Modeling Learners' Social Centrality and Performance through Language and Discourse

Nia M. Dowell
Department of Psychology
Institute for Intelligent Systems
University of Memphis
365 Innovation Drive
Memphis, TN 38152
+1 901-678-5102
ndowell@memphis.edu

Arthur C. Graesser
Department of Psychology
Institute for Intelligent Systems
University of Memphis
365 Innovation Drive
Memphis, TN 38152
+1 901-678-5102
a-graesser@memphis.edu

Thieme A. Hennis
Delft Extension School
Delft University of Technology
2628 BX, Delft
+31651855220
t.a.hennis@tudelft.nl

Oleksandra Skrypnyk
School of Education
University of South Australia
Adelaide, Australia
+61 402918694
olesandra.skrypnyk@mymail.unisa.edu.au

Shane Dawson
Learning and Teaching Unit
University of South Australia
Adelaide, Australia
+61 883027850
shane.dawson@unisa.eduau

Pieter de Vries
Systems Engineering Department
Participatory Systems Design
Delft University of Technology
2628 BX Delft, Netherlands
+31651517278
pieter.devries@tudelft.nl

Srećko Joksimović
School of Interactive Arts and Technology
Simon Fraser University
Burnaby, Canada
+1604-375-2496
sjoksimo@sfu.ca

Dragan Gašević
Schools of Education and Informatics
University of Edinburgh
Edinburgh, United Kingdom
+44 131 651 6243
dragan.gasevic@ed.ac.uk

Vitomir Kovanović
Schools of Education and Informatics
University of Edinburgh
+1604-375-2496
v.kovanovic@ed.ac.uk

ABSTRACT

There is an emerging trend in higher education for the adoption of massive open online courses (MOOCs). However, despite this interest in learning at scale, there has been limited work investigating the impact MOOCs can play on student learning. In this study, we adopt a novel approach, using language and discourse as a tool to explore its association with two established measures related to learning: traditional academic performance and social centrality. We demonstrate how characteristics of language diagnostically reveal the performance and social position of learners as they interact in a MOOC. We use Coh-Metrix, a theoretically grounded, computational linguistic modeling tool, to explore students' forum postings across five potent discourse dimensions. Using a Social Network Analysis (SNA) methodology, we determine learners' social centrality. Linear mixed-effect modeling is used for all other analyses to control for individual learner and text characteristics. The results indicate that learners performed significantly better when they engaged in more expository style discourse, with surface and deep level cohesive integration, abstract language, and simple syntactic structures. However, measures of social centrality revealed a different picture. Learners garnered a more significant and central position in their social network when they engaged with more narrative style discourse with less overlap between words and ideas, simpler syntactic structures and abstract words. Implications for further research and practice are discussed regarding the misalignment between these two learning-related outcomes.

Keywords

Social Centrality, Learning, Discourse, Coh-Metrix, MOOCs

1. INTRODUCTION

Advances in educational technologies and a desire for increased access to learning, are enabling the development of pedagogical environments at scale, such as Massive Open Online Courses (MOOCs) [41]. Open online courses have the potential to advance education on a global level, by providing the masses with broader access to lifelong learning opportunities. Additionally, the insulated nature of the MOOC web-based platforms allows valuable learning dynamics to be detailed at unprecedented resolution and scale. As such, the digital traces left by learners are regarded as a gold mine that can offer powerful insights into the learning process, resulting in the advancement of educational sciences and substantially improved learning environments.

While the scale of the data has grown, making sense of data from the learning environments is not a novel effort. Prior to the arrival of MOOCs, similar endeavors were undertaken at smaller scale in the domains of computer-supported collaborative learning and intelligent tutoring systems, among others. The volume of student behavior and performance data produced in those interactions motivated the fields of educational data mining (EDM) and learning analytics (LA) [37]. Both of these research communities have leveraged this fine-grained data and aligned with educational

theory. The EDM community offer methods for exploring learners and educational settings, while LA focuses on the measurement, collection, and analyses that aim at optimizing the learning process [38]. That said, inquiring into MOOCs and other unexplored learning environments requires inputs from both communities. Direct application of methodologies, theoretical frameworks, and established analytics require deeper understanding of the relationships between parts of the whole, to enable drawing the relevant parallels with existing research.

