# Chapter 11: Modeling Educational Discourse with Natural Language Processing

Nia Dowell,[1] Vitomir Kovanović[2]

[1] School of Education, University of California, Irvine, United States
[2] Education Futures, University of South Australia, Adelaide, Australia

DOI: 10.18608/hla22.011

## ABSTRACT

The broadening adoption of technology enhanced learning environments has substantially altered the manner in which educational communication takes place, with most people engaging in some form of online asynchronous or synchronous conversation every day. The language and discourse artifacts emerging from these technological environments is a rich source of information into learning processes and outcomes. This chapter describes the current landscape of natural language processing (NLP) tools and approaches available to researchers and practitioners to computationally discern patterns in large quantities of text-based conversations that take place across a variety of educational technology platforms. The capabilities of NLP are particularly important as, in the field of learning analytics, we desire to effectively and efficiently learn about the process of learning by observing learners, and then subsequently use that information to improve learning. We conclude the chapter with a discussion around the emerging applications (i.e., sensing technologies, breakthroughs in AI, and cloud computing) and challenges of NLP tools to educational discourse.

**Keywords:** Natural Language Processing (NLP), computational linguistics, discourse analysis

The rapid growth of social media, online communities and learning platforms has dramatically changed the manner in which communication takes place. Conversation technologies are omnipresent in today's organizational environment, from email, text messaging, and wikis to more sophisticated knowledge management systems; all of which are leveraged to support social, business, and educational functions. Educational environments in particular have become increasingly reliant on computer-mediated communication, relying on video conferencing, synchronous chats, and asynchronous forums, in both small (with 5–20 learners) and massive (with hundreds or even thousands of learners) environments. These platforms, which are designed to support or even supplant traditional instruction, have become commonplace across all levels of education, and as a result created big data in education [64, 82].

The language and discourse artifacts emerging from these environments is a rich source of information into learning processes. It is important to clarify what we mean by discourse. Our definition of discourse includes both oral and chat-based communication between two or more individuals (e.g., peer-peer, peer-teacher communicative interactions). Indeed, the importance of communication for the learning process has been a consistent narrative in the learning sciences and learning analytics research [112]. The fundamental role of language is represented in the scope of chapters devoted to various language and discourse processes such as social network analysis (cf. chapter X), reading (cf. chapter X), writing (cf. chapter X), a general overview of analysis approaches (cf. chapter X), and multi-party interaction (i.e., peer-peer interactions, peer-agent, or peer-teacher), which is the focus of the current chapter. As evident in these chapters, language provides a powerful and measurable behavioral signal that can be used to capture the semantic, structural and sociocognitive interaction patterns that characterize learning related phenomenon including cognitive, metacognitive, motivational, social and affective dimensions of student engagement [7, 62].

Conventional approaches to quantifying and characterizing language and discourse characteristics have traditionally required human examination (i.e., manual content analysis) [71], which is known to carry biases and other methodological limitations [72]. In particular, the laborious nature of these tasks make them no longer a viable option with the increasing scale of online interaction data (Graesser et al., 2018) [84, 108, 126]. Advances in artificial intelligence methods, such as Natural language Processing (NLP) [63], have made it possible to automatically i) harness vast amounts of communication data being produced in technology-mediated learning environments, ii) quantify aspects of human cognition, affective and social processes in text-based human-to-human and human-to-agent conversations that iii) would otherwise not be possible for human coders to capture, given the multifaceted discourse characteristics of human interaction.

## 1 ANALYZING CONVERSATIONAL INTERACTIVE DISCOURSE USING NATURAL LANGUAGE PROCESSING

While discourse analysis can involve the analysis of different kinds of data (e.g., video, audio, text), the most widely used techniques for discourse analysis focus on the analysis of written, textual information. Within the field of Text Mining (Aggarwal & Zhai, 2012) and Natural Language Processing [63] there have been many techniques developed which can be used for the analysis of discourse data. In this section we will examine the ways in which techniques from these two fields have been used to analyze educational discourse.

The simplest forms of discourse analysis involve *bag-of-words approaches* [2] and calculation of the N-gram frequencies, which are sequences of consecutive N-words (i.e., unigrams: one word, bigrams: two words, trigrams: three words). Extracted N-gram frequencies are then used as *input features* for the development of various analytical models, such as discourse classification or clustering systems. For instance, Kovanović et al. [68] used unigram, bigram, and trigram counts as features for the classification of discussion messages according to the level of cognitive presence [43], a theoretical construct that captures the development of students' critical thinking. Similar approaches have been used, for example, for detecting student reflection [118], student's knowledge states [80], detection of relevant/irrelevant questions [13], classification of dialogue acts [38], and collaborative problem-solving [108]. In all of these cases, extracted bag-of-words N-gram features were used to represent discourse for the purpose of analytical model development.

While bag-of-words representations (i.e., frequencies of the extracted N-grams) depend on the content of the input data, *dictionary-based approaches* utilize a predefined list of words (or phrases), and represent the input data through frequencies of the different word groups. One of the most widely used dictionary-based tools is LIWC (Linguistic Inquiry and Word Count) [97, 116], which calculates the frequencies of words from over 100 word categories. An important benefit of such approaches is that those categories are empirically validated and representative of important psychological processes, making them easier to interpret and use for research purposes. Within the context of educational research, LIWC has been used, for instance, to assess students' cognitive load [65], predicting student performance and engagement in MOOCs [105, 124, 131] and traditional face-to-face courses [106], cognitive presence detection [61, 92], reflection [45, 69, 76, 78, 119], and social interactions [4, 35, 128].

In addition to simple, word-based representations, there is a whole range of techniques for representing discourse using the different linguistic properties of the input text [84]. Such techniques range from the simple counts of the number of words, sentences or paragraphs to more complex measures of different linguistic properties. In this regard, one of the widely used tools is Coh-Metrix [50, 85], which provides over 200 different linguistic metrics of the input text. In addition to providing simple word, sentence and paragraph counts, Coh-Metrix also provides a wide range of linguistic and coherence indices, including text readability, lexical diversity, use of connective words, syntactic complexity and pattern density, part-of-speech category use, and semantic overlap of input sentences/paragraphs. Coh-metrix has been used in a wide range of studies of educational discourse (see Dowell, Graesser, and Cai [29] for an overview).

Another class of NLP technique for representing discourse focuses on understanding the semantic structure of the input text. Such techniques focus on capturing the meaning of the textual data, and use that semantic information to model the discourse. These techniques typically involve extracting a specific number of hidden, or *latent*, topics in a large collection of textual documents and associating these topics to each of the documents in the collection. The input for such algorithms is the *document-term matrix* (DTM), which is a matrix where rows represent documents, columns represent all words (used across all documents), and values word frequencies in the documents.

One of the earliest and most widely-used semantic analysis techniques is Latent Semantic Analysis (LSA) [74] which is a technique for decomposing DTM into a product of two smaller matrices (document-topic and topic-word matrices) using singular value decomposition (SVD), a simple linear algebra transformation algorithm. Thus, each document represents a combination of latent topics, and each latent topic is characterized through word frequency distribution. LSA has been widely used in education [75], for a wide range of problems from automated essay grading [42], team communication [24], and use of online discussions [14]. LSA is also utilized by Coh-Metrix to calculate the semantic overlap between the sentences and paragraph as a means of assessing cohesiveness of the written text [50].

While LSA has been widely used for semantic analysis of educational discourse, the recent development in statistical machine learning brought several new techniques that often produce results superior to those by LSA. Those include *probabilistic topic modeling algorithms* [8, 115], which derive document-topic and topic-word associations through the use of generative models and Markov-Chain Monte Carlo (MCMC) simulations [47]. The most notable algorithm in this domain is Latent Dirichlet Allocation (LDA) [9], which enables realistic modeling of uneven topic distribution across documents (as often the case in practice). LDA has been widely used in humanities [17] and social sciences [101] including education. Within learning analytics field, LDA and topic modeling have been primarily used for modeling students' online communication [14, 15, 41, 53, 107, 121], and student writings [46, 111], but also for the analysis of student course enrollment data [91].

