JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 8(1), 113–125. http://doi.org/10.18608/jla.2021.7118

Deep Networks for Collaboration Analytics: Promoting Automatic Analysis of Face-to-Face Interaction in the Context of Inquiry-Based Learning

Joni Lämsä[1], Pablo Uribe[2], Abelino Jiménez[3], Daniela Caballero[4], Raija Hämäläinen[5], Roberto Araya[6]

Abstract

Scholars have applied automatic content analysis to study computer-mediated communication in computer-supported collaborative learning (CSCL). Since CSCL also takes place in face-to-face interactions, we studied the automatic coding accuracy of manually transcribed face-to-face communication. We conducted our study in an authentic higher-education physics context where computer-supported collaborative inquiry-based learning (CSCIL) is a popular pedagogical approach. Since learners' needs for support in CSCIL vary in the different inquiry phases (orientation, conceptualization, investigation, conclusion, and discussion), we studied, first, how the coding accuracy of five computational models (based on word embeddings and deep neural networks with attention layers) differed in the various inquiry-based learning (IBL) phases when compared to human coding. Second, we investigated how the different features of the best performing computational model improved the coding accuracy. The study indicated that the accuracy of the best performing computational model (differentiated attention with pre-trained static embeddings) was slightly better than that of the human coder (58.9% vs. 54.3%). We also found that considering the previous and following utterances, as well as the relative position of the utterance, improved the model's accuracy. Our method illustrates how computational models can be trained for specific purposes (e.g., to code IBL phases) with small data sets by using pre-trained models.

Notes for Practice

- Knowing the inquiry-based learning (IBL) phase of the learners' process may help with the provision of timely guidance for computer-supported collaborative inquiry-based learning (CSCIL).
- Instead of using classical algorithms to analyze computer-mediated interaction, we present computational models based on advanced natural language processing techniques for automatically coding orientation, conceptualization, investigation, conclusion, and discussion phases from authentic face-to-face conversations.
- Our methodology shows the potential of word embeddings and deep networks with attention mechanisms in content analyzing face-to-face conversations.
- Our results suggest our methodology may be combined with automatic speech recognition methods to analyze face-to-face conversations and support real-time orchestration of CSCIL activities.

Keywords

Collaboration analytics, computational models, computer-supported collaborative learning, CSCL, CSCIL, deep networks, inquiry-based learning, word embedding

Submitted: 09/04/20 — **Accepted:** 19/01/21 — **Published:** 09/04/21

Corresponding author [1]Email: joni.lamsa@jyu.fi Address: Department of Education, University of Jyväskylä, P.O. Box 35, FI-40014 University of Jyväskylä, Finland. ORCID ID: https://orcid.org/0000-0001-7995-4090

[2]Email: pablo.uribe@student.ecp.fr Address: Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Periodista José Carrasco 75, Santiago, Chile. ORCID ID: https://orcid.org/0000-0002-4194-9189

[3]Email: abjimenez@gmail.com Address: Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Periodista José Carrasco 75, Santiago, Chile. ORCID ID: https://orcid.org/0000-0002-7041-284X

[4]Email: daniela.caballero@ciae.uchile.cl Address: Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Periodista José Carrasco 75, Santiago, Chile. ORCID ID: https://orcid.org/0000-0002-7319-3910

[5]Email: raija.h.hamalainen@jyu.fi Address: Department of Education, University of Jyväskylä, P.O. Box 35, FI-40014 University of Jyväskylä, Finland. ORCID ID: https://orcid.org/0000-0002-3248-9619

[6]Email: roberto.araya.schulz@gmail.com Address: Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Periodista José Carrasco 75, Santiago, Chile. ORCID ID: https://orcid.org/0000-0003-2598-8994

1. Introduction

Computer-supported collaborative inquiry-based learning (CSCIL) has the potential to support the development of competencies, such as collaboration, problem-solving, and critical thinking (Donnelly et al., 2014), that people graduating from higher education need today. To achieve the full potential of CSCIL and the reported benefits, however, scholars widely agree that guidance is required (Alfieri et al., 2011; Bell et al., 2010; de Jong & Lazonder, 2014). In CSCIL, the different types of guidance have been designed for the different inquiry-based learning (IBL) phases. For example, Bell et al. (2010) presented various technological tools aiming to support different phases of inquiry, such as hypothesis generation, planning, analysis, and conclusions. To provide timely guidance for learners, monitoring the phases of the learners' CSCIL process may provide essential insights for teachers and to improve technological learning environments. This monitoring could be supported by a better understanding of IBL phases. Our study is one attempt to address this goal. We present innovative computational models based on natural language processing (NLP) to automatically code manual transcripts of learners' face-to-face CSCIL conversations, with a particular focus on IBL phases (see Pedaste et al., 2015). To promote the automatic content analysis of face-to-face conversations across contexts, we pay attention to the relevant features of the computational models that carry valuable information about this automatizing procedure.

In the following section, we describe how researchers have so far applied automatic content analysis to study computer-supported collaborative learning (CSCL). Following that, we elaborate on how IBL in CSCL settings can benefit from collaboration analytics, particularly from automatic coding of IBL phases.

1.1. Automatic Content Analysis of Computer-Supported Collaborative Learning

The idea to automatize content analysis in CSCL contexts is not new. Over 15 years ago, Dönmez et al. (2005) applied automatic content analysis in the context of computer-mediated argumentative knowledge construction (see also Rosé et al., 2008). Dönmez et al. used over 1,000 text segments that they coded based on a coding scheme with seven dimensions. First, they tested k-nearest neighbors, a non-binary classifier, to assign a category to each text for each of the seven dimensions. Each dimension included between two and 35 categories, and the seven dimensions included 76 categories in total. Second, they ranked the best binary classifier by assessing various classifiers based on how they learned the 76 different categories. A few years later, Mu et al. (2012) took further steps and applied more advanced NLP techniques so that the proposed automatic content analysis solution would be 1) more context-independent and 2) able to automatically identify coded units of analysis. They presented a multilayer framework, Automatic Classification of Online Discussions with Extracted Attributes (ACODEA), that performed several classifications and followed a cascade from one layer to the next. The idea is to, first, extract attributes (syntactic and semantic) to segment and then, second, to code raw data.

In addition to the argumentative knowledge construction contexts, Xing et al. (2019) identified transformative and non-transformative discourse in a CSCIL context. They captured learner discourse from the computer-mediated interaction. First, humans coded 1,111 (of a total of 5,521) utterances. Next, they tested combinations from three different methods for extracting features from the text: 1) a qualitative insights-based method of regular expressions likely to belong to each category ("regex"), 2) a linguistic inquiry and word count program-based method (Pennebaker et al., 2015), and 3) a method based on a latent Dirichlet allocation topic model. In general, all three methods consisted of initially creating a fixed number of word sets (different for each method) and later computing, for each utterance, the number of words belonging to each set. After that, they evaluated four classical methods to be used as a classification algorithm: naïve Bayes, logistic regression, support vector machines, and decision trees (see Kotsiantis, 2007 for a detailed description of the models). Finally, they compared the precision and recall values of different combinations of textual features and classification algorithms. The precision value measures the relevance of the results given by the algorithm; that is, it is the ratio between the true positives and the sum of true positives and false positives. The recall value measures the fraction of the relevant results that the algorithm correctly classified; that is, the ratio between true positives and the sum of true positives and false negatives. For example, if every unit of analysis is automatically coded to the category X, the recall value of that category is 100% (no false negatives), but the precision value may provide poor values (depending on the number of false positives).

