International Society of the Learning Sciences

Taking Transactivity Detection to a New Level

James Fiacco, Ki-Won Haan, Anita Williams Woolley, Carolyn Rosé
jfiacco@cs.cmu.edu, khaan@andrew.cmu.edu, awoolley@andrew.cmu.edu, cprose@cs.cmu.edu
Carnegie Mellon University

Abstract: Transactivity is a valued collaborative process, which has been associated with elevated learning gains, collaborative product quality, and knowledge transfer within teams. Dynamic forms of collaboration support have made use of real time monitoring of transactivity, and automation of its analysis has been affirmed as valuable to the field. Early models were able to achieve high reliability within restricted domains. More recent approaches have achieved a level of generality across learning domains. In this study, we investigate generalizability of models developed primarily in computer science courses to a new student population, namely, masters students in a leadership course, where we observe strikingly different patterns of transactive exchange than in prior studies. This difference prompted both a reformulation of the coding standards and innovation in the modeling approach, both of which we report on here.

Introduction

Research shows that students benefit from rich discussions with other students in learning environments (Ferschke et al., 2015). Consequently, for more than a decade, researchers in the field of learning science have developed many frameworks for automated analysis of student discussion, as it has repeatedly been shown to be valuable for assessing student learning (McLaren et al., 2007; Dascalu et al., 2015; Joshi & Rosé, 2007; Rosé et al., 2008; McLaren et al., 2007; Ai et al., 2010; Gweon et al., 2013; Fiacco & Rosé, 2018), supporting group learning (Kumar et al., 2007), and enabling effective group assignments (Wen et al., 2016). Some work has explicitly addressed the issue of whether these frameworks generalize across domains (Mu et al., 2012; Fiacco & Rosé, 2018), which is critical to enabling educators in a variety of fields to leverage these tools. While cross-domain generalizability may sound like a purely technical problem, what we find in the current study is that the characteristics of different student populations and their learning processes, as well as the interplay between the two, are critical components. Here we report on the elements required to generalize technology developed for automated analysis of transactivity from one student population to another.

In this work, we focus on transactivity as a quality of conversational behavior where students explicitly build on ideas and integrate reasoning previously presented during the conversation. Transactivity stems from the Piagetian theory of learning. While its earliest formulations comprised a set of 18 different codes (Berkowitz & Gibbs, 1979), applications within the CSCL community, aiming to achieve success with automation, have targeted much simpler operationalizations defined by the presence of two requirements. First, the speaker must demonstrate a reasoning attempt. Second, the speaker must reference ideas or concepts presented earlier in the conversation. Prior datasets for transactivity typically identify the presence or absence of these requirements in a binary fashion (Joshi & Rosé, 2007; Wen et al., 2016). This approach has been successful in extant work which has focused on assignments that made use of very short collaborative discussions (Fiacco & Rosé, 2018), informal posts in discussion forums (Nelimarkka & Vihavainen, 2015), or team-based project support (Wen et al., 2016). However, the complexity of the language articulated within these previous works was limited, with some studies finding transactive exchange in only 60% of posts in a discussion forum (Sankaranarayanan et al., 2018). By contrast, in the current dataset, which includes masters level students in social science courses, about 80% of posts were rated as transactive by human coders using the simple definition, resulting in a lack of useful differentiation. Therefore, we turned our attention to adding more nuance to our operationalization of transactivity for automated analysis in order to better differentiate students and conversations.

Here we present a new dataset for transactivity detection based on a more detailed conceptual framework and measure. We then answer the following research questions targeted at automatic transactivity detection with respect to this new operationalization: First, can previous state-of-the-art models of transactivity detection apply to the domain of current event discussion forums in social science courses, and what phenomena exists in that domain that distinguish it from transactivity datasets? And, second, how can we capture these differences in a model that can better detect transactivity on this new dataset?