Drawing on this, this paper adopts a novel approach, which uses language and discourse as a tool to explore its association with two established measures of learning, namely traditional academic performance and social centrality. Specifically, we are investigating the extent to which characteristics of language diagnostically reveal the performance and social position of students as they interact in a MOOC. As a methodological contribution, we adopt a theoretically grounded computational linguistics modeling approach to explore students' forum posting, within a MOOC, across five potent discourse dimensions. In line with current practice, we implement a Social Network Analysis (SNA) methodology to monitor and detect learners' social centrality. Students' performance in the course, i.e. course grade, is represented by an aggregate measure combining scores for the essays submitted during the MOOC, and a final peer-evaluated, open-ended written-assignment. Linear mixed-effects modeling approach is used for all other analyses to control for individual learner and text characteristics. This design allows us to contrast the linguistic profiles of high performing learners and centrally situated learners. Consequently, we gain insights into the qualitative differences between these two different learning-related outcomes. Finally, we explored whether the discourse features characterizing learning-related outcomes varied within different learner population, namely across all learners in the MOOC and within a subset of active learners.

The subsequent sections of the paper are organized as follows. First, we provide a brief overview of language and discourse situated within the contexts of psychological frameworks of comprehension and learning. Then, the following two sections address the traditional application of social network analysis, including theoretical foundations, as well as interpretations applied in MOOCs research. We then move into the methodological features of the current investigation, and conclude the paper with a detailed discussion of the results in the context of theory, as well as a general discussion of the theoretical, methodological, and practical implications for the EDM and LA community.

2. THEORETICAL BACKGROUND

2.1 Language and Discourse

Across academic fields, there has been a burgeoning literature demonstrating the usefulness of language and discourse in predicting a number of psychological, affective, cognitive, and social phenomena, ranging from personality to emotion to learning to successful group interactions (e.g. [6,10,26]). Within the educational contexts, there are many critical learning-related constructs that cannot be directly measured, but can be inferred from measurable signals like language and other behavioral patterns. Working with these barriers, we are continually pushing beyond the boundaries of established implementation. In that realm, it is particularly important that these endeavors be guided by established theory. A number of psychological models of discourse comprehension and learning, such as the construction-integration, constructionist, and indexical-embodiment models, lend themselves nicely to the exploration of learning related phenomena in computer-mediated educational environments. These psychological frameworks have identified the representations, structures, strategies, and processes at multiple levels of discourse [16,23,40]. Five levels have commonly been offered in these frameworks: (1) words, (2) syntax, (3) the explicit textbase, (4) the situation model (sometimes called the mental model), and (5) the discourse genre and rhetorical structure (the type of discourse and its composition). In the learning context, learners can experience communication misalignments and comprehension breakdowns at different levels. Such breakdowns and misalignments have important implications for the learning process. In this paper we adopt this multilevel approach to the analysis of language and discourse.

With regard to analytical approaches, there has been extensive knowledge gleamed from manual content analyses of learners' discourse during educational interactions, however, these methods are no longer a viable option with the increasing scale of educational data. As such, researchers have been incorporating automated linguistic analysis, including more shallow level word counts and deeper level discourse analysis approaches. Both levels of linguistic analysis are informative. Content analysis using word-counting methods allows getting a fast overview of learners' participation levels, as well as assessing specific words. For instance, a study by Wen and colleagues [43] is an example of incorporating word counts (LIWC) of theory-informed and carefully selected words with manual message coding. Their work links specific (and thus identifiable and countable) words used by the students with the degree of their engagement and commitment to remain in the course.

To extend analysis of learning-related phenomena beyond the shallow level word counts, one needs to conduct a deeper level discourse analysis employing sophisticated natural language processing techniques, e.g. syntactic parsing and cohesion computation. For example, Dowell and colleagues [11] explored the possibility of using discourse features to predict student performance during collaborative learning interactions. Their results indicated that students who engaged in deeper cohesive integration and generated more complicated syntactic structures performed significantly better. In line with this, Cade and others [3] demonstrated that cognitive linguistic cues can be used in detecting students' socio-affective attitudes towards fellow students in CMCL environments. As a whole, these studies highlight the critical and complex role of language and discourse. This is, perhaps, not surprising, since language is a primary means for expressing and communicating information in computer-mediated learning environments.