Recent advancements in artificial neural networks (ANNs) and deep learning resulted in the development of some highly effective techniques for discourse representation. The most notable tool in this area is *Word2vec* [88], which utilises a two-layer shallow neural network to produce *word embeddings*, a vector-based representation of the text which preserves its semantics. Using word2vec, a semantic similarity of two texts can be easily calculated through calculation of the cosine similarity between their respective vectors. Word2vec has been used in learning analytics for a wide range of tasks, including grade predicting through the analysis of student lecture comments [81], short responses [83], and student misconceptions [87].

In addition to the development of more complex and sophisticated discourse representations, there has also been significant focus on capturing and modeling the inherent complexity and temporal dynamics of the learners conversations, such as those that take place in online collaborative learning, problem-solving, and online course forums [18, 52, 55, 104]. In particular, the sociocognitive aspects of learner's interactions reside in and evolve through the semantic connection between individual's utterances over time. As such, researchers have started to use innovative temporally sensitive NLP approaches to assess the socio-cognitive properties of online interactions.

The most representative approach of temporally sensitive NLP tools is *Group Communication Analysis* (GCA) [32], a computational approach for the analysis of multi-party discourse from computer-mediated peer to peer, team, and collaborative group interactions. In contrast with existing computational approaches to text analysis, GCA emphasizes emergent aspects of learner discourse interactions [70]. Temporal emergence of the discourse is integral to the methods behind GCA that capture temporal alignment, sequential ordering and coordination in meaning during human communication [26, 32](Hu et al., 2018).

To this end, GCA combines artificial intelligence methods, such as computational semantic models of cohesion, with temporally sensitive semantic analyses inspired by the cross- and auto-correlation measures from time-series analysis. These semantic space models, which rely on advanced artificial intelligence techniques, may be constructed via Latent Semantic Analysis (LSA) [75], a classic matrix-factorization method, or more current artificial neural network word embedding models such as Skip-gram (i.e. Word2vec, [89]) or Global Vectors of Words (i.e., "GloVe", [98]). Using this approach, GCA allows researchers to quantify discourse as a dynamic and evolving sociocognitive process that resides in the interaction between learner's communicative contributions.

## 2 CURRENT STATE OF DISCOURSE ANALYTICS

### 2.1 Small Scale Multi-Party Interactions

One of the most common NLP applications in the context of small-scale multi-party interactions involves examining the word level properties of student's communication. For instance, researchers have used the features from LIWC to explore sentiment [108], transformative discourse [127], and self and socially-shared regulation during collaboration [132]. Similarly, Latent Dirichlet Allocation has proved successful in transforming the topics of texts into values as a basis for representing cognitive information graphically [37]. Grammatical information can also provide valuable insights as shown by Sullivan and Keith's (2019) research [114], which highlights how parts of speech (POS) analysis can be used to uncover student sense-making activity during collaborative learning. Quantifying the occurrence of words in general and across different psychological categories provides information about the precise content of students' communication.

Other tools move beyond the explicit meaning and allow researchers to quantify more latent characteristics of student discourse interactions, such as Coh-Metrix [50, 85], TAALES [73], TAACO [19], and ReaderBench [22]. These systems provide a summative account of learner discourse at the student level (i.e. individual posts or totality of them per person) as well as at the group level (i.e. text of the overall thread transcript) along various text properties, such as cohesion (e.g., [28]), and narrativity [102]. These "bag of words" and more summative NLP methods offer several advantages regarding their simplicity and ability to provide specific information about the content of student discourse during computer-mediated collaborations, such as word level, syntactic, and cohesion properties of texts.

Collaborative interactions are fundamentally defined as a process that occurs over time [103], and characterized by the dynamic, emergent, adaptive, and interdependent nature of joint human communicative actions to produce meaning. However, the above NLP approaches traditionally ignored this character, choosing instead to examine relationships between relatively static input and outcome variables [126]. Temporally sensitive NLP approaches offer significant promise for the conceptualization of the ways in which collaboration unfolds over time and the inherent complexity [49, 51, 103, 104], which could substantially advance our understanding of multi-party collaborative interactions. In this context, Järvelä et al. [58] traced the occurrence of self-regulated learning (SRL) and socially shared regulated learning (SSRL) in the context of CSCL. They used temporal and sequential analysis of chat discussions and log file traces to find evidence of whether the students collaboratively planned regulatory activities were shared in practice. In practice, Järvelä et al. [58] matched each individual's SRL from the log file traces and his or her SSRL from the chat data and composed micro-level examples to demonstrate the interplay between self-regulation and socially shared regulation of learning. The main finding was that collaborating groups engaging in SSRL achieved better learning outcomes when compared with groups that did not.

More recently, GCA has been used to quantify the temporal properties of learners' socio-cognitive processes and communication dynamics in online multi-party interactions. This approach has provided substantial insights on the emergent sociocognitive roles learners occupy during collaborative interactions [30, 32, 34, 33], and deeper understanding of inclusivity and equality in online team interactions [26, 31, 6, 79]. For instance, Dowell and colleagues have uncovered differences in learners' interpersonal interaction patterns across ethnic populations, between male and female students [6], and the influence of gender group composition on equitable interpersonal discourse during STEM interactions [31]. Across these studies, GCA has revealed substantial intra- and interpersonal differences in women and URM's engagement, which could influence their sense of belonging in online STEM environments.

## 2.2 Scaling of Discourse Analytics

Advances in educational technologies and a desire for increased access to learning, have enabled the development of pedagogical environments at scale, such as Massive Open Online Courses (MOOCs) [62, 120]. Open online courses have the potential to advance education on a global level, by providing the masses with broader access to lifelong learning opportunities. Early research on the MOOC phenomena saw significant investment in understanding the makeup of the learner population, largely through demographic [36], performance, and activity-based measures [66]. The discourse artifacts emerging from these environments were primarily investigated from the network perspective, with Social Network Analysis (SNA; see chapter X for an overview) being a primary means of extrapolating meaning from this data. However, there has been an uptick in the application of NLP tools to understand temporal population trends (e.g., [27]), profiles [34], and various learning phenomena within MOOCs (e.g., engagement, [62]).

Some notable applications include the use of NLP to quantify aspects of learner-generated posts, as well as learners' cognitive, affective, and social processes. Identifying aspects and categories of students posts as enormous value given the scale of student discourse within MOOCs, and the associated teacher effort required. Wise et al. [125] work has focused on bringing order to the chaos in MOOC discussion forums. Their work used a bag of words approach (i.e., unigram and bigram) to classify students' posts into content vs. non-content related posts. Others have used similar approaches in conjunction with tools like LIWC and machine learning models to identify urgent posts that require more immediate teacher attention [3].

A major theme in the literature is the use of NLP for the assessment of learners' psychological processes in MOOCs and broader technology-mediated learning contexts. Interesting applications around affective detection hold significant potential given the important role of emotions in learning (see Graesser [48], Pekrun [96], and Perry and Souza [99] for a review). Sentiment analysis can be used as a first step for identifying complex emotions, such as excitement, frustration or confusion. Sentiment analysis is the process of identifying and classifying learners' opinions from a piece of text into different sentiments-for example, positive, negative, or neutral—or emotions such as happy, sad, angry, or disgusted to determine the user's attitude toward a particular topic or within a context. This can give an insight into how learners feel with the course to be able to perform modifications aimed at increasing learners' engagement and satisfaction, which is very important to ensure the success of the MOOC [90, 100].

Several researchers have highlighted the application of sentiment analysis in the context of scaled learner interactions (e.g., [1, 16, 129]). Some of the earlier work by Wen, Yang, and Rose [123], applied sentiment analysis techniques on student posts on three MOOCs. They observed a negative correlation between the ratio of positive to negative terms and dropout across time. In detecting different confusion states Yang et al. [130] relied on psychologically meaningful categories of words, extracted from online discussions using the LIWC as one of the classification features for retention. Their work highlighted that confusion reduced the likelihood of retention, but this could be reduced with confusion resolution and other supportive interventions. Others have explored student sentiment in scaled environments in relation to performance and student perceptions. For instance, Tucker, Pursel, and Divinsky [117], using word-sentiment lexicon, found that students' affective discourse was negatively related to their average grade. However, this relationship was modest and positively related to their quiz grades. Similar to Yang, Adamopoulos [1] employed AlchemyAPI to extract student sentiment from discussion forum messages and found student sentiment toward course instructors, assignments, and course materials have a positive effect on the course retention.

