation of the SVM classifier's errors, we independently ran the third and fourth stages while varying the input with all the possible combinations of nine and two group-work sessions in order to build the training and test sets. With 55 combinations in total, we obtained 55 confusion matrices. As different test sets had different numbers of utterances, each matrix was standardised to total 100. The output of the overall process was then an average confusion matrix of the 55 standardised matrices. This procedure allowed us to answer the RQ, as the confusion matrix described how well the IBL classifier was performing for each IBL phase.

Baseline

To test whether the classifier was gaining information from the text features, we defined two baselines that did not require the extraction of information from the utterances. The first baseline was a classifier that classified all the utterances as the more frequent phase. In this study, the discussion phase was the most frequent and represented 36.1% of the utterances in the CIBL sessions. The second baseline was a classifier that, in the first part of the sessions, classified the utterances as orientation, and, in the second part of the sessions, classified them as discussion. To define where in the transcription the baseline classifier should start to classify utterances as discussion, we used the training dataset to find the optimal threshold for each run of the third and fourth stage. The average threshold used for the second baseline was to classify the first 40% of the session as orientation and the remainder as discussion. The second baseline had an accuracy of 42.7%.

RESULTS

A comparison of the results of the manual and automatic content analyses is presented in the confusion matrix (Table 2). Each cell $C_{i,j}$ in the confusion matrix is the average number of utterances that were manually coded to IBL phase i and automatically coded to IBL phase j. Each cell $C_{i,j}$ represents an average across all the independent runs of the third and fourth stages of the automatic content analysis. The accuracy of the classifiers are the coincidences between the manual and automatic content analyses. As the resulting matrix in Table 2 is standardised to a total of 100, the accuracy is obtained by adding up the diagonal. The overall accuracy was 52.9% (SD = 4.8%). The precision of the classifier for each phase is the number of times that the automatically coded phase was the same as the manually coded phase. The precision of automatic coding of different IBL phases varied from 49% to 68% (orientation 50%, conceptualisation 49%, investigation 68%, conclusion 49%, and discussion 51%).

Table 2. Comparison of the results of the manual and automatic content analyses. The rows refer to the manual content analysis, and the columns refer to the automatic content analysis.

	Predicted Orientation	Predicted Conceptualisation	Predicted Investigation	Predicted Conclusion	Predicted Discussion
Orientation	16.3	0.4	1.1	0.0	6.8
Conceptualisation	2.8	2.2	0.8	0.0	5.7
Investigation	5.8	0.3	9.6	0.0	8.6
Conclusion	0.1	0.6	0.2	0.4	2.5
Discussion	7.6	0.9	2.5	0.4	24.4

In addition to the precision of the classifiers, there are other indexes that can be used to measure the reliability of the automatic coding. For example, the recall for each phase is the number of times that the automatic coding agreed with the manual coding, divided by the frequency of the phase. The recall of the automatic coding varied from 10% to 68% (orientation 66%, conceptualisation 19%, investigation 39%, conclusion 10%, and discussion 68%). These results show that the recall was rather low in the investigation phase compared to the orientation and discussion phases. This demonstrates that utterances manually coded to the investigation phase were automatically coded to the orientation and discussion phases many times, as shown in Table 2.

The results show that the precision of the automatic coding varied depending on the IBL phase. In particular, the precision of the investigation phase was higher than the precision of the other phases, i.e., when the utterance was automatically coded to the investigation phase, the utterance was also manually coded to the investigation phase with 68% probability (even though the recall was 39% for the investigation phase). The following is an utterance that was coded to the investigation phase both manually and automatically:

> Yes, it was something like 20, approximately [the number of collisions $N = 300$ in the Python programme]. Then it was a little bit more than 30. Let's now try when it [N] is 500. I'll try a couple of times: 11, 15, 25, 21, 15 … [manual coding: investigation; automatic coding: investigation]

In the investigation phase, the students had to, amongst other things, collect data to address the inquiry problem. The previous utterance shows an example of the data collection. During that collection, students collected the values of the displacement of an atom with different values of the number of collisions so they could infer the relationship between the displacement and time. Numbers were thus a characteristic of the investigation phase (cf. Topic 8 in Table 1). The following utterance demonstrates a challenge for the automatic coding: the utterance was manually coded to the discussion phase but automatically coded to the orientation phase.

> Would someone else like to tap [run the Python programme]? [manual coding: discussion; automatic coding: orientation]

Even though the classifier on which the automatic coding is based accounts for the previous and subsequent utterances as well as the relative position of the utterance in the whole discussion, the consideration of the overall context of the utterances is difficult to automatise. This specific utterance was manually coded to the discussion phase as it is about communicating and suggesting a new way to proceed with the inquiry problem (changing the student in charge of working with the Python programme). As can be seen from Table 2, the conceptualisation and conclusion phases were rare in the students' discussions compared to the other IBL phases, and there is thus not a representative example that illustrates inter-rater agreement between the manual and automatic coding in these phases.

DISCUSSION AND CONCLUSION

This study was a novel attempt to automate content analysis in authentic CIBL contexts in which students were working face-to-face in computer-supported settings. We compared the results of the manual and automatic content analyses, which were based on an SVM classifier.