2.2 Social Network Analysis in Educational Research

Social Network Analysis (SNA) is a methodology that is increasingly being used for analyzing learning-related phenomena, especially in online settings [25]. SNA has gained popularity with researchers who view social relationships between students as an aspect influencing overall educational experience and learning outcomes (i.e. [33]). Its methodology is grounded in systematic empirical data [4:8], as well as "motivated by a relational intuition based on ties connecting social actors" (*ibid.*). Studies that employ SNA, aim at revealing the role of social relationships in learning, around such issues as *who is central in a social learning network, who is talking to whom,* and *who is participating peripherally* and *how those interaction patterns influence learning* [4,25,42]. Due to such focus, SNA provides the

theoretical and methodological tools to understand activities and social processes that students and teachers engage with. [25,31]

Traditionally, the analyses of social networks of learners have been derived from participation in discussion forums in formal online courses. The relationship between learners' position in a social network and student academic performance is well documented, in this context [5,14,33]. The general finding in this literature shows more centrally situated learners tend to get higher final grades [33]. Moreover, Russo and Koesten [34] showed that network centrality (measured as in-degree and out-degree) is a significant predictor of cognitive learning outcome. Rizzuto and others [32] found that network density significantly predicted the scores reflecting course material comprehension. Reflective of the finding from these studies a students' position in a network also influences their overall sense of community [9]. These studies suggest, in the context of formal online learning, individuals who are centrally positioned in their network perform better, and feel a stronger sense of connection than students that are more peripheral in the network structure.

In the context of MOOCs, SNA is increasingly used to explore learning-related phenomena [13]. For example, Gilliani et al. [15] applied SNA to capture broad trends in communication and the roles of individuals in facilitating discussions [15]. Another example of SNA in MOOCs is a study by Yang and colleagues [44], which suggests that learners who join forums (i.e. networks of learners) earlier are likely to persist in the course, in contrast to their counterparts who joined later and found it difficult to form social bonds. This finding is parallel to prior findings in the domain of traditional online learning revealing that learners central to the social network tend to have a higher sense of belonging to the group [8]. However, there is research that suggests the interpretation of SNA in MOOCs requires further attention. For example, the relationship between student centrality in MOOC discussion forums and their academic performance (i.e., final grade), has been shown to be context dependent [21]. Jiang and colleagues [21] demonstrated that in Algebra MOOC, betweenness and degree centrality yielded significant correlation with the final grade, while none of the metrics analyzed (i.e., closeness, degree, and betweenness centrality) was significantly correlated with the learning outcome in a Financial Planning MOOC.

Automated linguistic analysis of student interactions, within computer-mediated learning environments, can compliment SNA techniques by adding rich contextual information to the structural patterns of learner interactions. However, the combination of these two analytical methods is relatively sparse in the literature, beyond a few noteworthy exceptions [22,36]. Similar to the current work, is Joksimović and colleagues' [22] analysis of students' interaction patterns in a distributed MOOC, i.e. learner interactions take place via social media, and the course is based on connectivist pedagogy. Their findings pinpoint specific discourse features that were predictive of a learners' accumulation of social capital.

2.3 Research Questions

To summarize, SNA is a widely used tool for exploring learning processes that take place in MOOCs, largely due to its theoretical foundation and established application in formal educational contexts. However, given the open nature of scaled online courses, the interpretation of SNA in MOOCs requires further attention. This study approaches language as the primary means for communication and a window into inferring learning-related phenomena. We apply discourse analysis as a proxy for providing qualitative information about the position of learners in the network and their performance. The analysis focuses around the following research questions: Which characteristics of language diagnostically reveal the performance and social position of students as they interact in a MOOC? And do these features operate similarly with different learner populations, namely across all learners in a MOOC and within a subset of active learners?