An emerging trend in research highlights the novel insights that can be gleamed through a combination of complementary analytic techniques, such as SNA and various NLP analytics. The research in this context used systems like Coh-Metrix and LIWC or analytical approaches such as GCA [32] and Epistemic Network Analysis (ENA; [109]) in conjunction with SNA to gain a more holistic understanding of learners discourse [20, 35, 44, 59, 60]. For instance, Coh-Metrix has been involved in pioneering research exploring the potential methodological and theoretical advantages of combining SNA and computational linguistic analyses [35, 59]. Joksimović and colleagues used Coh-Metrix to analyze learners' forum posts in a distributed (Twitter, blogs and Facebook) MOOC. Social Network Analysis was used to determine students' social centrality. Linear mixed-effect modeling was used to reveal the linguistic profiles associated with more centrality located learners. Overall, the results indicated that learners in the MOOC connected easier to individuals who use a more informal, narrative style, but still maintain a deeper cohesive structure to their communication. However, this linguistic profile cannot be immediately interpreted as beneficial for learning. Dowell et al. [35] used a similar methodological design, but also included a measure of student performance in the MOOC. Specifically, students who performed significantly better engaged in more expository style discourse, with surface and deep level cohesive integration, abstract language, and simple syntactic structures. However, linguistic profiles of the centrally positioned learners differed from the high performers. Learners with a more significant and central position in their social network engage using a more narrative style discourse with less overlap between words and ideas, simpler syntactic structures and abstract words [35, 59].

## 3 CHALLENGES AND FUTURE ADVANCES

Here we have provided a landscape view of computational methods available for researchers to understand and quantify learning related phenomena during computer-mediated communication, and situated these within the context of both small- and large-scale learner interactions. As illustrated by these applied examples, computational linguistic methods are now in full swing within the learning analytics and broader educational community [84]. Thus, as nicely articulated by Wise and Schwarz [126] the substantive question is not if we should embrace computational approaches to understanding multi-party interactions, but how to develop practices and norms around their use that maintain the community's commitment to theory and situational context. Looking forward, we propose it is unlikely that these computational advances and applications will slow, but instead, we are already seeing evidence of future innovations that will have very real implications for both researchers and practitioners, and the relationship between these groups. Below we outline a few of these emerging trends and associated challenges.

Educational discourse research includes both written and auditory discourse analysis, though we focused exclusively on computational methods for computer-mediated text interactions, however, the approaches taken to understanding learner interactions between these registers can differ significantly. Turn taking cues differ heavily between the modalities and shape social aspects of the environment such as power dynamics and inclusion. It is not uncommon for auditory discourse analysis to include these elements, usually through painstaking annotation of text and video transcripts captured from the educational setting. New sensor technologies promise to increase both the recording prevalence and the automation of analysis of technology-mediated speech discourse. Some researchers have already taken a step in this direction by using spoken language to computationally model complex collaboration processes (e.g., construction of shared knowledge, negotiation/coordination, and maintaining team function) [5, 113], effective communication [57], agreeableness [77] and speaker's influence [93]. For instance, Hung and Gatica-Perez connected team cohesion to the audio-visual features within task-oriented groups [54]. This is driven in part by the development of low-cost software and consumer appliances aimed at more natural human computer interaction. For instance, IBM Watson Speech to Text service [56] can aid researchers by generating a transcript from video based multi-party interactions with start and stop times for each utterance spoken by each learner. Similarly, the Amazon Echo Dot, designed for home automation tasks, is a small and inexpensive device which contains an array of seven directional microphones and can capture speaker direction, record audio, and respond to queries based on speech recognition. Depth sensing cameras, popularized by the Microsoft Kinect device but now available from various vendors, form three dimensional maps of a learner based on their physical appearance and have the capability to do facial recognition, detect gaze direction, and detect facial expression.

The implication of such inexpensive yet highly capable sensing technologies rests primarily in the significant opportunity for researchers who study in-person or video based discourse interactions [21, 25]. In addition to potentially lowering manual coding costs and effort (i.e., human annotation of text and video transcripts captured from the educational setting) when categorizing educational discourse processes, the low cost and small size of such devices makes it conceivable that future educational spaces might be built with data analytics in mind [94]. For instance, one could imagine even very large classrooms being outfitted with such technologies which might enable the analysis of (and thus interventions for) active learning approaches. Regardless of whether such equipment becomes ubiquitous in educational spaces or used for research studies alone, it provides an opportunity for educational researchers to rethink data capture and analysis methods, with an eye towards how one might distill large volumes of fine grained data into constructs of interest [10].

Modern computational processing power has created revolutionary advances in NLP. A major player in the field was revealed by Google and is a breakthrough artificial intelligence technology called BERT (Bidirectional Encoder Representations from Transformers; [23], which has garnered significant attention in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering, Natural Language Inference, and others. BERT's main technical innovation is applying the bidirectional training of Transformer, a widely used attention model, to language modeling. This is in contrast to previous efforts which examine a text sequence either from left to right or combined for training. Devlin et al. [23] highlight how a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. However, this revolutionary AI appears to have a significant issue, as articulated by the NY Times " It could be picking up on biases in the way a child mimics the bad behavior of his parents" [86]. That is, BERT, like many other similar NLP approaches, learns linguistic representations from tons of digitized information, such as old books, Wikipedia entries and news articles. This has created non-trivial issues as these societal artifacts carry decades of biases as well as the current biases within our society [12]. An illustration of the problematic behavior are the recurrently appearing occupational stereotypes that the word *'homemaker'* is related to the word *'woman'* as the word *'programmer'* is to the word *'man'* [11, 122].

Recent studies have aimed to detect, analyze and mitigate gender bias in different NLP tools and applications including word embeddings, but these issues remain and should be carefully thought about when implementing any NLP techniques.

Nonetheless, the advances in the computing domain open up several opportunities for researchers aiming at improving education [64, p.127]. For instance, pervasive sensing and data analytics offer the ability to do real-time capture, inference, and intervention. While the vast majority of current educational discourse analysis is done in a post-hoc fashion, there is a growing trend towards real-time software analytics augmentation [67]. For instance, most learning management systems (LMSs) now have clickstream-style logging of learner interactions which is available instantly to researchers. This native functionality is being integrated by data specification bodies groups such as IMS Global who are now actively engaged in reflecting real-time data interoperability needs in educational data standards. This work has the potential to increase dramatically the number and variety of educational technologies that provide data about learner interactions with systems (including discourse interactions) in an *in-situ* fashion. Those includes the provision of feedback to both students and instructors [67] as well as integration with other real-time analytics systems such as social network analysis [39], epistemic network analysis [44, 107] or Group Communication Analysis [32]. Educational discourse analysis also poses some potentially high challenges for researchers with regard to ethics and privacy preservation [95, 110]. While a discussion of these important issues is beyond the scope of the current work, there have been efforts towards the development of different solutions and frameworks for privacy protection in learning analytics [40].

Educational discourse analysis is a broad research area, and takes place in primary, secondary, higher and emerging education environments. In this Chapter, we have provided an overview of the developing field of educational NLP analysis, and a map of emerging opportunities and challenges educational researchers face with socio-technical advances. As we have outlined, sociotechnical advances have already influenced the scale of discourse data and computational methods used by educational researchers. For instance, the increase in blended, MOOC, and informal educational environments has changed the scale of discourse data, wherein researchers now regularly utilize automated linguistic analysis and machine learning approaches to handle the increasing amount of discourse data produced within these educational environments. As these sociotechnical changes continue, we hope this discussion draws attention not only to future research opportunities immediately available in the field, but also the necessary technical, computational, sociological, and linguistic developments needed to handle the changing nature of discourse, the computational infrastructure resources needed for real-time analysis of educational discourse, and the relationships between educational researchers, institutional educational technologies, and third party vendors, which are imperative to enable next-generation educational NLP scholarly work.

## REFERENCES


[1] Panagiotis Adamopoulos. "What makes a great MOOC? An interdisciplinary analysis of student retention in online courses". In: ICIS 2013 Proceedings. aisel.aisnet.org, 2013. URL: https://aisel.aisnet.org/icis2013/proceedings/BreakthroughIdeas/13/.

[2] Charu C Aggarwal and Chengxiang Zhai. "An Introduction to Text Mining". In: *Mining Text Data*. Ed. by Charu C Aggarwal and Chengxiang Zhai. Springer US, Jan. 2012, pp. 1–10. URL: http://link.springer.com.proxy.lib.sfu.ca/chapter/10.1007/978-1-4614-3223-4_1.