We show that, despite being highly functional on simpler datasets, the existing state-of-the-art model fails on our new dataset owing to the higher degree of abstractness in the conversations analyzed. We then present a new model based on this dataset that leverages the structure of the source data to more accurately predict more

nuanced transactivity phenomena. The work illustrates how variations in student and course content result in different expressions of transactivity and that successful models must reflect those differences.

Transactivity coding

We collected communication data from 198 students in a master's degree program from a university in the Northeastern U.S. As part of a leadership course assignment, instructors provided students with a weekly article related to some of the topics learned from class for seven weeks. Students were instructed to post their thoughts on an online discussion board and also provide a response to at least three other classmates' posts. We extracted the response data, nested in each student's post thread, from the course platform in a json file format for each discussion topic. Across six different discussion topics, 3,415 replies in total were collected.

Building on prior work on transactivity coding (Berkowitz & Gibbs, 1979; Gweon, Jain, McDonough, Raj, & Rosé, 2013), we operationalized and coded the transactivity of the responses. Overall, the responses were found to be more elaborative than previous work that coded transactivity using a binary approach (i.e., transactive or non-transactive) and automatically annotated transactivity (Joshi & Rosé, 2007; Wen et al., 2016). But more importantly, while in prior studies the prevalence of transactivity was fairly low, in this population nearly 80% of contributions satisfied the simple definition of transactivity. The software infrastructure and nature of the task did not differ significantly from prior studies. Thus, we concluded that the substantial shift in conversational practices was due to the different student population and assignment, namely masters students in a social science course discussing current events.

To develop an appropriate framework, we proceeded to regroup the transacts of Berkowitz and Gibbs (1979) according to their roles (functions) in collaborative learning, develop a new coding scheme, and measure the level of transactivity or the extent to which an individual expended effort to represent and operate on their partners' reasoning. In their original framework, Berkowitz and Gibbs (1979) identified 18 types of transacts or dialogue behaviors, which are classified as higher or lower order transacts. Higher-order forms are operational transacts (e.g., counter consideration) that work on partners' reasoning through logical analysis and integration. Lower-order forms include representational (e.g., juxtaposition) and elicitational transacts (e.g., feedback request), which do not entail any transformations of partners' reasoning. Moreover, the transacts feature either competitive (e.g., competitive paraphrase) or non-competitive (e.g., paraphrase) forms, which can be focused on either partner and the dyad's positions.

Building from this framework, we developed new transacts and grouped them into three dimensions (functions): *active listening* (acknowledgment), *idea extension* (elaboration), and *challenging views* (qualification). First, we focused on *active listening* as it is conducive to creating an environment of mutual respect (Itzchakov, Kluger, & Castro, 2017) and psychological safety where the partners feel their contributions are valued and respected (Azmitia & Montgomery, 1993). In examining *active listening*, we coded whether the responders put in the effort to acknowledge their partners' ideas and thoughts by paraphrasing and/or asking them for further explanation. Second, more learning happens when discussions are disequilibrating, where individuals are exposed to something new from the interactions and experience cognitive perturbations (Berkowitz & Oster, 1985; De Lisi & Golbeck, 1999). Accordingly, *idea extension* evaluates the extent to which the individuals were elaborative in presenting their own reasoning processes in relation to their partners' original ideas or asked thought-provoking questions about their counterpart's contribution. Third, considering that cognitive perturbations could be more salient when there are conflicting views, thereby increasing the likelihood of transactive exchanges, we evaluated *challenging views* to assess the strength and clarity of the partners' challenging of their counterparts' argument.

In coding the data, each dimension was rated independently, although multiple dimensions might apply to a particular response. Specifically, for *active listening*, a binary rating was used, while for *idea extension* and *challenging views*, a 3-point scale (0: Not at all, 1: A little, 2: A lot) was used. In short, each statement was evaluated for the focal individuals' perceived exertion of effort to make sense of the meaning of their partners' argument (*active listening*) and build on their partner's reasoning (*idea extension, challenging views*).