3. METHODS

3.1 Participants

The study analyzed forum discussion posted on the edX platform, within the course NGI101x Next Generation Infrastructures (NGIx). It ran for 8 weeks in the period of April 22 – July 8, 2014. The subject area of the analyzed MOOC fell under the domain of applied non-life soft sciences [2]; the course objective was to introduce the complexity of infrastructure systems, familiarize students with the main concepts within the area, as well as with the practical approaches to the infra-systems analysis. In total 16,091 participants enrolled and 517 received certificate of completion (passed). To pass the course the students needed to receive a score of 0.7 (out of 1) or higher. The grade was derived from the submission of 3-6 open-ended papers (60% of the grade) and a final issue paper (40% of the grade) that was peer assessed by several co-learners. The dataset for the analysis in this study included 1,754 participants (N_{post}=7,244, M=4.13, SD=9.85, $Q1$=1.0, $Q3$=4.0, Min=1.0, Max=180), i.e. all those who used the course forum. Forum data was collected from the edX platform in the JSON format, and included all the information specified within the edX discussion forums data documentation[1].

3.2 Analyses

3.2.1 Social Network Analysis

Although other approaches have been proposed, the most common approach for extracting social networks from online discussions is to consider each message as directed to the previous one in the thread [25,31]. In the current study, we followed the approach suggested in [24,25,31], among others. Specifically, social graph representing interaction within the discussion forum included all the students who posted a message(s). For example, author A1 initiated the discussion, and author A2 posted a message directly into the thread, in reply to A1's initial thread message, we would add directed edge A2->A1. Then, if author A3 replied to the message posted by author A2, we would include a direct edge A3->A2 to the graph. If author A4 started a nested discussion as a reply to A1's initial post, then A4 would have a direct edge to A1. The concept of centrality has been commonly used to assess the importance of an individual node within a social network [12,42]. The following well-established SNA measures [42], that capture various notions of a graph structural centrality, were calculated for each learner in the social network extracted:

- **Degree Centrality** – the number of edges a node has in a network;
- **Closeness Centrality** – the distance of an individual node in the network from all the other nodes;
- **Betweenness Centrality** – the number of shortest paths between any two nodes that pass via a given node.

Degree centrality is generally used to capture the "potential for activity in communication" [12:219] or the *popularity* [31] of a node in a social network. Betweenness centrality, on the other hand, represents a *potential for influence* over the information

[1] http://devdata.readthedocs.org/en/latest/internal_data_formats/discussion_data.html

flow, as it *bridges* the parts of the network that were disconnected otherwise [12,31,42]. Finally, the concept of closeness centrality refers to the distance between a learner and the other participants of the network. In a MOOC, closeness centrality can be interpreted as the extent to which a learner is in the middle of what is happening on the forum. The relationship between students' linguistic properties and their position in the social network, measured through the three properties described above, has been investigated in this study. The social network variables were analyzed using *igraph 0.7.1* [7], a comprehensive R software package for complex social network analysis research.

3.2.2 Coh-Metrix Analyses

Prior to Coh-Metrix analyses, the logs were cleaned and parsed to facilitate a student level evaluation. Thus, text files were created that included all contributions from a single learner, yielding a total of 1,754 text files, one for each student. All files were then analyzed using Coh-Metrix. Coh-Metrix (www.cohmetrix.com) is a computational linguistics facility that provides measures of over 100 measures of various types of cohesion, including co-reference, referential, causal, spatial, temporal, and structural cohesion [18,26]. Coh-Metrix also has measures of linguistic complexity, characteristics of words, and readability scores. Currently, Coh-Metrix is being used to analyze texts in K-12 for the Common Core standards and states throughout the U.S. More than 50 published studies have demonstrated that Coh-Metrix indices can be used to detect subtle differences in text and discourse [26].

There is a need to reduce the large number of measures provided by Coh-Metrix into a more manageable number of measures. This was achieved in a study that examined 53 Coh-Metrix measures for 37,520 texts in the TASA (Touchstone Applied Science Association) corpus, which represents what typical high school students have read throughout their lifetime [17]. A principal components analysis was conducted on the corpus, yielding eight components that explained an impressive 67.3% of the variability among texts; the top five components explained over 50% of the variance. Importantly, the components aligned with the language-discourse levels previously proposed in multilevel theoretical frameworks of cognition and comprehension [16,23,40]. These theoretical frameworks identify the representations, structures, strategies, and processes at different levels of language and discourse, and thus are ideal for investigating trends in learning-oriented conversations. Below are the five major dimensions, or latent components, that may be useful for understanding trends in learning-oriented, but inherently social, conversations:

- **Narrativity.** The extent to which the text is in the narrative genre, which conveys a story, a procedure, or a sequence of episodes of actions and events with animate beings. Informational texts on unfamiliar topics are at the opposite end of the continuum.
- **Deep Cohesion.** The extent to which the ideas in the text are cohesively connected at a deeper conceptual level that signifies causality or intentionality.
- **Referential Cohesion.** The extent to which explicit words and ideas in the text are connected with each other as the text unfolds.
- **Syntactic Simplicity.** Sentences with few words and simple, familiar syntactic structures. At the opposite pole are structurally embedded sentences that require the reader to hold many words and ideas in working memory.
- **Word Concreteness.** The extent to which content words that are concrete, meaningful, and evoke mental images as opposed to abstract words.

3.2.3 Data Preparation

The students' performance, linguistic and network data were merged to facilitate subsequent statistical analyses. Following this, the scores were centered and normalized by removing any outliers. Specifically, the normalization procedure involved Winsorising the data based on each variable's upper and lower percentile. Finally, we were interested in exploring whether the discourse features characterizing learning-related outcomes varied within different learner population, namely across all learners in the MOOC and within a subset of active learners. To enable this analysis, we created two datasets. The *All Learner* dataset contained data for the full 1,754 students that participated in the MOOC. We operationalized active students as those learners who made 4 or more posts in the MOOC. The cut-off point was chosen because the top 25% of learners made 4 or more posts. The resulting *Active Learner* dataset contained the data for those top 471 learners.

3.2.4 Statistical analyses

A mixed-effects modeling approach was adopted for all analyses due to the structure of the data (e.g., inter-individual and word count variability) [30]. Mixed-effects models include a combination of fixed and random effects and can be used to assess the influence of the fixed effects on dependent variables after accounting for any extraneous random effects. The primary analyses focused on identifying the association between the discourse features, namely, Narrativity, Deep Cohesion, Referential Cohesion, Syntax Simplicity, and Word Concreteness and the learning outcomes, measured through learners' social centrality and grades. Therefore, we identified two sets of dependent measures in the present analyses: (1) learners' social centrality (Closeness, Degree, and Betweenness) and (2) learners' performance in the course (the final grade). The independent variables in all models were the five discourse features of interest.

Additionally, the influence of language on learning and social capital might vary depending on relevant learner characteristics. For instance, discourse may play a more meaningful role, for student performance and social position in a network, for more active learners than less active learners [25]. This would be in line with Gillani and others [15] conclusion that suggests the social network extracted from the learner interactions "was a noise-corrupted version of the "true" network" (p.2). Thus, we decided to further refine our analysis and create social graph only for those learners who actively participated in discussions (for the cut-off point see Section 4.2). This resulted in an additional four models, labeled as *Active Learners*, exploring the influence of language on learners' social centrality (three models) and performance (one model) for the most active participants in the course.

It is important to note that in addition to constructing the models with the five discourse features as fixed effects, *null models* with the random effects (*learner* and *word count*) but no fixed effects were also constructed. A comparison of the null random effects only model with the fixed-effect models allows us to determine whether discourse predicts social centrality and performance above and beyond the random effects. Akaike Information Criterion (AIC), Log Likelihood (LL) and a likelihood ratio test were used to determine the best fitting and most parsimonious model. In addition, we also estimate effect sizes for each model, using a pseudo R^2 method, as suggested by Nakagawa and Schielzeth [28]. For mixed-effects models, R^2 can be characterized into two varieties: marginal R^2 and conditional R^2. Marginal R^2 is associated with variance explained by fixed factors, and conditional R^2 is can be interpreted as the variance explained

by the entire model, namely random and fixed factors. Both marginal (R^2_m) and conditional (R^2_c) R^2 convey unique and relevant information regarding the model fit and variance explained, and so we report both here. The lme4 package in R [1] was used to perform all the required computation.