In this paper, we complement the existing studies on CSCIL in an authentic higher education physics setting where IBL is one of the most popular pedagogical approaches when CSCL is used (Jeong et al., 2019). Before we present the detailed aims of the study, we briefly conceptualize CSCIL.

1.2. Computer-Supported Collaborative Inquiry-Based Learning

CSCIL is a process in which technological resources facilitate and mediate negotiation among learners (Tan, 2018) when they follow scientific practices to solve problems (Pedaste et al., 2015). CSCIL processes include different phases with various activities, and many scholars have presented IBL frameworks featuring such variety (e.g., Bybee et al., 2006; White & Frederiksen, 1998). In this study, we apply the contemporary, concise, and widely used IBL framework that Pedaste et al. (2015) developed. These authors conducted a systematic literature review to synthesize various existing IBL models, resulting

in a process of inquiry comprising orientation, conceptualization, investigation, conclusion, and discussion phases. First, in the orientation phase, learners familiarize themselves with the given inquiry problem, its main variables and concepts, and available technological resources. Second, in the conceptualization phase, learners may identify dependent and independent variables to propose research questions (a sub-phase) or generate hypotheses (a sub-phase). Third, in the investigation phase, learners may explore the problem by planning the data collection based on the research question (a sub-phase), experiment by planning the data collection in order to test their hypotheses (a sub-phase), and analyze and interpret the data (a sub-phase). Fourth, in the conclusion phase, learners provide solutions to their research question or check whether the data support the hypotheses. Fifth, in the discussion phase, learners may communicate their findings and conclusions (a sub-phase) and reflect their CSCIL throughout the process or at the end of the process (a sub-phase). The IBL processes are not linear from the orientation to the discussion phases — instead, learners can move back and forth between the different phases. For example, when they start to investigate the problem, they may notice a need to reformulate their research questions or hypotheses (i.e., to re-conceptualize the problem).

Some of the phases, for example, the conceptualization phase, are particularly challenging for learners (Zacharia et al., 2015). While various tools to guide the different phases of CSCIL have been developed (see Bell et al., 2010; Zacharia et al., 2015), which in addition consider contextual factors (such as the nature of inquiry problems or the age of learners), implementing timely guidance requires that the phase of the learners' CSCIL processes be known. There are technological learning environments that allow the implementation of pre-structured CSCIL assignments based on the different IBL phases (e.g., Go-Lab; van Joolingen et al., 2005), but CSCIL taking place in face-to-face interaction may be less predictable (see Lämsä et al., 2018; 2020) and more fragmented than the computer-mediated interaction (Sins et al., 2011). Thus, there is a need to examine whether collaboration analytics can be applied to study these dynamic and sometimes fragmented face-to-face conversations. One promising approach to implement the collaboration analytics when studying CSCIL processes taking place in face-to-face interactions is to combine automatic speech recognition and automatic coding of the different IBL phases from learners' face-to-face conversations. While the current automatic speech recognition systems are promising (e.g., Kronholm et al., 2017), researchers have conducted automatic content analysis mostly in computer-mediated settings, and face-to-face interaction has been underrepresented in this field of study.

2. Research Aims

Within the last 15 years, researchers have increasingly executed collaboration analytics by conducting automatic content analysis in different computer-mediated CSCL contexts. These studies have indicated the many advantages of automatic coding of computer-mediated interaction by using various classical methods and algorithms (e.g., the possibility to analyze large data sets). On the other hand, CSCL still frequently takes place in face-to-face interaction. To design better models for CSCIL taking place in face-to-face interaction, we may have to make "key traces of activity visible to learners and their instructors or available to computational analysis" (Maldonado et al., 2019, p. 1044) by applying collaboration analytics.

In this study, we first aimed to examine whether and how automatic content analysis based on computational models with advanced NLP techniques (word embeddings and deep neural networks [DNN] with attention layers) could be applied to the study of face-to-face interaction in a CSCIL context. Second, since we assumed that the automatic coding of an IBL phase from an utterance requires computational models that can take many features into account, we aimed to identify these relevant features that carry valuable information to promote automatizing the content analysis of face-to-face interaction. First, we considered the linguistic context in which the utterance takes place — that is, the linguistic content of the previous and following utterances. Second, we considered the temporal context in which the utterance takes place — that is, whether the utterance emerges at the beginning of CSCIL, for example, when some of the IBL phases (e.g., orientation) are more probable than others (e.g., conclusion). Third, we considered that the quantitative context, which refers to the number of words in the utterance, might be indicative of some of the IBL phases. For example, in the discussion phase, the length of the utterance on average may be longer than in the conclusion phase since learners communicate their conclusions and reflect their CSCIL in the discussion phase. The sensitivity analysis of these different features may promote collaboration analytics research across the face-to-face interaction contexts since it can offer valuable insights on which parts of the interaction are relevant for determining the phase of the learning process based on manual transcriptions. This analysis may thus help researchers for whom analyzing video and audio recordings may be more time-consuming than analyzing transcripts. To address these aims, we answered the following research questions (RQs):

RQ1. How does the accuracy of the automatic coding differ in the various IBL phases when compared with human coding?

RQ2. How do the different features of the computational model improve the accuracy of the automatic coding in the different IBL phases?

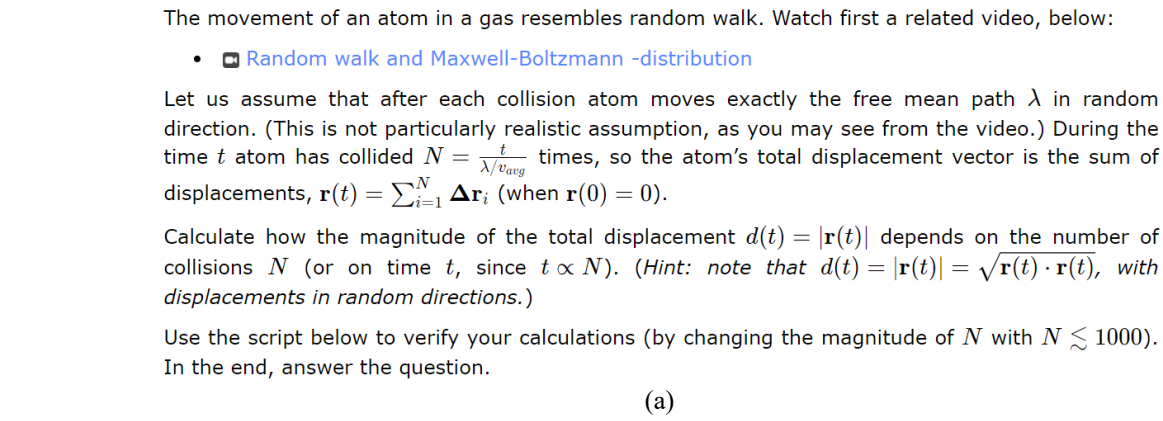
ISSN 1929-7750 (online).