[3] Omaima Almatrafi, Aditya Johri, and Huzefa Rangwala. "Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums". In: *Comput. Educ.* 118 (Mar. 1, 2018), pp. 1–9. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2017.11.002. URL: http://www.sciencedirect.com/science/article/pii/S0360131517302373.

[4] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. "Talk to me: Foundations for successful individual-group interactions in online communities". In: *CHI '06*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Journal Abbreviation: CHI '06. New York, NY, USA: ACM, 2006, pp. 959–968. DOI: 10.1145/1124772.1124916. URL: http://doi.acm.org/10.1145/1124772.1124916 (visited on 05/11/2013).

[5] Kathleen T. Ashenfelter. *Simultaneous analysis of verbal and nonverbal data during conversation: symmetry and turn-taking*. University of Notre Dame, 2007. URL: https://www.academia.edu/download/53965124/AshenfelterK042007D.pdf.

[6] Yiwen author=Lin, Nia M. Dowell, Andrew Godfrey, Heeryung Choi, and Christopher Brooks. "Modeling gender differences in intra- and interpersonal dynamics during online learning collaborative interactions". In: Proceedings of the 9th International Conference for Learning Analytics & Knowledge. Ed. by D Azcona and R Chung. Temple, AZ: ACM, 2019.

[7] Roger Azevedo. "Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues". In: *Educ. Psychol.* 50.1 (Jan. 2, 2015). Publisher: Routledge, pp. 84–94. ISSN: 0046-1520. DOI: 10.1080/00461520.2015.1004069. URL: https://doi.org/10.1080/00461520.2015.1004069.

[8] David M. Blei. "Probabilistic topic models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: http://doi.acm.org/10.1145/2133806.2133826.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation (LDA)". In: *Journal of Machine Learning Research* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[10] Paulo Blikstein. "Multimodal learning analytics". In: Proceedings of the third international conference on learning analytics and knowledge. 2013, pp. 102–106.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, pp. 4356–4364. URL: https://dl.acm.org/doi/10.5555/3157382.3157584 (visited on 03/04/2020).

[12] Ewa S. Callahan and Susan C. Herring. "Cultural bias in Wikipedia content on famous persons". In: *Journal of the American Society for Information Science and Technology* 62.10 (Oct. 7, 2011), pp. 1899–1915. ISSN: 0002-8231. DOI: 10.1002/asi.21577. URL: http://doi.wiley.com/10.1002/asi.21577.

[13] Suleyman Cetintas, Luo Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez. "Microblogging in a classroom: Classifying students' relevant and irrelevant questions in a microblogging-supported classroom". In: *IEEE Transactions in Learning Technology* 4.4 (2011), pp. 292–300. ISSN: 1939-1382. DOI: 10.1109/TLT.2011.14. URL: http://dx.doi.org/10.1109/TLT.2011.14.

[14] Bodong Chen. "Visualizing semantic space of online discourse: The knowledge forum case". In: *LAK '14*. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. Journal Abbreviation: LAK '14. New York, NY, USA: ACM, 2014, pp. 271–272. DOI: 10.1145/2567574.2567595. URL: http://doi.acm.org/10.1145/2567574.2567595.

[15] Ming Ming Chiu and Nobuko Fujita. "Statistical discourse analysis: A method for modeling online discussion processes". In: *Journal of Learning Analytics* 1.3 (Nov. 8, 2014), pp. 61–83. ISSN: 1929-7750. URL: http://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/4074.

[16] Ruth Cobos, Francisco Jurado, and Álvaro Villén. "Moods in MOOCs: Analyzing emotions in the content of online courses with edX-CAS". In: 2019 IEEE Global Engineering Education Conference (EDUCON). Apr. 2019, pp. 1467–1474. DOI: 10.1109/EDUCON.2019.8725107. URL: http://dx.doi.org/10.1109/EDUCON.2019.8725107.

[17] Daniel J. Cohen and Joan Fragaszy Troyano. "Special issue on topic modeling in digital humanities". In: *Journal of Digital Humanities* 2.1 (2012).

[18] Ulrike Cress, Gerry Stahl, Sten Ludvigsen, and Nancy Law. "The core features of CSCL: Social situation, collaborative knowledge processes and their design". In: *International Journal of Computer-Supported Collaborative Learning* 10.2 (June 1, 2015), pp. 109–116. DOI: 10.1007/s11412-015-9214-2. URL: https://doi.org/10.1007/s11412-015-9214-2.

[19] Scott A. Crossley, Kristopher Kyle, and Danielle McNamara. "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion". In: *Behav. Res. Methods* 48.4 (Dec. 1, 2016), pp. 1227–1237. ISSN: 1554-351X. DOI: 10.3758/s13428-015-0651-7. URL: https://doi.org/10.3758/s13428-015-0651-7.

[20] Scott A. Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. "Combining click-stream data with NLP tools to better understand MOOC completion". In: *LAK '16*. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. Journal Abbreviation: LAK '16. New York, NY, USA: ACM, 2016, pp. 6–14. DOI: 10.1145/2883851.2883931. URL: http://doi.acm.org/10.1145/2883851.2883931.

[21] Sidney K. D'Mello, Nigel Bosch, and Huili Chen. "Multimodal-multisensor affect detection". In: *Processing, Architectures, and Detection …* (2018). Publisher: dl.acm.org. URL: https://dl.acm.org/doi/abs/10.1145/3107990.3107998.

[22] Mihai Dascalu, Stefan Trausan-Matu, Danielle McNamara, and Philippe Dessus. "ReaderBench: Automated evaluation of collaboration based on cohesion and dialogism". In: *Intern. J. Comput.-Support. Collab. Learn* 10.4 (Dec. 1, 2015), pp. 395–423. ISSN: 1556-1607. DOI: 10.1007/s11412-015-9226-y. URL: http://link.springer.com.proxy.lib.umich.edu/article/10.1007/s11412-015-9226-y (visited on 12/08/2016).

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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). 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.pdf.

[24] Andy Dong. "The latent semantic approach to studying design team communication". In: *Design Studies* 26.5 (Sept. 2005), pp. 445–461. ISSN: 0142-694X. DOI: 10.1016/j.destud.2004.10.003. URL: http://www.sciencedirect.com/science/article/pii/S0142694X05000050 (visited on 11/11/2015).

[25] Patrick J. Donnelly, Nathaniel Blanchard, Andrew M. Olney, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. "Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context". In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. dl.acm.org, 2017, pp. 218–227. URL: https://dl.acm.org/doi/abs/10.1145/3027385.3027417.

[26] Nia M. Dowell. "Preparing for the future: Group communication analysis as a tool to facilitate adaptive support during digitally-mediated team interactions". In: In Proceedings of the 3rd International Conference on Artificial Intelligence + Adaptive Education. 2019.

[27] Nia M. Dowell, Christopher Brooks, Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević. "The changing patterns of MOOC discourse". In: *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. Ed. by Claudia Urrea, Justin Reich, and Candice Thille. Cambridge, M.A.: ACM, 2017, pp. 283–286.

[28] Nia M. Dowell, Whitney L. Cade, Yla R. Tausczik, James W. Pennebaker, and Arthur C. Graesser. "What works: Creating adaptive and intelligent systems for collaborative learning support". In: *Twelfth International Conference on Intelligent Tutoring Systems.* Ed. by Stefan Trausan-Matu, Kristy E. Boyer, Martha Crosby, and Kitty Panourgia. Berlin: Springer, 2014, pp. 124–133.

[29] Nia M. Dowell, Arthur C. Graesser, and Zhiqiang Cai. "Language and discourse analysis with cohmetrix: Applications from educational material to learning environments at scale". In: *Journal of Learning Analytics* 3.3 (Dec. 2016), pp. 72–95. ISSN: 1929-7750. DOI: 10.18608/jla.2016.33.5. URL: https://learning-analytics.info/journals/index.php/JLA/article/view/4330.

[30] Nia M. Dowell, Yiwen Lin, Andrew Godfrey, and Christopher Brooks. "Exploring the relationship between emergent sociocognitive roles, collaborative problem- solving skills and outcomes: A Group Communication Analysis". In: *Journal of Learning Analytics* 7.1 (2020), pp. 38–57.