4. RESULTS AND DISCUSSION

4.1 Discourse and Learning

First, we assessed the relationship between learners discourse patterns and performance in the MOOC. The likelihood ratio tests indicated that both the *All Learner* and *Active Learner* models yielded a significantly better fit than the null model with $\chi^2(5) = 82.57$, $p = .001$, $R^2_m = .05$, $R^2_c = .93$, and $\chi^2(5) = 85.44$, $p = .001$, $R^2_m = .21$, $R^2_c = .95$, respectively. A number of conclusions can be drawn from this initial model fit evaluation and inspection of R^2 variance. First, the model comparisons imply that the discourse features were able to add a significant improvement in predicting the learners' performance above and beyond individual participant characteristics. Second, for the *All Learner* model, discourse and individual participant features explained about 93% of the predictable variance, with 5% of the variance being accounted for by the discourse features. However, the discourse features alone were able to explain a total of 21% of predictable variance in active learners' performance. The observed difference in variance suggests discourse features are more accurate at predicting active learners' performance than that of learners who are less active in the course. It is important to note that the difference in the explained variance for the *All Learner* and *Active Learner* models is not a result of the students simply being more prolific, because we controlled for number of words. Instead the findings might be reflecting a more substantive difference for the active students' potency of thought integration, complexity and communication style, beyond the observation that they are communicating more, compared to the overall learner population. Table 1 shows the discourse features that were predictive of learning performance for both the *All Learner* and *Active Learner* models. As can be seen from Table 1, all five levels of discourse were predictive of learning performance for the *All Learner* models, and four of the five levels were predictive of learning in the *Active Learner* models. Specifically, learners who engaged in more expository style discourse with referential and deep level cohesive integration, abstract language, and simple syntactic structures performed significantly better in the course.

Narrative discourse expresses events and actions performed by characters that unfold over time, as is typical in everyday oral communication, folktales, drama, and short stories [35]. In contrast to narrative, expository language is decontextualized and generally informs the audience about new concepts, broad truths, and technical material as in the case of academic articles and college textbooks. The genre of a text can be particularly revealing with regard to its difficulty. For example, narrative text is substantially easier to read, comprehend, and recall than informational or expository text [16]. From a constructionist theory [19,20] view, this is because expository discourse frequently presents abstract categories and less familiar information that require learners to have extensive background knowledge about the topics in order to generate the inferences necessary for comprehension [39]. As a reminder, our measure of narrativity/expository is a single continuum, wherein higher numbers indicate narrative style discourse and lower numbers indicate expository style discourse. Thus, the negative findings for Narrativity (Table 1) can be extrapolated to conclude that learners who articulated their responses in a more expository style, mirroring the informational nature of their class material, extracted enough information about the subject to generate inferential processing. Such interpretation is in line with other research showing knowledgeable students develop more comprehensive representations from material than less knowledgeable students [27], and can inferentially relate the information they derive from text better than readers with less background knowledge.

In line with Kintsch's [23] construction-integration theory, Coh-Metrix distinguishes between multiple types of cohesion which fall under two main forms, namely textbase (i.e. referential cohesion) and situation model cohesion (i.e. deep cohesion). Referential or textbase cohesion is primarily maintained through the bridging devices, i.e. the overlap in words, or semantic references. In this context, the findings for referential cohesion suggest that learners who perform better, construct their messages using more bridging devices

A theory of situation model cohesion has been described by [45] that characterizes it as knowledge elaborations that are product of incorporating information derived from the explicit texts with background world knowledge. Coh-Metrix analyzes the situation model dimension on causation, intentionality, space, and time [26]. With regard to the findings for deep cohesion, this suggests that students who are learning are engaging in deeper integration of topics with their background knowledge, generating more inferences to address any conceptual and structural gaps, and consequentially increasing the probability of comprehension. The results for syntax show that simple syntactic structures were associated with better performance. However, this finding was not significant in the *Active Learner* model.

Table 1. Descriptive Statistics and Mixed-Effects Model Coefficients for Predicting Performance with Language

Measure	All Learner Model				Active Learner Model			
	M	***SD***	**β**	***SE***	***M***	***SD***	**β**	***SE***
Narrativity	0.00	1.00	-.20**	.02	-0.23	0.69	-.60**	.07
Deep Cohesion	0.00	1.00	.08**	. 02	0.27	0.55	.19*	.08
Referential Cohesion	0.00	1.00	.08**	. 02	-0.26	0.64	.35**	.07
Syntax Simplicity	0.00	1.00	.07**	. 02	0.36	0.67	.08	.07
Word Concreteness	0.00	1.00	-.13**	. 02	-0.25	0.51	-.35**	.09

Note: * $p < .05$; ** $p < .001$. Mean (***M***). Standard deviation (***SD***). Fixed effect coefficient (**β**). Standard error (*SE*). All Learner Model *N*= 1754, Active Learner Model *N*= 471.