Extended operationalization of new transactivity dimensions

Active listening

Active listening involves the focal participant's (Ego) acknowledgment of their partners' (Alter) contribution as is, in a non-judgmental manner. The transacts for *active listening* include "paraphrasing" and "soliciting clarification." That is, evaluators code whether Ego made the effort to paraphrase Alter's message and asked for further explanation to better understand Alter's point of view. Importantly, in soliciting clarification, Ego is not asking Alter to justify their reasoning or explore the ideas Ego proposed. The main criteria is: "Did Ego attempt

to identify Alter's ideas and thoughts?" *Active listening* is coded "Yes (1)" when there is paraphrasing and/or soliciting clarification transacts; otherwise, it is coded "No (0)".

Coding examples for *active listening*

- Example of "Yes (1)"
  Alter: "… their job is to return value to the shareholders… That being said I don't think that only extroverted, or introverted people can do this. It just changes the way the company is set up and the culture that inherently stems from the leadership."
  Ego: "I agree with you that the main focus of hiring new leaders should be whether they can return value to the shareholders, but as you say introverted and extroverted leaders will set up different cultures. While these two cultures may be able to return the same value…"

Explanation: Ego paraphrased Alter's ideas in a clear way.

- Example "No (0)"
  Alter: "… Hiring new talent is an excellent way to gain access to these skills, but this should be in addition to retraining current staff, not in lieu of training. Some companies are able to fully hire a new staff, but many won't be able to do this…"
  Ego: "I think you make great points but there is one to add…"

Explanation: Here Ego did not demonstrate the way Ego understood Alter's argument.

Idea extension

In evaluating *idea extension,* coders annotated the extent to which Ego elaborated Alter's ideas by (1) exploring parallel lines of thought (i.e., agreement-based idea extension) and/or (2) qualifying Alter's argument (i.e., disagreement-based *idea extension*). Notably, Ego may demonstrate both forms of extension. First, for agreement-based extension, the transacts include "exploration," "exploration request," and "application." That is, Ego can provide additional evidence and thoughts either declaratively (exploration) or interrogatively (exploration request), and apply Alter's ideas to different contexts (application), such that Alter's argument becomes clearer and more generalizable. Second, for disagreement-based extension, the transacts include "critique" and "counter-argument." Ego can critically evaluate Alter's reasoning in a declarative or interrogative way (critique) and present opposing arguments (counter-argument)—uncovering the assumptions and exploring alternatives—such that Alter's argument becomes more robust and competitive. In evaluating this dimension, coders ask: "To what extent did Ego demonstrate their reasoning process?" Specifically, agreement-based *idea extension* is coded "A lot (2)" when Ego illustrated their argument with examples, demonstrated logical thinking, and/or integrated multiple ideas. Moreover, disagreement-based extension is coded "A lot (2)" when Ego explicated why Alter's argument may not be supported and/or provided clear evidence to support their counter-argument.

Coding examples for *idea extension*

- Example of "A little (1)"
  Alter: "…that extroversion became a cultural ideal and if extroversion is indeed the perceived ideal, maybe we have CEOs who learned how to be extroverted on the job because that is what is expected of them…"
  Ego: "You make a very interesting point. CEO extroversion could be a result of society's perception that it is crucial or more important than the other aspects and traits you mention. I agree that these are equally if not more important to hiring decisions. It would be interesting to see how this cultural ideal varies across countries/societies."

Explanation: Ego provided an additional thought, which needs to be developed.

- Example of "A lot (2)"
  Alter: "I don't believe that standardized tests should be used for college admissions, hiring, or anyplace else. Different people may have different skill sets that standardized tests don't take into account. Moreover, people may not have the same opportunity to be as prepared as they can for these tests."
  Ego: "The problem with eliminating standardized testing to remove bias is that there isn't a less biased criteria to replace it with. Ultimately, the bias shown on standardized testing is the result of general disadvantages that impact all parts of the student's application. In fact, when you consider recommendation letters…, essays…, and extracurricular activities that low income students simply can't afford, standardized testing is actually one of the less biased parts of the application…We also should do what we can to reduce the inequalities that cause all of these problems."