[31] Nia M. Dowell, Yiwen Lin, Andrew Godfrey, and Christopher Brooks. "Promoting inclusivity through time-dynamic discourse analysis in digitally-mediated collaborative learning". In: Proceedings of the 20th International Conference on Artificial Intelligence in Education. Ed. by Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin. Chicago, IL: ACM, 2019.

[32] Nia M. Dowell, Tristan Nixon, and Arthur C. Graesser. "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multi-party interactions". In: *Behav. Res. Methods* (2018). ISSN: 1554-351X. DOI: 10.3758/s13428-018-1102-z. URL: http://dx.doi.org/10.3758/s13428-018-1102-z.

[33] Nia M. Dowell and Oleksandra Poquet. "SCIP: Identifying learner roles through group communication and interpersonal network positioning in scaled digital environments". In: *Journal of Computers in Human Behavior* ().

[34] Nia M. Dowell, Oleksandra Poquet, and Christopher Brooks. "Applying group communication analysis to educational discourse interactions at scale". In: *Proceedings of the 13th International Conference on the Learning Sciences: Full Crossover Paper*. Ed. by Judy Kay and Rose Luckin. Vol. 2. London, England: International Society of the Learning Sciences, 2018, pp. 1815–1822.

[35] Nia M. Dowell, Oleksandra Skrypnyk, Srećko Joksimović, Arthur C. Graesser, Shane Dawson, Dragan Gašević, Pieter de Vries, Thieme Hennis, and Vitomir Kovanović. "Modeling learners' social centrality and performance through language and discourse". In: *Proceedings of the 8th International Conference on Educational Data Mining*. Ed. by Cristóbal Romero and M Pechenizkiy. International Educational Data Mining Society., 2015, pp. 250–257.

[36] Ezekiel J. Emanuel. "Online education: MOOCs taken by educated few". In: *Nature* 503.7476 (2013), p. 342. ISSN: 0028-0836.

[37] Melanie Erkens, Daniel Bodemer, and H. Ulrich Hoppe. "Improving collaborative learning in the classroom: Text mining based grouping and representing". In: *International Journal of Computer-Supported Collaborative Learning* 11.4 (Dec. 1, 2016), pp. 387–415. ISSN: 1556-1615. DOI: 10.1007/s11412-016-9243-5. URL: https://doi.org/10.1007/s11412-016-9243-5.

[38] Aysu Ezen-Can, Joseph F Grafsgaard, James C. Lester, and Kristy Elizabeth Boyer. "Classifying student dialogue acts with multimodal learning analytics". In: *LAK '15*. Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '15. New York, NY, USA: ACM, 2015, pp. 280–289. DOI: 10.1145/2723576.2723588. URL: http://

doi.acm.org/10.1145/2723576.2723588 (visited on 10/28/2016).

[39] Rebecca Ferguson and Simon Buckingham Shum. "Social learning analytics: Five approaches". In: *LAK '12*. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '12. New York, NY, USA: ACM, 2012, pp. 23–33. DOI: 10.1145/2330601.2330616. URL: http://doi.acm.org.proxy.lib.sfu.ca/10.1145/2330601.2330616.

[40] Rebecca Ferguson, Tore Hoel, Maren Scheffel, and Hendrik Drachsler. "Ethics and privacy in learning analytics [Special issue]". In: *Journal of Learning Analytics* 3.1 (Apr. 23, 2016), pp. 5–15. ISSN: 1929-7750.

[41] Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, and Vitor Rolim. "Towards combined network and text analytics of student discourse in online discussions". In: *Lecture Notes in Computer Science*. Artificial Intelligence in Education. Ed. by Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay. Journal Abbreviation: Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2018, pp. 111–126. DOI: 10.1007/978-3-319-93843-1_9. URL: http://dx.doi.org/10.1007/978-3-319-93843-1_9.

[42] Peter W. Foltz, Darrell Laham, Thomas K. Landauer, Peter W. Foltz, Darrell Laham, and Thomas K. Landauer. "Automated essay scoring: Applications to educational technology". In: vol. 1999. 1999, pp. 939–944. (Visited on 09/03/2015).

[43] D. Randy Garrison, Terry Anderson, and Walter Archer. "Critical thinking, cognitive presence, and computer conferencing in distance education". In: *American Journal of Distance Education* 15.1 (Jan. 1, 2001). Publisher: Routledge, pp. 7–23. ISSN: 0892-3647. DOI: 10.1080/08923640109527071. URL: https://doi.org/10.1080/08923640109527071.

[44] Dragan Gašević, Srećko Joksimović, Brendan R Eagan, and David Williamson Shaffer. "SENS: Network analytics to combine social and cognitive perspectives of collaborative learning". In: *Comput. Human Behav.* 92 (Mar. 1, 2019), pp. 562–577. ISSN: 0747-5632. DOI: 10.1016/j.chb.2018.07.003. URL: http://www.sciencedirect.com/science/article/pii/S074756321830325X.

[45] Dragan Gašević, Negin Mirriahi, and Shane Dawson. "Analytics of the effects of video use and instruction to support reflective learning". In: *LAK '14*. Proceedings of the Fourth International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '14. New York, NY, USA: ACM, 2014, pp. 123–132. DOI: 10.1145/2567574.2567590. URL: http://doi.acm.org/10.1145/2567574.2567590.

[46] Andrew Gibson and Kirsty Kitto. "Analysing reflective text for learning analytics: An approach using anomaly recontextualisation". In: *LAK '15*. Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '15. New York, NY, USA: ACM, 2015, pp. 275–279. DOI: 10.1145/2723576.2723635. URL: http://doi.acm.org/10.1145/2723576.2723635.

[47] Walter R. Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995. URL: https://www.taylorfrancis.com/books/9780429170232.

[48] Arthur C. Graesser. "Emotions are the experiential glue of learning environments in the 21st century". In: *Learning and Instruction* (July 30, 2019), p. 101212. ISSN: 0959-4752. DOI: 10.1016/j.learninstruc.2019.05.009. URL: http://www.sciencedirect.com/science/article/pii/S0959475219303640.

[49] Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. "Advancing the science of collaborative problem solving". In: *Psychological Science in the Public Interest* 19.2 (Nov. 2018). Publisher: journals.sagepub.com, pp. 59–92. ISSN: 1529-1006. DOI: 10.1177/1529100618808244. URL: http://dx.doi.org/10.1177/1529100618808244.

[50] Arthur C. Graesser, Danielle McNamara, Max M. Louwerse, and Zhiqiang Cai. "Coh-metrix: Analysis of text on cohesion and language". In: *Behavior Research Methods, Instruments, & Computers* 36.2 (2004), pp. 193–202. ISSN: 0743-3808. URL: http://www.ncbi.nlm.nih.gov/pubmed/15354684.

[51] Paul Hager and David Beckett. *The Emergence of Complexity: Rethinking Education as a Social Science*. Jan. 1, 2019. ISBN: 978-3-030-31837-6. DOI: 10.1007/978-3-030-31839-0. URL: https://www.researchgate.net/publication/336456931_The_Emergence_of_Complexity_Rethinking_Education_as_a_Social_Science (visited on 03/05/2020).

[52] Peter Holtz, Joachim Kimmerle, and Ulrike Cress. "Using big data techniques for measuring productive friction in mass collaboration online environments". In: *International Journal of Computer-Supported Collaborative Learning* 13.4 (Dec. 1, 2018), pp. 439–456. ISSN: 1556-1615. DOI: 10.1007/s11412-018-9285-y. URL: https://doi.org/10.1007/s11412-018-9285-y.

[53] I-Han Hsiao and Piyush Awasthi. "Topic facet modeling: Semantic visual analytics for online discussion forums". In: *LAK '15*. Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '15. New York, NY, USA: ACM, 2015, pp. 231–235. DOI: 10.1145/2723576.2723613. URL: http://doi.acm.org/10.1145/2723576.2723613.

[54] Hayley Hung and Daniel Gatica-Perez. "Estimating cohesion in small groups using audio-visual nonverbal behavior". In: *IEEE Trans. Multimedia* 12.6 (Oct. 2010). Publisher: ieeexplore.ieee.org, pp. 563–575. ISSN: 1941-0077. DOI: 10.1109/TMM.2010.2055233. URL: http://dx.doi.org/10.1109/TMM.2010.2055233.