3. Methods

3.1. Participants, Context, and Data

We collected the data from an introductory physics course offered by a Finnish university. The course content related to thermodynamics. The participants were 55 undergraduate students enrolled in the course. The students worked in small groups of five, so we focused on the CSCIL processes of 11 groups. The groups collaboratively solved inquiry problems weekly in a technological learning environment so that they worked face-to-face (see details of the course structure from Koskinen et al., 2018). In this study, we focused on the inquiry assignment (see Figure 1) to which the groups devoted, on average, the most time (22 min, *SD* = 11 min). This choice was justifiable to get more data (utterances) to train the computational models. The aim of the assignment was to study how the displacement of an atom in two-dimensional gas depends on time by using a Python program (Figure 1). In our previous studies (Lämsä et al., 2018; 2020), we identified tasks that learners should do in the different IBL phases. In the orientation phase, learners should identify the main concepts of the assignment and become familiar with the Python program. In the conceptualization phase, learners should determine the dependent variable (total displacement) and independent variables (amount of time and number of collisions). In the investigation phase, learners should plan the data collection procedure with the Python program, implement the procedure, and analyze and interpret the collected data. In the conclusion phase, learners should offer and evaluate solutions to the given question based on the data. Finally, in the discussion phase, learners should elaborate the findings and conclusions as well as reflect on the joint CSCIL process. CSCIL processes are rarely linear, but learners move back and forth between the different phases (see Section 1.2 and Lämsä et al., 2018; 2020).

We captured the CSCIL processes of 11 groups solving the random walk problem by screen-capturing videos of the computer screen and audio recordings. Based on the screen-captured videos and audio recordings, we manually transcribed the group conversations. The Finnish transcripts included on average 180 utterances per group.



The movement of an atom in a gas resembles random walk. Watch first a related video, below:

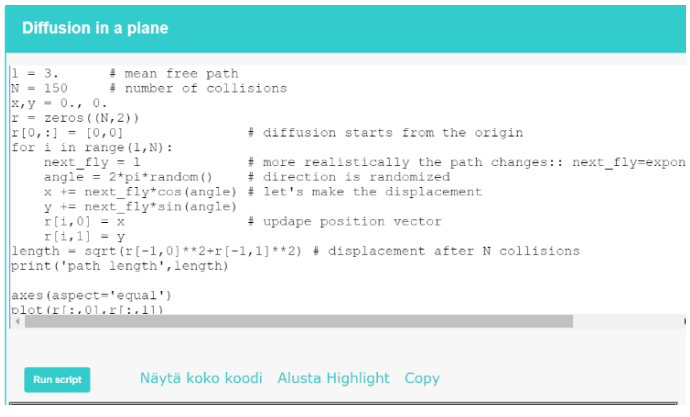
- Random walk and Maxwell-Boltzmann -distribution

Let us assume that after each collision atom moves exactly the free mean path λ in random direction. (This is not particularly realistic assumption, as you may see from the video.) During the time t atom has collided $N = \frac{t}{\lambda / v_{avg}}$ times, so the atom's total displacement vector is the sum of displacements, $\mathbf{r}(t) = \sum_{i=1}^{N} \mathbf{\Delta r}_i$ (when $\mathbf{r}(0) = 0$).

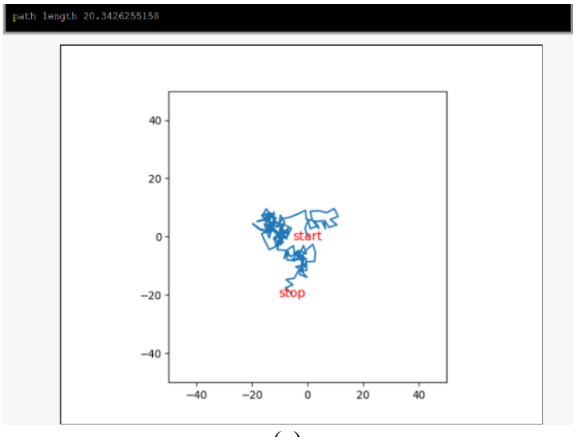
Calculate how the magnitude of the total displacement $d(t) = |\mathbf{r}(t)|$ depends on the number of collisions N (or on time t, since $t \propto N$). (*Hint: note that* $d(t) = |\mathbf{r}(t)| = \sqrt{\mathbf{r}(t) \cdot \mathbf{r}(t)}$, *with displacements in random directions.*)

Use the script below to verify your calculations (by changing the magnitude of N with $N \lesssim 1000$). In the end, answer the question.

(a)



(b)



(c)

Figure 1. (a) The assignment of the inquiry problem. (b) The Python program that learners used to solve the inquiry problem. (c) The output of the Python program, including the total displacement of an atom and plot of the atom's path.

ISSN 1929-7750 (online).

3.2. Analysis for RQ1: How Does the Accuracy of the Automatic Coding Differ in the Various IBL Phases When Compared with Human Coding?

To answer RQ1, the first author, who is familiar with the IBL framework (Lämsä et al., 2018; 2020), identified episodes from the transcripts that captured a "unit of meaning" (Henri, 1992; see details in Lämsä et al., 2018; 2020). An episode typically included a few utterances. Next, the first author conducted a theory-driven content analysis (Neuendorf, 2002) of all the episodes and coded them to the different IBL phases (orientation, conceptualization, investigation, conclusion, and discussion); here, each utterance in the episode was assigned the same code. To improve the reliability of this coding procedure, the first author consulted with the experts in content analysis and with those familiar with the IBL framework. The author and experts also consulted video and audio recordings when needed. We used these expert codings as a baseline for the precision and recall values of the human and automatic coding. In the following section, we explain both coding procedures.

3.2.1. Human Coding

In the human coding, a coder from outside of this study independently analyzed the transcripts of six groups (without access to the video and audio data) by coding the episodes the authors had identified. The coder assigned each utterance in the episode the same code. The transcripts from these six groups included 1,293 utterances that formed 65% of the whole data set. The human coder had not studied and acquainted himself with the IBL framework beforehand, but he read the article by Pedaste et al. (2015) and had a coding manual that included descriptions of the different IBL phases. To improve the validity of the coding, the manual also included authentic example episodes for each IBL phase and a comment on why the episode belonged to a specific IBL phase. To analyze the accuracy of the human coding in the various IBL phases, we calculated the precision and recall values separately for each IBL phase.

3.2.2. Automatic Coding

In the automatic coding, we split our data so that we used the transcripts from nine groups for training our computational models and another two group transcripts for testing these models. We used all the possible combinations of nine and two groups, leading to 55 different combinations, which we used for training and testing. Unlike with the human coding, the training and testing of the models did not include the information about the episodes to which individual utterances belonged. Altogether, we built five different models for which we calculated the overall accuracy and the precision and recall values separately for each IBL phase. In the following paragraphs, we elaborate on how we used advanced NLP techniques to build our five computational models. Further details are available at https://github.com/pabloveazul/CIBL.