Explanation: Ego raised an alternative perspective and provided supporting pieces of evidence.

Challenging views

Challenging views gauges whether Ego was clear and extensive in stating their opposing position to Alter. Coders focused on the choice of words/phrases and the sentence structure to evaluate the clarity and strength of the challenge. Notably, coders do not evaluate if Ego's argument is relevant and well-reasoned, which is the focus of *idea extension*. The main criteria here is: "To what extent did Ego qualify Alter's argument"?

Coding examples of *challenging views*

- Example of "A little (1)"
  Alter: "I believe that firms should include retraining initiatives as they transform their businesses… Retraining is a difficult journey, but it is one that can be mutually beneficial for companies and their employees."
  Ego: "I agree that re-training employees will typically be worthwhile. But, should re-training be available to all employees?…"

Explanation: Ego agreed with Alter's view in general; Ego qualified one aspect of the argument.

- Example of "A lot (2)"
  Alter: "… I would have to assume that the team would be next to impossible to rectify in due time to complete the deadline for WS1, and I would respectfully decline the position…"
  Ego: "You have only described opportunities for James. The bar has been set low by the poor performance of the group which has been operating without a strong leader. James can be the new spark that keep everyone on track..."

Explanation: Ego qualifies Alter's argument directly, explaining how it can be interpreted in a different way.

Evaluation of interrater reliability

Two independent raters coded three dimensions of transactivity for a sample of 180 responses. To be comprehensive, interrater reliability was assessed using three measures, including intraclass correlations (ICC), Krippendorff's alpha, and weighted Cohen's kappa. For ICC, ICC(2, *k*) or a two-way random effects model was used (McGraw & Wong, 1996; Shrout & Fleiss, 1979). For ordinal variables, ICC is recommended (Hallgren, 2012). ICC is also suitable for nominal and continuous variables. Krippendorff's alpha (Hayes & Krippendorf, 2007) was also computed as a measure for assessing inter-rater reliability for all types of variables. Moreover, whereas Cohen's (1960) kappa is only suitable for nominal or categorical variables, weighted Cohen's (1968) kappa allows estimating the reliability for ordinal variables.

The results demonstrated excellent absolute-agreement ICC values for all dimensions: *active listening* (.89), *idea extension* (.87), and *challenging views* (.91). Krippendorff's alpha was found to be acceptable across dimensions: *active listening* (.80), *idea extension* (.76), and *challenging views* (.80). Last, Cohen's kappa showed moderate to strong levels of agreement: *active listening* (.80), *idea extension* (.69), and *challenging views* (.77). Given these values, we were confident in moving forward with our plan to have only one of the two raters code the responses that are required to train the machine for automatic detection of transactivity. A sample of 910 comments, consisting of a similar number of comments for each discussion topic, were coded to be used for deep learning, as discussed in the following section.

Automated transactivity detection experiments

Our goal was to find a model that most accurately predicts the various facets of transactivity that we have defined in our dataset. To this end, we started with an implementation of the previous state-of-the-art in transactivity detection to evaluate its ability to detect our more nuanced operationalization of transactivity in our data. We analyzed the data to identify reasons for the discrepancy in performance of the baseline model on each dataset. Our findings informed a new detector for transactivity to address the shortcomings of the baseline model. Below we provide an evaluation of the new transactivity detector.

Results for each experiment for transactivity detection were obtained via a 10-fold cross-validation where each fold was randomly assigned but consistent throughout the different conditions.

Baseline: Transferable attention model for transactivity detection

The model, called the Transferable Attention Model by Fiacco & Rosé (2018) is a variant of the Decomposable Attention Model for Recognizing Textual Entailment by Parikh et al. (2016), where the model is pre-trained on the RTE task after which the final layers are re-randomized and the model is allowed to fine-tune on the small transactivity dataset. Full implementation details of the model are discussed in Fiacco & Rosé (2018).