[55] Laszlo Hunyadi and István Szekrényes. *The Temporal Structure of Multimodal Communication: Theory, Methods and Applications*. Springer, July 24, 2019. 162 pp. ISBN: 978-3-030-22895-8. URL: https://play.google.com/store/books/details?id=jAilDwAAQBAJ.

[56] *IBM*. 2019. URL: https://www.ibm.com/cloud/watson-speech-to-text.

[57] Hirofumi Inaguma, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. "Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue". In: Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction. dl.acm.org, 2016, pp. 11–15. URL: https://dl.acm.org/doi/abs/10.1145/3005467.3005472.

[58] Sanna Järvelä, Paul A Kirschner, Ernesto Panadero, Jonna Malmberg, Chris Phielix, Jos Jaspers, Marika Koivuniemi, and Hanna Järvenoja. "Enhancing socially shared regulation in collaborative learning groups: Designing for CSCL regulation tools". In: *Educational Technology Research and Development* 63.1 (2015), pp. 125–142.

[59] Srećko Joksimović, Nia Dowell, Oleksandra Skrypnyk, Vitomir Kovanović, Dragan Gašević, Shane Dawson, and Arthur C. Graesser. "How do you connect? Analysis of social capital accumulation in connectivist MOOCs". In: *Proceedings from the 5th International Learning Analytics and Knowledge (LAK) Conference*. New York, NY: ACM, 2015, pp. 66–68.

[60] Srećko Joksimović, Nia M. Dowell, Oleksandra Poquet, Vitomir Kovanović, Dragan Gašević, Shane Dawson, and Arthur C Graesser. "Exploring development of social capital in a CMOOC through language and discourse". In: *The Internet and Higher Education* 36 (Supplement C Jan. 1, 2018), pp. 54–64. ISSN: 1096-7516. DOI: 10.1016/j.iheduc.2017.09.004. URL: http://www.sciencedirect.com/science/article/pii/S1096751617304554.

[61] Srećko Joksimović, Dragan Gašević, Vitomir Kovanović, Olusola Adesope, and Marek Hatala. "Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions". In: *The Internet and Higher Education* 22 (2014), pp. 1–10. DOI: 10.1016/j.iheduc.2014.03.001. URL: http://dx.doi.org/10.1016/j.iheduc.2014.03.001.

[62] Srećko Joksimović, Oleksandra Poquet, Vitomir Kovanović, Nia Dowell, Caitlin Mills, Dragan Gašević, Shane Dawson, Arthur C. Graesser, and Christopher Brooks. "How do we model learning at scale? A systematic review of research on MOOCs". In: *Rev. Educ. Res.* (2018). ISSN: 0034-6543. DOI: 10.3102/0034654317740335. URL: https://doi.org/10.3102/0034654317740335 (visited on 12/17/2017).

[63] Anne Kao and Stephen R Poteet. *Natural language processing and text mining*. London: Springer, 2007. ISBN: 978-1-84628-175-4. URL: http://www.amazon.com/Natural-Language-Processing-Text-Mining/dp/184628175X.

[64] Ben Kei Daniel. *Big Data and Learning Analytics in Higher Education Current Theory and Practice*. Cham: Springer International Publishing : Imprint: Springer, 2017. ISBN: 978-3-319-06520-5. DOI: 10.1007/978-3-319-06520-5. URL: http://dx.doi.org/10.1007/978-3-319-06520-5.

[65] M. Asif Khawaja, Fang Chen, Christine Owen, and Gregory Hickey. "Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design". In: *Human-Computer Interaction – INTERACT 2009*. Ed. by Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler. Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 1, 2009, pp. 485–489. ISBN: 978-3-642-03654-5. URL: http://link.springer.com.proxy.lib.sfu.ca/chapter/10.1007/978-3-642-03655-2_54.

[66] René F. Kizilcec, Chris Piech, and Emily Schneider. "Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses". In: *LAK '13*. Proceedings of the Third International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '13. New York, NY, USA: ACM, 2013, pp. 170–179.

[67] Simon Knight and Karen Littleton. "Discourse centric learning analytics: Mapping the terrain". In: *Journal of Learning Analytics* 2.1 (Feb. 18, 2015), pp. 185–209. ISSN: 1929-7750. URL: http://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/4043.

[68] Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, and Marek Hatala. "Automated cognitive presence detection in online discussion transcripts". In: Workshop Proceedings of LAK 2014. Indianapolis, IN, USA, 2014. URL: http://ceur-ws.org/Vol-1137/.

[69] Vitomir Kovanović, Srećko Joksimović, Negin Mirriahi, Ellen Blaine, Dragan Gašević, George Siemens, and Shane Dawson. "Understand students' self-reflections through learning analytics". In: *LAK '18*. Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18). Journal Abbreviation: LAK '18. New York, NY, USA: ACM, 2018, pp. 389–398. DOI: 10.1145/3170358.3170374. URL: http://dx.doi.org/10.1145/3170358.3170374.

[70] Steve W. J. Kozlowski. "Enhancing the effectiveness of work groups and teams: A reflection". In: *Perspect. Psychol. Sci.* 13.2 (Mar. 2018), pp. 205–212. ISSN: 1745-6916. DOI: 10.1177/1745691617697078. URL: http://dx.doi.org/10.1177/1745691617697078.

[71] Klaus H. Krippendorff. *Content nalysis: An Introduction to Its Methodology*. Sage Publications, Dec. 2003.

[72] Klaus H. Krippendorff. "Reliability in content analysis". In: *Hum. Commun. Res.* 30.3 (2004), pp. 411–433. ISSN: 0360-3989. DOI: 10.1111/j.1468-2958.2004.tb00738.x. URL: http://dx.doi.org/10.1111/j.1468-2958.2004.tb00738.x.

[73] Kristopher Kyle and Scott A. Crossley. "Automatically assessing lexical sophistication: indices, tools, findings, and application". In: *TESOL Q* 49.4 (Dec. 25, 2015), pp. 757–786. ISSN: 0039-8322. DOI: 10.1002/tesq.194. URL: http://doi.wiley.com/10.1002/tesq.194.

[74] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis". In: *Discourse Process.* 25.2 (1998), pp. 259–284. ISSN: 0163-853X. DOI: 10.1080/01638539809545028. URL: http://www.tandfonline.com/doi/abs/10.1080/01638539809545028.

[75] Thomas K. Landauer, Danielle McNamara, Simon Dennis, and Walter Kintsch. *Handbook of latent semantic analysis*. New York: Lawrence Erlbaum Assoc Incorporated, 2007. 545 pp. ISBN: 978-0-8058-5418-3. URL: http://books.google.com/books?id=JbzCzPvzpmQC.

[76] Gilly Leshed, Jeffrey T. Hancock, Dan Cosley, Poppy L. McLeod, and Geri Gay. "Feedback for guiding reflection on teamwork practices". In: *GROUP '07*. Proceedings of the 2007 international ACM conference on Supporting group work. Journal Abbreviation: GROUP '07. New York, NY, USA: ACM, 2007, pp. 217–220. DOI: 10.1145/1316624.1316655. URL: http://doi.acm.org.proxy.lib.sfu.ca/10.1145/1316624.1316655.

[77] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. "Acoustic-prosodic entrainment and social behavior". In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies. dl.acm.org, 2012, pp. 11–19. URL: https://dl.acm.org/citation.cfm?id=2382032.

[78] Chi-Wei Lin, Meei-Ju Lin, Chin-Chen Wen, and Shao-Yin Chu. "A word-count approach to analyze linguistic patterns in the reflective writings of medical students". In: *Med. Educ. Online* 21 (2016), p. 29522. ISSN: 1087-2981. DOI: 10.3402/meo.v21.29522. URL: http://dx.doi.org/10.3402/meo.v21.29522.

[79] Yiwen Lin and Nia M. Dowell. "Does gender really matter?: Exploring differences in emerging discourse styles during digitally-mediated collaborative interactions". 2019.

[80] Mihai Lintean, Vasile Rus, and Roger Azevedo. "Automatic detection of student mental models based on natural language student input during metacognitive skill training". In: *International Journal of Artificial Intelligence in Education* 21.3 (Aug. 2011), pp. 169–190. ISSN: 1560-4292. DOI: 10.3233/JAI-2012-022. URL: http://dx.doi.org/10.3233/JAI-2012-022.