The average accuracy of the SVM classifier (52.9%, SD = 4.8%) was 15% lower than the agreement between human coders (67.7%). Regarding the baselines, the first classified all the utterances as discussion and had an accuracy of 36.1% (SD = 5.8%). The second baseline classified the first part of the IBL sessions as orientation and the rest as discussion and had an accuracy of 42.7% (SD = 5.3%). The SVM classifier performed significantly better than the baselines; therefore, the topic description of the sentences provided information that enabled the SVM classifier to distinguish the IBL phases. Overall, these results highlight the potential for using automatic content analysis both in CIBL contexts and in face-to-face interaction in general.

There are still issues that future research should consider. First, the results for the recall and precision per phase indicate that the SVM classifier was biased against the most frequent phases. This is known as an unbalanced dataset problem. In our study, the precision was notably higher in the investigation phase than in the other phases. The recall of investigation was 39%, and the classifier frequently confused the investigation phase with orientation or discussion. When the classifier correctly coded investigation, topics 8 (numbers and units), 19 (so, for example, same, or, …) and 33 (question mark, units, how, large, calculate, ...) were present. In the future, this bias could be addressed by gathering more examples of the less-frequent phases (conceptualisation and conclusion).

Second, we treated the results of the manual content analysis as reliable so they could be compared with the results of the automatic content analysis. However, inter-rater agreement in the manual content analysis indicated that the codes for many utterances were not unambiguous, but were decided after careful joint consideration. In future, research could check whether disagreement between computer coding and human coding appears more frequently in utterances in which there is disagreement between the human coders. Third, to build the numerical representation of the utterances, we used an LDA model. As our dataset was small (11 group discussions), we trained the LDA model with physics textbooks, which contained only some of the language that appears in authentic face-to-face discussions. In the future, we will integrate transcriptions from student discussions into the LDA training.

Combining emerging automatic content analysis methods with ASR applications could be beneficial for both researchers and teachers. Automatic content analysis could support researchers with preliminary coding so they can focus more on tasks that cannot be automated, such as designing experiments and the interpretation of results, instead of time-consuming data transcription and coding. As productive CIBL activities do not necessarily emerge without assistance (Alfieri et al., 2010; Kobbe et al., 2007), a tool that can detect the stage of group discussions in real time could help in adapting support to the needs of different groups. An advantage of building the feature vectors based on LDA and using the SVM classifier is that the analysis is not necessarily language-specific. Even though we trained the LDA model using Finnish textbooks, our novel methodological approach is applicable to automatising content analysis in any language in which there are available textbooks (or other similar materials to train the LDA model) matching the language and subject of the CIBL activities.

ACKNOWLEDGEMENTS

We would like to express our gratitude for the financial aid provided by Basal Funds for Centers of Excellence Project BF 0003 from CONICYT in Chile and the Academy of Finland [grant numbers 292466 and 318095, the Multidisciplinary Research on Learning and Teaching profiles I and II of JYU].

REFERENCES

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1–18. doi: 10.1037/a0021017

Araya, R., Plana, F., Dartnell, P., Soto-Andrade, J., Luci, G., Salinas, E., & Araya, M. (2012). Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm. *British Journal of Educational Technology, 43*(6), 836–846.

Araya R., Jiménez A., & Aguirre C. (2018). Context-based personalized predictors of the length of written responses to open-ended questions of elementary school students. In Sieminski A., Kozierkiewicz, A., Nunez, M., & Ha, Q. (eds.), *Modern Approaches for Intelligent Information and Database Systems. Studies in Computational Intelligence*, vol. 769, pp. 135–146. Cham, Switzerland: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan.), 993–1022.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval, 11*(2–3), 143–296.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Caballero, D., Araya, R., Kronholm, H., Viiri, J., Mansikkaniemi, A., Lehesvuori, S., Virtanen, T., & Kurimo, M. (2017) ASR in classroom today: Automatic visualization of conceptual network in science classrooms. In Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., & Pérez-Sanagustín, M. (eds.), *Data driven approaches in digital education. EC-TEL 2017. Lecture Notes in Computer Science*, vol. 10474, pp. 541–544. Cham, Switzerland: Springer.

Donnelly, P., Blanchard, N., Olney, A., Kelly, S., Nystrand, M., & D'Mello, S. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference*, pp. 218–227. New York, NY: Association for Computing Machinery.

Dowell, N., Nixon, T., & Graesser, A. (2018). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multi-party interactions. *Behavior Research Methods, 51*, 1007–1041.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher, 47*(7), 451–464.

Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämäläinen, R., Häkkinen, P., & Fischer, F. (2007). Specifying computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning, 2*(2/3), 211–224.

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Mannonen, J. (2020). The potential of temporal analysis: Combining log data and lag sequential analysis to investigate temporal differences between scaffolded and non-scaffolded group inquiry-based learning processes. *Computers & Education, 143*.

Lämsä, J., Hämäläinen, R., Koskinen, P., & Viiri, J. (2018). Visualising the temporal aspects of collaborative inquiry-based learning processes in technology-enhanced physics learning. *International Journal of Science Education*, *40*(14), 1697–1717.

Neuendorf, K. A. (2002). *The content analysis guidebook* (1st ed.). Thousand Oaks, CA: Sage Publications.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., . . . Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61.

Rosé, C., Wang, Y., 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. *Computer-Supported Collaborative Learning, 3*, 237–271

Stevens, R. (1912) The question as a measure of efficiency in instruction: A critical study classroom practice. *Teachers College Contributions to Education, 48*.

Wang, H., Duh, H., Li, N., Lin, T., & Tsai, C. (2014). An investigation of university students' collaborative inquiry learning behaviors in an augmented reality simulation and a traditional simulation. *Journal of Science Education and Technology, 23*(5), 682–691.