While the entailment task takes in a premise and a hypothesis statement to train the model with the hypothesis statement being the statement to be determined if the entailment relation holds, in the transactivity task, the premise is replaced by the context and the hypothesis is replaced by the message. The message is the text that is to be labeled as transactive and the context is the text for which the message is responding to.

For experiments on our dataset, the message was the post that is to be determined to show one of the aspects of transactivity while the context is the post to which that message responded. Note that the message and context may not be temporally adjacent as determination for message response was made via the forum response tree and participants can respond directly to prior posts.

Comparisons of transactivity data with respect to transferable attention model

The first research question we sought to address stems from a comparison between the data used by Fiacco & Rosé (2018) to train the Transferable Attention Model and our new dataset from class discussions. We noted that previous datasets used far more concrete language, while we observed more abstract language in our new dataset. Concreteness of language is characterized by referring to specific objects, people, or actions while abstractness is defined as language referring to concepts and ideas.

Table 1: Abstractness for datasets relevant to transactivity detection; scale 0 (concrete) to 1 (abstract)

Dataset	Text Abstractness
SNLI (Bowman et al., 2015)	0.334
MultiNLI (Williams et al., 2018)	0.530
Powerplant Transactivity Corpus	0.538
Masters Student Corpus	0.583

In Table 1, we present the abstractness of each dataset based on the average abstractness of inputs using the methodology from Brysbaert et al. (2014). We evaluate the transferable attention model using an alternative entailment pre-training dataset, the Multi-genre Natural Language Inference corpus (MultiNLI; Williams et al., 2018) which we found to be considerably more abstract than the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) which was the pre-training corpus for the original Transferable Attention Model. This pretraining corpus was hypothesized to improve the model's performance by better representing the more abstract text found in the masters student data.

Transformer model for transactivity detection

One of the key shortcomings of the Transferable Attention Model is its inability to take into account word order. This is especially relevant to the *challenging views* dimension as negation is common within examples of that dimension and the meaning of a negation is highly word order dependent. To address this, we propose to use a class of models from the Natural Language Processing literature called transformers (Vaswani et al., 2017). The benefit of this type of model is that it combines the capability for self-attention with sequential reasoning to build a numerical representation of a sequence of text that can be used to classify that sequence.

Specifically, we use the pretrained transformer model, RoBERTa (Liu et al., 2019) which incorporates some optimizations of the BERT transformer model (Devlin et al., 2019). This model, like the GloVe embeddings used in the Transferable Attention Model, was pretrained on an enormous amount of general text data and will be fine-tuned on both the entailment pretraining task and the transactivity task. The model was then fine-tuned on the Recognizing Textual Entailment task similarly to the Transferable Attention Model. This fine-tuned model was the based model for each of the cross-validation folds. For each fold, the model was further fine-tuned on the transactivity data with a separate classification head as the entailment classifier.

Evaluation

We evaluated the potential to automate analysis using the extended transactivity definition proposed here, beginning with the best published approach from Fiacco & Rosé (2018), and comparing its approach to three other variants. From Table 2, it is evident that pretraining the Transferable Attention Model on the MultiNLI dataset

had a large positive effect ($p < 0.05$) on all of the dimensions of transactivity. The increase was most notable for *active listening* while there were only modest improvements for *idea extension*, and *challenging views*.

Table 2: Cohen's kappa scores of transactivity detection models on 10 fold cross-validation

Model	Active listening	Idea extension	Challenging views
Transferable Attention Model (Fiacco & Rosé, 2018)	0.239	0.399	0.316
Transferable Attention Model+MultiNLI	0.314	0.429	0.340
Transferable Attention Model+MultiNLI+Self Attn.	**0.715**	0.656	0.461
RoBERTa+MultiNLI	0.651	**0.660**	**0.668**

Even more dramatic is the increase in performance from the redefinition of inputs for the Transferable Attention Model to make the model perform self-attention rather than attending between context and content. Furthermore, the RoBERTa model is able to significantly improve upon the performance on the *challenging views* dimension. However, it did not significantly improve on *idea extension* and underperformed on *active listening.* All differences between rows in Table 2 are significant ($p < 0.05$).