[81] Jingyi Luo, E. Shaymaa Sorour, Kazumasa Goda, and Tsunenori Mine. "Predicting student grade based on free-style comments using Word2Vec and ANN by considering prediction results obtained in consecutive lessons". In: Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015). 2015.

[82] Collin F. Lynch. "Who prophets from big data in education? New insights and new challenges". In: *Educ. Res. Eval.* 15.3 (Nov. 1, 2017). Publisher: SAGE Publications, pp. 249–271. ISSN: 1380-3611. DOI: 10.1177/1477878517738448. URL: https://doi.org/10.1177/1477878517738448.

[83] Jenny McDonald and Adon Christian Michael Moskal. "Quantext: Analysing student responses to short-answer questions". In: *Me, Us, IT* (2017), pp. 133–137. URL: http://2017conference.ascilite.org/wp-content/uploads/2017/11/Concise-MCDONALD.pdf.

[84] Danielle McNamara, Laura K. Allen, Scott A. Crossley, Mihai Dascalu, and Cecile A Perret. "Natural language processing and learning analytics". In: *Handbook of Learning Analytics*. Ed. by C. Lang, G. Siemens, A. F. Wise, and D. Gaevic. First Edition. Alberta, Canada: Society for Learning Analytics Research, 2017, pp. 93–104. ISBN: 978-0-9952408-0-3. URL: https://www.solaresearch.org/hla-17/hla17-chapter8/.

[85] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, M.A.: Cambridge University Press., 2014.

[86] Cade Metz. "We teach A.I. systems everything, including our biases". In: *The New York Times* (Nov. 11, 2019). ISSN: 0362-4331. URL: https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html.

[87] Joshua J. Michalenko, Andrew S. Lan, and Richard G. Baraniuk. "Data-mining textual responses to uncover misconception patterns". In: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale. 2017, pp. 245–248. URL: https://dl.acm.org/doi/abs/10.1145/3051457.3053996?casa_token=RfdA6yhHfgsAAAAA:R2hQ0irRKzJxze4_49ZnB8ahxGLPtw45Ck5zt_il2T7uacWhQJgq3kOjiCJDdfjEkgzRDYBykjdt.

[88] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". ISBN: 1301.3781 Publication Title: arXiv [cs.CL]. Jan. 16, 2013. URL: http://arxiv.org/abs/1301.3781.

[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed representations of words and phrases and their compositionality". In: *NIPS'13*. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Journal Abbreviation: NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[90] Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J Muñoz-Merino, Iria Estévez-Ayres, and Carlos Delgado Kloos. "Sentiment analysis in MOOCs: A case study". In: 2018 IEEE Global Engineering Education Conference (EDUCON). Apr. 2018, pp. 1489–1496. DOI: 10.1109/EDUCON.2018.8363409. URL: http://dx.doi.org/10.1109/EDUCON.2018.8363409.

[91] Benjamin Motz, Thomas Busey, Martin Rickert, and David Landy. "Finding topics in enrollment data". In: Proceedings of the 11th International Conference on Educational Data Mining. Buffalo, NY: International Educational Data Mining Society, 2018, pp. 424–430.

[92] Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins. "Automated analysis of cognitive presence in online discussions written in Portuguese". In: *Lecture Notes in Computer Science*. Lifelong Technology-Enhanced Learning. Ed. by Viktoria Pammer-Schindler, Mar Pérez-Sanagustín, Henrik Drachsler, Raymond Elferink, and Maren Scheffel. Journal Abbreviation: Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2018, pp. 245–261. DOI: 10.1007/978-3-319-98572-5_19. URL: http://dx.doi.org/10.1007/978-3-319-98572-5_19.

[93] Fumio Nihei, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. "Predicting influential statements in group discussions using speech and head motion information". In: Proceedings of the 16th International Conference on Multimodal Interaction. dl.acm.org, 2014, pp. 136–143. URL: https://dl.acm.org/doi/abs/10.1145/2663204.2663248.

[94] Amy Ogan. "Reframing classroom sensing: promise and peril". In: *Interactions* 26.6 (2019). Publisher: ACM New York, NY, USA, pp. 26–32. ISSN: 1072-5520. URL: https://dl.acm.org/doi/fullHtml/10.1145/3358902.

[95] Abelardo Pardo and George Siemens. "Ethical and privacy principles for learning analytics". In: *Br. J. Educ. Technol.* 45.3 (May 1, 2014), pp. 438–450. ISSN: 0007-1013. DOI: 10.1111/bjet.12152. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12152.

[96] Reinhard Pekrun. "Inquiry on emotions in higher education: progress and open problems". In: *Studies in Higher Education* 44.10 (Oct. 3, 2019). Publisher: Routledge, pp. 1806–1811. ISSN: 0307-5079. DOI: 10.1080/03075079.2019.1665335. URL: https://doi.org/10.1080/03075079.2019.1665335.

[97] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin, Sept. 16, 2015. URL: https://repositories.lib.utexas.edu/handle/2152/31333.

[98] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[99] Gabriela Perry and Napoliana Souza. "Identification of affective states in MOOCs: A systematic literature review". In: *1* 6.12 (Dec. 31, 2018), pp. 39–55. ISSN: 2411-2933. DOI: 10.31686/ijier.Vol6.Iss12.1250. URL: https://ijier.net/index.php/ijier/article/view/1250.

[100] Trang Phan, Sara G. McNeil, and Bernard R. Robin. "Students' patterns of engagement and course performance in a Massive Open Online Course". In: *Comput. Educ.* 95 (Apr. 1, 2016), pp. 36–44. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2015.11.015. URL: http://www.sciencedirect.com/science/article/pii/S0360131515300877.

[101] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. "Topic modeling for the social sciences". In: NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond. Whistler, Canada, Dec. 2009.

[102] Joseph M. Reilly and Bertrand Schneider. "Predicting the quality of collaborative problem solving through linguistic analysis of discourse". In: Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019). ERIC, 2019, pp. 149–157. URL: https://files.eric.ed.gov/fulltext/ED599226.pdf.

[103] Peter Reimann. "Time is precious: Variable- and event-centred approaches to process analysis in CSCL research". In: *Computer Supported Learning* 4.3 (June 27, 2009), pp. 239–257. ISSN: 1556-1607. DOI: 10.1007/s11412-009-9070-z. URL: http://link.springer.com/article/10.1007/s11412-009-9070-z (visited on 12/01/2015).

[104] Bernard P. Ricca, Nicole Bowers, and Michelle E. Jordan. "Seeking emergence through temporal analysis of collaborative-group discourse: A complex-systems approach". In: *The Journal of Experimental Education* (July 8, 2019). Publisher: Routledge, pp. 1–17. ISSN: 0022-0973. DOI: 10.1080/00220973.2019.1628691. URL: https://doi.org/10.1080/00220973.2019.1628691.

[105] Carly Robinson, Michael Yeomans, Justin Reich, Chris Hulleman, and Hunter Gehlbach. "Forecasting student achievement in MOOCs with natural language processing". In: *LAK '16*. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. Journal Abbreviation: LAK '16. New York, NY, USA: ACM, 2016, pp. 383–387. DOI: 10.1145/2883851.2883932. URL: http://doi.acm.org/10.1145/2883851.2883932 (visited on 10/28/2016).

[106] Rebecca L. Robinson, Reanelle Navea, and William Ickes. "Predicting final course performance from students' written self-introductions: A LIWC analysis". In: *J. Lang. Soc. Psychol.* (Feb. 2013), p. 0261927X13476869. ISSN: 0261-927X. DOI: 10.1177/0261927X13476869. URL: http://jls.sagepub.com/content/early/2013/02/07/0261927X13476869.

[107] Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, and Dragan Gašević. "Analysing social presence in online discussions through network and text analytics". In: Proceedings of the 19th IEEE International Conference on Advanced Learning Technologies (ICALT'19). New Jersey, USA: IEEE, 2019, pp. 163–167. DOI: 10.1109/ICALT.2019.00058. URL: http://dx.doi.org/10.1109/ICALT.2019.00058.

[108] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning". In: *International Journal of Computer-Supported Collaborative Learning* 3.3 (Sept. 1, 2008), pp. 237–271. ISSN: 1556-1615. DOI: 10.1007/s11412-007-9034-0. URL: https://doi.org/10.1007/s11412-007-9034-0.