We started to build the models by preprocessing the transcripts so that computational algorithms could handle them better. The preprocessing included 1) converting raw digits into words (for example, we turned "10" into "ten"), 2) removal of punctuation marks (except for the "?" which we considered as a word), 3) tokenization of the words so that we considered the 2,000 words that appeared most frequently in the transcripts (in total, we found 3,500 distinct words), and 4) padding of the utterances to 20 words so that the length of all the utterances was the same (we truncated 254 utterances and padded 1,700 utterances with the token "0").

Concerning the classifier algorithm, our procedure was based on two relevant components. First, we obtained textual features through a word embedding model (see Mikolov et al., 2013). This model obtained high-dimensional vector representations from words — called word embeddings — by capturing their syntactic and semantic relationships (Tshitoyan et al., 2019). Since the Finnish transcripts of the 11 groups' conversations represented a relatively small corpus of around 20,000 words, we used a pre-trained word embedding model trained on a large Finnish corpus of over 4.5 billion words (see the TurkuNLP project; Luotolahti et al., 2015). Second, we built an automatic classification algorithm consisting of a DNN with an embedding layer and an attention layer on top of it. To build our five models, we explored different configurations of the widely used embedding and attention layers (Hu, 2020). We used a categorical cross-entropy loss function and an adaptive moment estimation optimizer (Kingma & Ba, 2014) during the training process.

The embedding layer associates its respective high dimensional vector from the TurkuNLP project to each word input. These vectors can be adjusted through backpropagation during the training process, when desired, to be a trainable embedding (TE) configuration or to remain static during the training process as a static embedding (SE) configuration. It is important to note that in the first case, word embeddings will be adjusted according to the conversations and the language use within our data set. However, the cost for that is adding a huge number of trainable parameters to the model, specifically in the embedding layer. In our case, we gave the previous, current, and following utterances as inputs for the embedding layer, and we concatenated their outputs into a single sequence of 60-word vectors e = [e1, . . ., e60]. There were also valid reasons to include an attention mechanism in our model. A standard neural network consists of a series of layers, which sequentially apply nonlinear transformations to the output of the previous layer (or to the input in the case of the first layer) to produce fixed-dimensional hidden states, until the last layer computes the output. When the input belongs to a large space, this model may encounter difficulties in controlling the interaction among components (Kim et al., 2017).

In our case, words were numerically represented as high dimensional vectors, so that utterances could be interpreted as a sequence of vectors (or a matrix) of a large space. For example, merely adding these vectors to reduce the dimensionality would lead to an information bottleneck, as we will explore further. Adding an attention layer, however, provides an alternative approach. The attention layer consists of an internal inference mechanism that performs a soft selection over the previous representations (Kim et al., 2017). For instance, for previous hidden states consisting of a sequence of word vectors, this mechanism will effectively soft select the important words according to the classification task. In our study, we tested two attention mechanisms, a simple attention (SA) configuration and a differentiated attention (DA) configuration, in addition to a model with no attention mechanism (see Appendix). Incorporating these layers adds interpretability to the models, allowing us to identify which parts of the textual features are the most relevant for the classification task according to our algorithms (see Appendix).

First, the SA configuration mechanism (see formulae 1–5 below) consists of a single-layer perceptron (SLP) with a one-dimensional output w_t — called an attention weight — that is applied word-wise (i.e., to each word vector e_t). These attention weight outputs are later normalized using a softmax function to obtain probabilities (noted *att(e)*) proportional to their exponentials. This probability distribution corresponds to a soft selection of words where each probability can be interpreted as the importance of each word for the classification task. Next, each word vector e_t is multiplied by its attention probability $att(e)_t$ to obtain a weighted sequence. Finally, the resulting vectors are summed up into one single vector s, called the context vector. This context vector is fed up, together with the number of words n and the relative position r of the current utterance, as an input into a multi-layer perceptron (MLP) with one hidden layer and a five-dimensional output layer, so that the final prediction is modelled as a softmax distribution over the five different categories.

$$(1)\quad w_t = SLP(e_t)$$
$$(2)\quad att(e)_t = softmax(w)_t = exp(w_t)/\sum_{i=1}^{60} exp\,(w_i)$$
$$(3)\quad s = \sum_{i=1}^{60} att(e)_i e_i$$
$$(4)\quad y = MLP(s, r, n)$$
$$(5)\quad \hat{y} = softmax(y)$$

However, one may think that the soft selection of the words should depend on the final coding. For instance, an utterance corresponding to the conceptualization phase may contain word vectors corresponding to concepts, while an utterance corresponding to the investigation phase may contain word vectors associated with numbers. Thus, second, we tested the DA configuration mechanism (see formulae 6–10), where each category c is connected with one independent attention mechanism; that is, five different SLPs (noted as SLP_c) perform a soft selection of word vectors, after which five different context vectors s_c are obtained independently through the same process explained previously. Next, each context vector is fed up together with the number of words n and the relative position r of the utterance as an input into an independent MLP (noted as MLP_c) with one hidden layer and a single output y_c. Finally, a softmax layer is applied over the five different outputs to obtain the final prediction as a probability distribution. We present an example of the different attention mechanisms in the Appendix.

$$(6)\quad w_t^c = SLP_c(e_t)$$
$$(7)\quad att_c(e)_t = softmax(w^c)_t = exp\,(w_t^c)/\sum_{i=1}^{60} exp\,(w_i^c)$$
$$(8)\quad s_c = \sum_{i=1}^{60} att_c(e)_i e_i$$
$$(9)\quad y_c = MLP_c(s_c, r, n)$$
$$(10)\quad \hat{y} = softmax(y_1, y_2, y_3, y_4, y_5)$$

Third, we tested a model with no attention mechanism (see formulae 11–13), which simply performs a non-weighted sum of all word embeddings into one single vector. This vector is later fed up together with the other inputs into an MLP with one hidden layer and a five-dimensional output layer. We later model the prediction as a softmax distribution over the five different categories.

$$(11)\quad s = \sum_{i=1}^{60} e_i$$
$$(12)\quad y = MLP(s, r, n)$$
$$(13)\quad \hat{y} = softmax(y)$$

3.3. Analysis for RQ2: How Do the Different Features of the Computational Model Improve the Accuracy of the Automatic Coding in the Different IBL Phases?

To answer RQ2, we made a sensitivity analysis by evaluating our best performing model. We removed the input features of the model and compared the performance of the reduced model with the full model. We also separately examined how the

different features of the model (the linguistic context, temporal context, and quantitative context) improved the accuracy of the automatic coding. First, concerning the linguistic context, we compared the reduced model with the model whose input included the previous and following utterances but neither the relative position nor the number of words of the current utterance. Second, concerning the temporal context, we compared the reduced model with the model whose input included the relative position of the current utterance but neither the previous and following utterances nor the number of words of the current utterance. Third, concerning the quantitative context, we compared the reduced model with the model whose input included the number of words of the current utterance but neither the previous and following utterances nor the relative position of the current utterance.

4. Results

In Section 4.1, we compare the results of the automatic and human coding when identifying different IBL phases from learners' authentic face-to-face conversations. We present the results of the automatic coding separately for the various tested computational models. In Section 4.2, we analyze how the different features of the best performing computational model improve the accuracy of the coding.