Discussion and conclusions

In this study, we uncovered some important considerations that must be taken into account when modeling approaches are used for automated detection of constructs such as transactivity. The line of experimentation reported here was prompted by an observation that a previously published demonstration of domain generality could not be generalized to a substantially different student population with distinctive discourse practices. Our findings point to necessary adjustments, first at the level of operationalization of the construct and then at the level of modeling approach -- with synergistic considerations between the two -- in order to achieve success.

In particular, our findings reveal a larger increase in performance for the *active listenin*g dimension between the baseline Transferable Attention Model and the version that used the MultiNLI pretraining as compared to the *idea extension* or *challenging views* dimensions. We attribute this largely to the vocabulary of the NLI datasets as compared to the masters student data. The masters student data is far more verbose and abstract than the SNLI dataset as compared to MultiNLI dataset. *Active listening* is a relatively simple task as compared to *challenging views* or *idea extension* as it is frequently signaled by agreement or disagreement. As the SNLI dataset is based off of image descriptions, there is little opportunity for that kind of language to occur. The MultiNLI corpus pulls data from a far broader range of genres and may expose the model to more relevant sentence forms. For *idea extension* or *challenging views*, the limiting factor was not as much the vocabulary, but how the model was able to use the data it had.

There was a large jump in performance across all dimensions of transactivity by redefining the Transferable Attention Model as a self-attention model as opposed to attending between the content and its context. While in data with less abstract contributions, the important factor for detecting transactivity may be ensuring that there are aligned phrases between the content and the context, in our masters student dataset, it appears to be more important for the model to understand what the responder is contributing. This result aligns with our qualitative observations that the masters students had deeper contributions and more structured responses as compared to the contributions in prior datasets. Detecting transactivity, in this case, is more about evaluating how the response is formed, regardless of the context.

This insight is reinforced by the performance of the RoBERTa based model that also uses sequential information to preserve the word order and sentence structure within the embedding. For c*hallenging views*, word order is critical to understand the content of a contribution as challenging one's view often involves negation. Negation can drastically change the meaning of a text segment depending on where it occurs. Adding the capability to do word order allowed the RoBERTa model to perform comparably between *idea extension* and *challenging views* while the Transferable Attention Model demonstrated a large gap between the two.

However, an interesting result was that the RoBERTa-based model performed worse than the Transferable Attention Model on *active listening*. A possible explanation of this comes from a qualitative analysis of the data where many of the *active listening* examples (for both the positive and negative cases) had a consistent structure where a student would express agreement or disagreement and then give an example. For the cases that reflected *active listening*, the example used specific language referring to content in the previous post (e.g. "I agree that re-training employees would be worthwhile.") For the cases that did not show *active listening*, the examples tended to use generalization or non-specific language (e.g. "I agree with what you said.") This difference can be modeled very well by simple self-attention; the model only needs to determine if the words attended to are

generic or specific. Adding considerably more information via the RoBERTa model may make the distinction less clear.

Finally, the work reported here, to investigate the transfer of a successful automated analysis approach for transactivity from one context to another, is important for the community if resources are to be used efficiently through sharing. We began by recounting some history in application of the construct of transactivity to research in CSCL and the reasons why automation is valuable to the community. We then presented the contrasting case of masters students in social science to illustrate how population differences may be associated with substantial differences in discourse practices which may render earlier definitions unable to differentiate between students. A more nuanced operationalization and corresponding automation approach was therefore needed, which we have presented along with an evaluation in this paper. In future work, it would be valuable to explore how population differences impact desiderata for operationalization of other constructs related to collaboration process; it could be fruitful to identify how differences in population characteristics such as personality, age, academic/professional field or discussion context necessitate changes in the analysis approach. These further point to the need and potential value for a more coordinated effort across the CSCL community to provide sharable resources for automatic collaborative process analysis.

Acknowledgments

This research was funded in part by NSF grants ACI-1443068 and IIS 1546393 and funding from the Schmidt Foundation.