[109] David Shaffer, David Hatfield, Gina Navoa Svarovsky, Padraig Nash, Aran Nulty, Elizabeth Bagley, Ken Frank, André A Rupp, and Robert Mislevy. "Epistemic network analysis: A prototype for 21st-century assessment of learning". In: *International Journal of Learning and Media* 1.2 (May 1, 2009), pp. 33–53. DOI: 10.1162/ijlm.2009.0013. URL: http://dx.doi.org/10.1162/ijlm.2009.0013 (visited on 01/14/2016).

[110] Sharon Slade and Paul Prinsloo. "Learning analytics ethical issues and dilemmas". In: *American Behavioral Scientist* 57.10 (Oct. 1, 2013), pp. 1510–1529. ISSN: 0002-7642. DOI: 10.1177/0002764213479366. URL: http://abs.sagepub.com.ezproxy.is.ed.ac.uk/content/57/10/1510 (visited on 06/19/2016).

[111] Vilaythong Southavilay, Kalina Yacef, Peter Reimann, and Rafael A. Calvo. "Analysis of collaborative writing processes using revision maps and probabilistic topic models". In: *LAK '13*. Proceedings of the Third International Conference on Learning Analytics and Knowledge. Journal Abbreviation: LAK '13. New York, NY, USA: ACM, 2013, pp. 38–47. DOI: 10.1145/2460296.2460307. URL: http://doi.acm.org.proxy.lib.sfu.ca/10.1145/2460296.2460307.

[112] J. Michael Spector, M. David Merrill, Jeroen van Merrienboer, and Marcy P. Driscoll, eds. *Handbook of Research on Educational Communications and Technology: Third Edition*. 3 edition. New York: Routledge, Dec. 2007. ISBN: 978-0-415-96338-1.

[113] Angela E. B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, and Sidney K. D'Mello. "I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving". In: *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW 2019). Publisher: ACM New York, NY, USA, pp. 1–19. URL: https://dl.acm.org/doi/abs/10.1145/3359296.

[114] Florence R. Sullivan and P. Kevin Keith. "Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions". In: *Br. J. Educ. Technol.* 50.6 (Nov. 26, 2019), pp. 3047–3063. ISSN: 0007-1013. DOI: 10.1111/bjet.12875. URL: https://

onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12875.

[115] Yizhou Sun, Hongbo Deng, and Jiawei Han. "Probabilistic models for text mining". In: *Mining Text Data*. Ed. by Charu C. Aggarwal and Chengxiang Zhai. Springer US, Jan. 1, 2012, pp. 259–295. ISBN: 978-1-4614-3222-7. URL: http://link.springer.com.proxy.lib.sfu.ca/chapter/10.1007/978-1-4614-3223-4_8.

[116] Yla R. Tausczik and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of Language Society and Psychology* 29.1 (Mar. 2010), pp. 24–54. ISSN: 0261-927X. DOI: 10.1177/0261927X09351676. URL: http://jls.sagepub.com.proxy.lib.sfu.ca/content/29/1/24.

[117] Conrad Tucker, Barton K Pursel, and Anna Divinsky. "Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes". In: 2014 ASEE Annual Conference & Exposition. June 15, 2014, pp. 24.907.1–24.907.14. URL: https://peer.asee.org/22840.pdf.

[118] Thomas Daniel Ullmann. "Automated detection of reflection in texts. A machine learning based approach". PhD thesis. The Open University, 2015.

[119] Thomas Daniel Ullmann, Fridolin Wild, and Peter Scott. "Comparing automatically detected reflective texts with human judgements". In: 2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning. 2013, pp. 101–116. URL: http://oro.open.ac.uk/id/eprint/37830 (visited on 09/22/2017).

[120] Taylor Walsh and William G Bowen. *Unlocking the Gates: How and Why Leading Universities Are Opening Up Access to Their Courses*. Princeton: Princeton University Press, Jan. 17, 2011. 320 pp. ISBN: 978-0-691-14874-8. URL: http://www.amazon.com/Unlocking-Gates-Leading-Universities-Opening/dp/0691148740.

[121] Xu Wang, Miaomiao Wen, and Carolyn P Rosé. "Towards triggering higher-order thinking behaviors in MOOCs". In: *LAK '16*. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. Journal Abbreviation: LAK '16. New York, NY, USA: ACM, 2016, pp. 398–407. DOI: 10.1145/2883851.2883964. URL: http://doi.acm.org/10.1145/2883851.2883964 (visited on 05/10/2016).

[122] Kellie Webster, Marta R Costa-jussà, Christian Hardmeier, and Will Radford. "Gendered Ambiguous Pronoun (GAP) shared task at the gender bias in NLP workshop 2019". In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. aclweb.org, 2019, pp. 1–7. URL: https://www.aclweb.org/anthology/W19-3801.pdf.

[123] M Wen, D Yang, and C Rose. "Sentiment analysis in MOOC discussion forums: What does it tell us?" In: *Educational data mining 2014* (2014). Publisher: Citeseer. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.728.1722&rep=rep1&type=pdf.

[124] Miaomiao Wen, Diyi Yang, and Carolyn Rosé. "Linguistic reflections of student engagement in massive open online courses". In: *Proceedings 14th International Conference on Web and Social Media*. Ann Arbor, MI: AAAI, 2014, pp. 525–534. URL: http://meefen.github.io/notes/2014/05/01/Wen2014/ (visited on 01/24/2015).

[125] Alyssa Friend Wise, Yi Cui, Wanqi Jin, and Jovita Vytasek. "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling". In: *The Internet and Higher Education* 32 (Jan. 1, 2017), pp. 11–28. ISSN: 1096-7516. DOI: 10.1016/j.iheduc.2016.08.001. URL: http://www.sciencedirect.com/science/article/pii/S1096751616300513.

[126] Alyssa Friend Wise and Baruch B. Schwarz. "Visions of CSCL: Eight provocations for the future of the field". In: *International Journal of Computer-Supported Collaborative Learning* 12.4 (Dec. 1, 2017), pp. 423–467. ISSN: 1556-1615. DOI: 10.1007/s11412-017-9267-5. URL: https://doi.org/10.1007/s11412-017-9267-5.

[127] Wanli Xing, Vitaliy Popov, Gaoxia Zhu, Paul Horwitz, and Cynthia McIntyre. "The effects of transformative and non-transformative discourse on individual performance in collaborative-inquiry learning". In: *Comput. Human Behav.* 98 (Sept. 1, 2019), pp. 267–276. ISSN: 0747-5632. DOI: 10.1016/j.chb.2019.04.022. URL: http://www.sciencedirect.com/science/article/pii/S0747563219301700.

[128] Xiaoxi Xu, Tom Murray, Beverly Park Woolf, and David Smith. "If you were me and i were you - Mining social deliberation in online communication". In: Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013). Memphis, TN, USA: International Educational Data Mining Society, July 2013, pp. 208–216.

[129] Wenfei Yan, Nia Dowell, Caitlin Holman, Stephen S Welsh, Heeryung Choi, and Christopher Brooks. "Exploring learner engagement patterns in Teach-Outs: Using topic, sentiment and on-topicness to reflect on pedagogy". In: Proceedings of the 9th International Conference for Learning Analytics & Knowledge. Ed. by D Azcona and R Chung. Temple, AZ: ACM, 2019.

[130] Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. "Exploring the effect of confusion in discussion forums of massive open online courses". In: Proceedings of the 2nd ACM conference on Learning@Scale (L@S 2015). ACM,

2015, pp. 121–130. DOI: 10.1145/2724660.2724677.

[131] Jaebong Yoo and Jihie Kim. "Predicting Learner's Project Performance with Dialogue Features in Online Q&A Discussions". In: *Intelligent Tutoring Systems*. Ed. by Stefano A Cerri, William J Clancey, Giorgos Papadourakis, and Kitty Panourgia. Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 1, 2012, pp. 570–575. ISBN: 978-3-642-30949-6. URL: http://link.springer.com/chapter/10.1007/978-3-642-30950-2_74 (visited on 02/22/2014).

[132] Juan Zheng, Wanli Xing, and Gaoxia Zhu. "Examining sequential patterns of self- and socially shared regulation of STEM learning in a CSCL environment". In: *Comput. Educ.* 136 (July 1, 2019), pp. 34–48. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2019.03.005. URL: http://www.sciencedirect.com/science/article/pii/S0360131519300636.