4.1. Automatic and Human Coders Perform Well in the Most Frequent IBL Phases

Table 1 shows the averaged precision and recall values of five computational models and the human coder. When focusing on the human codings, the precision and recall values in Table 1 are based on transcripts from six groups (1,293 utterances). The overall accuracy (54.3%) indicates moderate agreement with the baseline codings. When focusing on the automatic coding results, we compared the performance of five computational models whose overall accuracy varied between 48.8% and 58.9%. Thus, the overall accuracy of the best performing model, differentiated attention with pre-trained static embeddings (MLP+SEDA), was 4.6 percentage points higher than that of the human coder (58.9% vs. 54.3%). We showcase in the Appendix an example utterance of the orientation phase that illustrates how the differentiated attention mechanism performed better than the simple attention mechanism.

Table 1 also shows that, first, the accuracy of both the automatic and the human coding was better in the orientation, investigation, and discussion phases than in the conceptualization and conclusion phases. Since the conceptualization and conclusions phases were rarer in the transcripts than the orientation, investigation, and discussion phases, the poor accuracy in the conceptualization and conclusion phases did not significantly decrease the overall accuracy. Second, the precision value in the investigation phase was the highest in the automatic coding. While the recall value of the investigation phase was the highest in human coding, the recall value of the discussion phase was the highest in every computational model. The significantly higher recall value of the discussion phase compared with the precision value indicates that the number of false positives (utterances that are automatically coded to the discussion phase but belong to another IBL phase) is higher than the number of false negatives (utterances of the discussion phase that are automatically coded to another IBL phase) in the computational models. Furthermore, if we compare the results from the TE and SE configurations and exclude the conceptualization and conclusion phases, the MLP+TE model shows similar results to the MLP+SE model. However, when we incorporate a simple attention layer, the MLP+SESA model performs better than the MLP+TESA and the previous models.

Table 1. Averaged Precision (P) and Recall (R) Values of Five Computational Models and the Human Coder

Model/ Phase	MLP+SE		MLP+SESA		MLP+SEDA		MLP+TE		MLP+TESA		Human Coder	
	P	R	P	R	P	R	P	R	P	R	P	R
Orientation	44.8	44.0	58.1	63.5	60.1	66.6	47.6	52.9	44.8	44.0	69.8	44.4
Conceptualization	42.5	16.6	38.9	9.0	47.3	19.2	40.7	25.5	42.5	16.6	26.7	57.4
Investigation	55.5	49.3	66.9	63.3	67.2	64.2	56.5	50.9	55.5	49.3	64.8	80.9
Conclusion	16.7	2.0	31.5	3.0	24.2	5.2	22.9	6.5	16.7	2.0	24.0	51.2
Discussion	48.5	66.8	52.7	69.0	55.4	67.9	52.2	61.9	48.5	66.8	62.0	40.1

Note. All numeric values in the table are expressed as percentages (%). The values of the human coder are based on coding 65% of the data. Abbreviations of the model names are as follows: static embeddings (SE), trainable embedding (TE), simple attention (SA), differentiated attention (DA), and multilayer perceptron (MLP).

4.2. Considering the Linguistic and Temporal Context Improves the Accuracy of Automatic Coding

Table 2 shows the associations of the model features of the differentiated attention with pre-trained static embeddings (MLP+SEDA) with the averaged precision and recall values in the different IBL phases. First, the accuracy of the model that considered the linguistic context was close to the full model except in the orientation phase. Second, taking the temporal context into account significantly improved the precision and recall values in the orientation and conclusion phases. Third,

taking the quantitative context into account slightly improved the recall value of the conclusion phase; otherwise, this model feature did not have a significant effect on the coding accuracy. Thus, it seems that the linguistic and temporal contexts were essential features when considering the accuracy of the computational model, while the quantitative context of the utterance did not significantly improve the accuracy.

Table 2. Precision (P) and Recall (R) Average Values of the MLP+SEDA Model on the Test Set with Model Features Considering the Temporal, Linguistic, and Quantitative Contexts

Model/ Phase	Full model		Model considering linguistic context		Model considering temporal context		Model considering quantitative context		Reduced model	
	P	R	P	R	P	R	P	R	P	R
Orientation	60.1	66.6	51.7	50.7	55.4	65.3	46.1	40.6	46.5	43.8
Conceptualization	47.3	19.2	45.9	17.4	40.6	11.3	36.4	11.7	42.6	12.3
Investigation	67.2	64.2	66.0	65.1	69.9	49.7	70.5	48.9	69.8	49.6
Conclusion	24.2	5.2	22.0	64.2	39.9	24.2	36.2	21.4	36.5	17.5
Discussion	55.4	67.9	52.3	67.7	50.2	67.7	45.4	70.7	46.2	70.5

Note. All numeric values in the table are expressed as percentages (%).

Finally, Table 2 shows that the full model compared with the reduced model did not improve the precision value in the investigation phase. The recall value in the investigation phase, however, was higher in the full model than in the most reduced model (64.2% vs. 49.6%). These findings are explained by the confusion matrix (Table 3), which shows that the full model correctly coded more utterances of the investigation phase than the reduced model (16.0% vs. 12.0%) and that the full model had fewer false negatives than the reduced model (8.9% vs. 12.7%, see the "Investigation" row in Table 3). The full model, however, had slightly more false positives — that is, the utterances (whatever their IBL phase) were more frequently coded into the investigation phase — than the reduced models (7.8% vs. 5.3%, see the "Predicted investigation" column in Table 3). Contrary to the investigation phase, the full model improved the precision value but not the recall value of the discussion phase compared to the most reduced model. The full model had fewer false positives as it coded fewer utterances of the orientation (5.6%) and investigation phases (5.9%) into the discussion phase than the most reduced model (12.0% and 9.3%).

Table 3. Confusion Matrix

Full model & reduced model/ Phase	Predicted orientation	Predicted conceptualization	Predicted investigation	Predicted conclusion	Predicted discussion
Orientation	**17.0 & 11.0**	0.8 & 0.5	1.7 & 1.3	0.1 & 0.1	5.6 & 12.0
Conceptualization	2.4 & 3.1	**2.1 & 1.4**	1.4 & 0.8	0.0 & 0.2	5.0 & 5.6
Investigation	2.6 & 2.8	0.4 & 0.4	**16.0 & 12.0**	0.0 & 0.1	5.9 & 9.3
Conclusion	0.0 & 0.6	0.2 & 0.2	0.5 & 0.1	**0.2 & 0.6**	2.7 & 2.0
Discussion	6.0 & 5.9	0.7 & 0.6	4.2 & 3.1	0.4 & 0.6	**24.0 & 25.0**

Note: All numeric values in the table are expressed as percentages (%). Each cell (row *i*, column *j*) is the average number of utterances that belong to the IBL phase *i* and that the full and reduced MLP+SEDA model predicted to belong to the IBL phase *j*. The correctly coded utterances in each IBL phase have been bolded.