References

Ai, H., Sionti, M., Wang, Y.C., & Rosé, C. (2010). Finding transactive contributions in whole group classroom discussions. In Proceedings of the 9th International Conference of the Learning Sciences-Volume 1 (pp. 976–983).

Ankur P. Parikh and Oscar Täckström and Dipanjan Das and Jakob Uszkoreit (2016). A Decomposable Attention Model for Natural Language InferenceCoRR, abs/1606.01933.

Azmitia, M., & Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. Social Development, 2, 202-221.

Berkowitz, M. W., & Gibbs, J. C. (1979). A preliminary manual for coding transactive features of dyadic discussion. Unpublished manuscript, Marquette University, Milwaukee.

Berkowitz, M. W., & Oser, F. (1985). Moral education: Theory and application. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bowman, S., Angeli, G., Potts, C., & Manning, C. D. (2015, September). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 632-642).

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods, 46(3), 904-911.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Dascalu, M., Trausan-Matu, S., McNamara, D., & Dessus, P. (2015). ReaderBench: Automated evaluation of collaboration based on cohesion and dialogismInternational Journal of Computer-Supported Collaborative Learning, 10(4), 395–423.

De Lisi, R., & Golbeck, S. (1999). Implications of Piagetian theory for peer learning. In O'Donnell, A. M., & King, A. (Eds.), Cognitive perspectives on peer learning (pp. 213–312). New York, NY: Routledge.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Ferschke, O., Yang, D., Tomar, G., & Rosé, C. (2015). Positive impact of collaborative chat participation in an edX MOOC. In International Conference on Artificial Intelligence in Education (pp. 115–124).

Fiacco, J., & Rosé, C. (2018, June). Towards domain general detection of transactive knowledge building behavior. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-11).

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

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8, 23-34.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77-89.

Itzchakov, G., Kluger, A. N., & Castro, D. R. (2017). I am aware of my inconsistencies but can tolerate them: The effect of high quality listening on speakers' attitude ambivalence. Personality and Social Psychology Bulletin, 43, 105-120.

Joshi, M., & Rosé, C. P. (2007). Using transactivity in conversation summarization in educational dialog. Proceedings of the SLaTE Workshop on Speech and Language Technology in Education

Kumar, R., Rosé, C., Wang, Y.C., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning supportFrontiers in artificial intelligence and applications, 158, 383.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30-46.

McLaren, B., Scheuer, O., De Laat, M., Hever, R., De Groot, R., & Rosé, C. (2007). Using machine learning techniques to analyze and support mediation of student e-discussionsFrontiers in Artificial Intelligence and Applications, 158, 331.

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 discussionsInternational journal of computer-supported collaborative learning, 7(2), 285–305.

Nelimarkka, M., & Vihavainen, A. (2015). Alumni & tenured participants in MOOCs: Analysis of two years of MOOC discussion channel activity. In Proceedings of the Second (2015) ACM Conference on Learning@Scale (pp. 85–93).

Rosé, C., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learningInternational journal of computer-supported collaborative learning, 3(3), 237–271.

Sankaranarayanan, S., Dashti, C., Bogart, C., Wang, X., Sakr, M., Rosé, C. (2018). When Optimal Team Formation is a Choice - Self-Selection versus Intelligent Team Formation Strategies in a Large Online Project-Based Course, *Proceedings of AI in Education 2018*

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Turney, P., Neuman, Y., Assaf, D., & Cohen, Y. (2011, July). Literal and metaphorical sense identification through concrete and abstract context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 680-690).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wen, M., Maki, K., Wang, X., Dow, S. P., Herbsleb, J., & Rosé, C. P. (2016). Transactivity as a Predictor of Future Collaborative Knowledge Integration in Team-Based Learning in Online Courses. In Barnes, T., Chi, M., & Feng, M. (Eds.), Proceedings of the 9th International Conference on Educational Data Mining (pp. 533–538).

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1112-1122).