5. Discussion

While CSCL researchers have used classical models and algorithms to illustrate the potential of automatic coding of computer-mediated interaction over the last 15 years, few studies have examined whether automatic coding could be applied to study CSCL in face-to-face interaction. We used the manual transcriptions of CSCIL processes and showed that the computational models based on advanced NLP techniques were able to code utterances to the different IBL phases despite the dynamic and fragmented nature of face-to-face interaction. We found, however, differences pertaining to the various IBL phases. While the accuracy of the automatic and human coding was better in the orientation, investigation, and discussion phases, there were challenges in both human and automatic coding of the conceptualization and conclusion phases (RQ1, see Table 1). The poorer performance in the conceptualization and conclusion phases relates to the infrequent emergence of these phases in learners' conversations. Thus, the computational models did not properly learn the characteristic features of these phases. As formulating research problems or hypothesis (conceptualization phase) and making justified and evidence-based conclusions may include similar features across contexts, the training set could have included utterances related to these two phases from various

contexts. Improving the accuracy in these two phases is crucial to promoting the guidance of CSCIL, as it is known that learners may need support to formulate proper research questions and hypotheses (van Joolingen et al., 2005) and to jointly build explanations for research problems (Matuk & Linn, 2018).

To promote collaboration analytics research across face-to-face interaction contexts, we also examined the best performing computational model to better understand what features of the model improved its accuracy (RQ2, see Table 2). First, the accuracy of the automatic coding improved when we considered the linguistic context in which the utterance emerged; that is, we took into account both the previous and the following utterances. When we considered the linguistic context, the recall value of the investigation phase improved by 15%. This finding might indicate that the computational models considering the linguistic context recognized many activities of the investigation phase (e.g., planning and data interpretation) in addition to mere data collection that was revealed by the use of numbers (learners collected data with the Python program; see Figure 1). Second, the accuracy of the automatic coding improved when we considered the temporal context. This improvement was particularly visible in the orientation and conclusion phases (see Table 2) that are typically present at the beginning (orientation) or end (conclusion) of the CSCIL process. The importance of the temporal context indicates that linguistically similar utterances may have different meanings in the various stages of the learning process (Lemke, 2000; Mercer, 2008). For example, the computational models had difficulties in making the distinction between the orientation and discussion phases without considering the temporality (see Tables 2 and 3). Our previous findings based on manual content analysis also show that considering the temporal context may be essential when designing guidance and analyzing its role in CSCIL (Lämsä et al., 2018; 2020).

Our study, however, has limitations. First, even though the accuracy of the MLP+SEDA was moderate in the orientation, investigation, and discussion phases, the automatic classifier performed poorly in the conceptualization and conclusion phases. When we consider the fragmented nature of face-to-face interaction, however, the precision and recall values of the MLP+SEDA model can be considered sufficient. In fact, the human coder also faced challenges in his analysis when he only had access to the transcripts and not to the video and audio data (see details in Section 3.2.1). Second, the results of the automatic and human codings were not entirely commensurable. The human coder outside of this study coded 65% of the data (six group-working sessions) from which we calculated the precision and recall values for the different IBL phases (Table 1), while the automatic coding results were based on the test sets of two group-working sessions. Moreover, the human coder coded the episodes, not individual utterances, contrary to the computational models (see details in Sections 3.2.1 and 3.2.2). Even though Knight and Littleton (2015, p. 122) have highlighted the challenge of focusing on individual utterances, we coded each utterance to an IBL phase. We recognized, however, the sequential nature of the interaction in which each utterance gets input from the previous utterances and provides output to the following utterances. We approximated this sequentiality by considering the previous and following utterances when coding the utterance under interest. Even though this approximation simplified the true sequentiality of the interaction (e.g., a learner may have referred to an utterance that had been said many turns previously), it significantly improved the performance of the model (see the effect of the linguistic context in Table 2). Third, the workflow of that approximation started by concatenating the previous, current, and following utterances. The next steps (attention, sum) were commutative so we lost the order of the previous, current, and following utterances. Thus, the linguistic context would be the same for any permutation of the previous, current, and following utterances (e.g., [A, B, C] or [C, A, B], see the context vector in formula 3).

Despite these limitations, our study provides several implications. In respect to methodological implications, our approach for automatic coding provides a novelty that can advance learning sciences and collaboration analytics in face-to-face interaction in particular. First, concerning the textual feature creation, our model incorporated an embedding layer that associated word vectors with words. These word vectors can be learned through backpropagation during the training process ("feature learning"). Thus, we did not need to create features from the words beforehand ("feature engineering") as in many other current models (e.g., Espinoza et al., 2019; Xing et al., 2019). We illustrated how the proposed attention layer improved the accuracy of the computational models and eased the interpretation of the results (see Appendix). This methodological choice also showed the potential of small data sets since we could use pre-trained textual features derived from an existing large data corpus (Turku NLP embeddings). We showed that these features are powerful enough for our classification task without the need to adjust them for our problem, as the SE models reported a similar or even better performance than the TE models (see Table 1). Second, concerning the classification algorithm, we used deep networks that can model more complex functions than many widely used classical algorithms (e.g., support vector machines or decision trees) and enabled us to have the advantages of the embedding and attention layers explained previously.

In respect to practical implications, our study may advance the design of adaptive guidance of CSCIL. Our approach for automatic coding could be combined with automatic speech recognition solutions that can transcribe learners' discussions (e.g., Kronholm et al., 2017). Thus, visualizations of the results of automatic coding (Lämsä et al., 2018; Thompson et al., 2013) could be presented in the dashboards of teachers or learners. From the teachers' perspective, these dashboards could support their orchestration of CSCIL activities in real time (e.g., facilitating the guidance of learners and provision of feedback

for learners; see van Leeuwen, 2015). From the learners' perspective, these dashboards could be used as awareness tools. Even though the computational models cannot have all the information available (such as the relative position of the utterance or the content of the following utterance) in real time, teachers might be able to estimate the length of the discussion with reasonable precision beforehand, and this estimation could be used in the model. Moreover, the lag of few utterances when visualizing the results of the automatic coding may not have significant implications considering these practical applications.

6. Conclusion and Future Work

In this study, we complemented previous literature focused on the automatic content analysis of CSCL. First, instead of focusing on the computer-mediated interaction, we automatically analyzed the transcripts of face-to-face conversations. Second, instead of using classical methods and algorithms to automatize the content analysis, we applied novel computational models with advanced NLP techniques (e.g., pre-trained static word embeddings and DNN with differentiated attention layers) and studied which features of the models may be essential in analyzing the face-to-face interaction. We illustrated how to capture IBL phases of face-to-face conversations with good accuracy even from a relatively small data set. Our findings also indicate that it is crucial to consider the temporality of learning when developing collaboration analytics. In addition to considering the temporal context in which the utterances take place, it may be useful to implement independent attention mechanisms for the previous, current, and following utterance inputs so that linguistic context can be distinguished from the utterance under interest.

Besides our focus on what was said in the conversations, scholars could also analyze prosodic features of speech straight from the audio data (how it is said, Hämäläinen et al., 2018; Smith et al., 2016). In CSCIL contexts, for example, learners may form common ground in the orientation phase by using cumulative patterns of talk (Mercer et al., 1999), and when the talk is cumulative, talkers may use small pitch variation and repeat specific word stress patterns (Hämäläinen et al., 2018). This information about the prosodic context could thus be added to the computational models (see RQ2). Besides the content and prosody of the speech of learners, many other properties may have a significant role in face-to-face interaction; for example, gestures (Schneider & Blikstein, 2014), facial expressions (Worsley & Blikstein, 2015), visual attention (Schneider & Pea, 2013, 2015), the learner's physical location and its variation regarding other learners, teachers, and learning resources (Howard et al., 2017). Using multimodal learning analytics (see, e.g., Olsen et al., 2020) to support automatic content analysis could thus decrease the gap between the baseline coding (in which video and audio recordings were used on demand, see Section 3.2) and automatic coding. Despite the apparent potential of multimodal data and its analysis in understanding the CSCIL activities (Noroozi et al., 2019), we believe that basic research is also needed to study different data modalities and analysis techniques separately.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the Academy of Finland [grant numbers 292466 and 318095, the Multidisciplinary Research on Learning and Teaching profiles I and II of University of Jyväskylä] and ANID/PIA/Basal Funds for Excellence (FB0003).

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. https://doi.org/10.1037/a0021017

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education*, *32*(3), 349–377. https://doi.org/10.1080/09500690802582241

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Powell, J. C., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins and effectiveness.* Biological Sciences Curriculum Study.

de Jong, T., & Lazonder, A. W. (2014). The guided discovery learning principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 371–390). Cambridge, UK: Cambridge University Press.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In T. Koschmann, D. D. Suthers, & T.-W. Chan (Eds.), *Learning 2005: The Next 10 Years! Proceedings of the 2005 Conference on Computer Support for Collaborative Learning* (CSCL '05), May 30–June 4 2005, Taipei, Taiwan (pp. 125–134). International Society of the Learning Sciences.

Donnelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and characteristics of computer-based science inquiry learning environments for precollege students. *Review of Educational Research*, *84*(4), 572–608. https://doi.org/10.3102/0034654314546954

Espinoza, C., Lämsä, J., Araya, R., Hämäläinen, R., Jiménez, A. G., Gormaz, R., & Viiri, J. (2019). Automatic content analysis in collaborative inquiry-based learning. In O. Levrini & G. Tasquier (Eds.), *The Beauty and Pleasure of Understanding: Engaging with Contemporary Challenges Through Science Education. Proceedings of the European Science Education Research Association Conference* (ESERA 2019), 26–30 August, 2019, Bologna, Italy (Part 18, pp. 2041–2050). University of Bologna. https://www.esera.org/publications/esera-conference-proceedings/esera-2019

Hämäläinen, R., De Wever, B., Waaramaa, T., Laukkanen, A., & Lämsä, J. (2018). It's not only what you say, but how you say it: Investigating the potential of prosodic analysis as a method to study teacher's talk. *Frontline Learning Research*, *6*(3), 204–227. https://doi.org/10.14786/flr.v6i3.371

Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *Collaborative learning through computer conferencing: The Najadan Papers* (pp. 117–136). Springer-Verlag.

Howard, S. K., Thompson, K., Yang, J., Ma, J., Pardo, A., & Kanasa, H. (2017). Capturing and visualizing: Classroom analytics for physical and digital collaborative learning processes. In B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (Eds.), *Making a Difference: Prioritizing Equity and Access in CSCL. Proceedings of the 12th International Conference on Computer Supported Collaborative Learning* (CSCL 2017) 18–22 June 2017, Philadelphia, PA, USA (Vol. 2, pp. 801–802). International Society of the Learning Sciences.

Hu, D. (2020). An introductory survey on attention mechanisms in NLP problems. In Y. Bi, R. Bhatia, & S. Kapoor (Eds.), *Intelligent systems and applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)* (pp. 432–448). Springer.

Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review*, *28*, 100284. https://doi.org/10.1016/j.edurev.2019.100284

Kim, Y., Denton, C., Hoang, L., & Rush, A. M. (2017). Structured attention networks. https://arxiv.org/abs/1702.00887

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980

Knight, S., & Littleton, K. (2015). Discourse centric learning analytics: Mapping the terrain. *Journal of Learning Analytics*, *2*(1), 185–209. https://doi.org/10.18608/jla.2015.21.9

Koskinen, P., Lämsä, J., Maunuksela, J., Hämäläinen, R., & Viiri, J. (2018). Primetime learning: Collaborative and technology-enhanced studying with genuine teacher presence. *International Journal of STEM Education, 5(*20), 1–13. https://doi.org/10.1186/s40594-018-0113-8

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*, *31*(3), 249–268.

Kronholm, H., Caballero, D., Mansikkaniemi, A., Araya, R., Lehesvuori, S., Pertilä, P., Virtanen, T., Kurimo, M., & Viiri, J. (2017). The automatic analysis of classroom talk. *Proceedings of the Annual FMSERA Symposium 2016*, Joensuu, Finland (pp. 142–151). https://journal.fi/fmsera/article/view/60940/27049

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*, 103674. https://doi.org/10.1016/j.compedu.2019.103674

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. https://doi.org/10.1080/09500693.2018.1506594

Lemke, J. L. (2000). Across the scales of time: Artifacts, activities, and meanings in ecosocial systems. *Mind, Culture, and Activity*, *7*(4), 273–290. https://doi.org/10.1207/S15327884MCA0704_03

Luotolahti, J., Kanerva, J., Laippala, V., Pyysalo, S., & Ginter, F. (2015). Towards universal web parsebanks. *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), 24–26 August 2015, Uppsala, Sweden (pp. 211–220). https://www.aclweb.org/anthology/W15-21

Maldonado, R. M., Worsley, M., Schneider, B., & Kharrufa, A. (2019). International workshop on collaboration analytics: Making learning visible in collaborative settings. In K. Lund, G. Niccolai, E. Lavoué, C. Hmelo-Silver, G. Gweon, & M. Baker (Eds.), *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings. Proceedings of the 13th International Conference on Computer Supported Collaborative Learning* (CSCL 2019), 17–21 June 2019, Lyon, France (Vol. 2, p. 1044). International Society of the Learning Sciences.

Matuk, C., & Linn, M. C. (2018). Why and how do middle school students exchange ideas during science inquiry? *International Journal of Computer-Supported Collaborative Learning, 13*(3), 263–299. https://doi.org/10.1007/s11412-018-9282-1

Mercer, N. (2008). The seeds of time: Why classroom dialogue needs a temporal analysis. *Journal of the Learning Sciences*, *17*(1), 33–59. https://doi.org/10.1080/10508400701793182

Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk and the development of reasoning in the classroom. *British Educational Research Journal*, *25*(1), 95–111. https://doi.org/10.1080/0141192990250107

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

Mu, J., Stegmann, K., Mayfield, E., Rosé, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, *7*(2), 285–305. https://doi.org/10.1007/s11412-012-9147-y

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

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, *51*(5), 1527–1547. https://doi.org/10.1111/bjet.12982

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* University of Texas at Austin. http://hdl.handle.net/2152/31333

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. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237–271. https://doi.org/10.1007/s11412-007-9034-0

Schneider, B., & Blikstein, P. (2014). Unraveling students' interaction around a tangible interface using gesture recognition. In J. Stamper et al. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July 2014, London, UK (pp. 320–323). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.3554729

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 375–397. https://doi.org/10.1007/s11412-013-9181-4

Schneider, B., & Pea, R. (2015). Does seeing one another's gaze affect group dialogue? A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. https://doi.org/10.18608/jla.2015.22.9

Sins, P. H. M., Savelsbergh, E. R., van Joolingen, W. R., & van Hout-Wolters, B. H. A. M. (2011). Effects of face-to-face versus chat communication on performance in a collaborative inquiry modeling task. *Computers & Education*, *56*(2), 379–387. https://doi.org/10.1016/j.compedu.2010.08.022

Smith, J., Bratt, H., Richey, C., Bassiou, N., Shriberg, E., Tsiartas, A., D'Angelo, C., & Alozie, N. (2016). Spoken interaction modeling for automatic assessment of collaborative learning. *Proceedings of Speech Prosody 8*, 31 May–3 June 2016, Boston, MA, USA (pp. 277–281). https://doi.org/10.21437/SpeechProsody.2016-57

Tan, E. (2018). Effects of two differently sequenced classroom scripts on common ground in collaborative inquiry learning. *Instructional Science*, *46*(6), 893–919. https://doi.org/10.1007/s11251-018-9460-6

Thompson, K., Ashe, D., Carvalho, L., Goodyear, P., Kelly, N., & Parisio, M. (2013). Processing and visualizing data in complex learning environments. *American Behavioral Scientist*, *57*(10), 1401–1420. https://doi.org/10.1177/0002764213479368

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior*, *21*(4), 671–688. https://doi.org/10.1016/j.chb.2004.10.039

van Leeuwen, A. (2015). Learning analytics to support teachers during synchronous CSCL: Balancing between overview and overload. *Journal of Learning Analytics*, *2*(2), 138–162. https://doi.org/10.18608/jla.2015.22.11

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118. https://doi.org/10.1207/s1532690xci1601_2

Worsley, M., & Blikstein, P. (2015). Using learning analytics to study cognitive disequilibrium in a complex learning environment. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 426–427). New York: ACM. https://doi.org/10.1145/2723576.2723659

Xing, W., Popov, V., Zhu, G., Horwitz, P., & McIntyre, C. (2019). The effects of transformative and non-transformative discourse on individual performance in collaborative-inquiry learning. *Computers in Human Behavior*, *98*, 267–276. https://doi.org/10.1016/j.chb.2019.04.022

Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S. A. N., Kamp, E. T., Mäeots, M., Siiman, L., & Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: A literature review. *Educational Technology Research and Development*, *63*(2), 257–302. https://doi.org/10.1007/s11423-015-9370-0

Appendix: Simple vs. Differentiated Attention Mechanisms

In Figures 2 and 3, we can see the results of the attention mechanisms in the simple and differentiated attention cases. Current utterance (number 120 in a group's transcript) is marked as black at the middle, while the previous and next utterances are grey on the left and right side of the current utterance, respectively. Blank spaces correspond to the padding token added during the data preprocessing. In the first case, a single curve represents the soft selection of words that the model performs using the simple-attention mechanism. In the second case, each coding category possesses an independent attention mechanism that performs the soft selection of words. Therefore, five different curves represent the differentiated-attention mechanism. The simple-attention model misclassified the utterance into the discussion phase, while the differentiated-attention mechanism model correctly classified it into the orientation phase. In particular the attention mechanism proper to the orientation phase (blue line) is giving high attention values to the words' "assignments" (translated from Finnish) of the previous utterance, in contrast to the simple attention case. This showcases how the differentiated attention mechanism captured an essential feature of the orientation phase: Students should familiarize themselves with the given assignment.

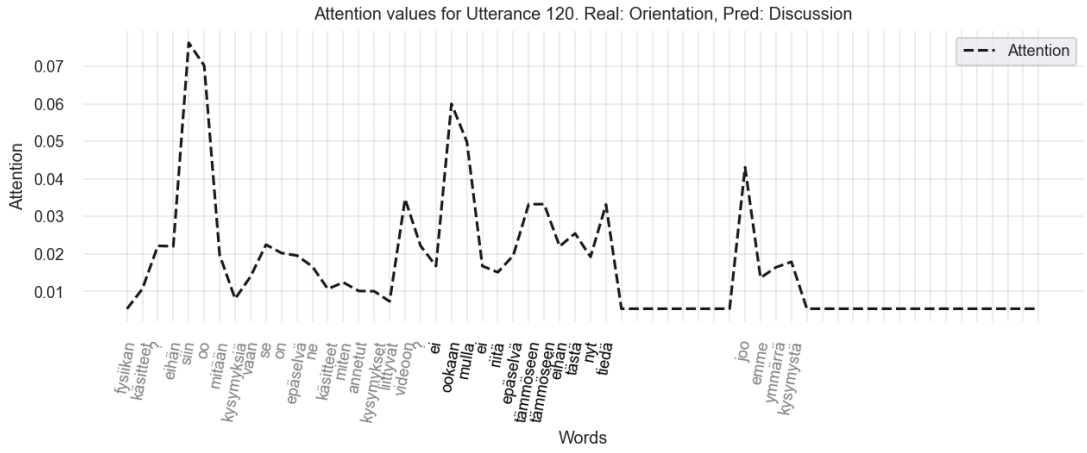


Figure 2. Example of the attention values of the simple attention mechanism.

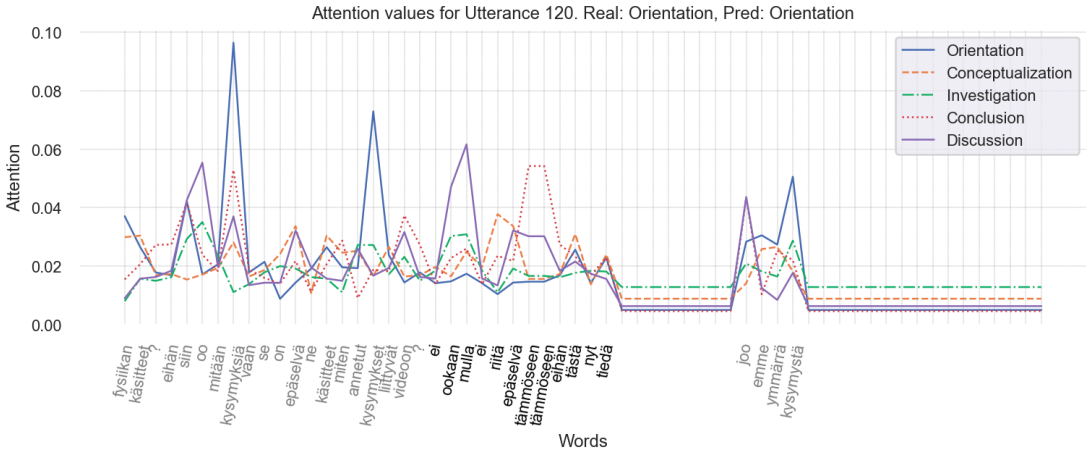


Figure 3. Example of the attention values of the differentiated attention mechanism.