Coh-Metrix measures psychological dimensions of words that influence language complexity. As a reminder, our measure of word concreteness is a single continuum, wherein scores are higher when a higher percentage of the content words are concrete, are meaningful, and evoked mental images – as opposed to being abstract. Thus, the negative findings for word concreteness show learners who engaged using more abstract language performed significantly better in the course. There are interesting interpretations from the view of Petty and Cacioppo's Elaboration Likelihood Model (ELM) [29]. The ELM outlines several factors that affect both the ability and motivation to elaborate on arguments contained in messages. If ability to process is impaired, or motivation to process is low, the elaboration and thought density of the learners' communication

would likely suffer. With the exception of syntax ease, the findings suggest students who adopt central route linguistic characteristics perform significantly better than those who use peripheral linguistic features.

4.2 Discourse and Social Centrality

Next, we investigated the relationship between learners' discourse patterns and their position in the social network. The likelihood ratio tests indicated that the *All Learner* models for Closeness, Betweenness and Degree yielded a significantly better fit than the null random effects only models with $\chi^2(5) = 135.74$, $p = .001$, $R^2_m = .07$, $R^2_c = .93$, $\chi^2(5) = 25.63$, $p = .0001$, $R^2_m = .01$, $R^2_c = .91$, and $\chi^2(5) = 62.19$, $p = .0001$, $R^2_m = .02$, $R^2_c = .94$, respectively. Similarly, for the *Active Learner* models, the likelihood ratio tests indicated that Closeness, Betweenness and Degree yielded a significantly better fit than the null models with $\chi^2(5) = 38.39$, $p = .0001$, $R^2_m = .08$, $R^2_c = .94$, $\chi^2(5) = 45.92$, $p = .0001$, $R^2_m = .09$, $R^2_c = .94$, and $\chi^2(5) = 63.78$, $p = .0001$, $R^2_m = .12$ and $R^2_c = .96$, respectively. Similar to the results for performance, the model comparisons imply that the discourse features were able to add a significant improvement in predicting the learners' social centrality above and beyond participant characteristics. In line with this, across the three *All Learner* models, our features explained about 92% of the predictable variance, with 10% of the variance being accounted for by the linguistic features. However, the discourse features were able to explain a total of 29% of predictable variance in active learners' social centrality. Again, this suggests discourse more accurately predicts active learners' position than less active learners. The details of the *All Learner* and *Active Learner* models are reported in Table 2 and Table 3. Interestingly, the pattern of discourse features associated with learners' social centrality differed from the one observed for students' performance in the MOOC. Instead, learners who garnered central positions in the network engaged in narrative discourse with lower referential cohesion, abstract words and simple syntactic structures. With the exception of word abstractness, this pattern is indicative of informal communication.

Across all learners, higher closeness centrality is characterized by more narrative style discourse with less overlap between words and ideas (i.e. low referential cohesion), simple syntactic structures and abstract words. For active learners, the pattern is similar, with only narrativity and referential cohesion being significant. The conventional interpretation of closeness centrality indicates the efficiency of an individual in passing the information directly onto all other individuals in the social network [12]. Due to the nature of MOOC centralized forums, it can be inferred that shorter distance to all the learners can be obtained, if the individual participates in many various discussion threads. Therefore, individuals who are more active and initiate more topical messages yielding replies from many other learners, or reply to many other discussions, would use language characterized by simpler structures, narrative style, and lower referential cohesion. Similar pattern for higher narrativity and lower referential cohesion has been observed in the discourse of learners with high degree and betweenness centrality in a distributed MOOC – a course where learner interactions take place on social media, rather than on the course platform [22]. Although conventionally betweenness centrality is associated with the brokering of information between sub-groups, this is questionable in the context of an online open centralized discussion forum.

These results suggest that learners who attained a more prominent social centrality position used more conversational style discourse. Most noteworthy is that these results do not mirror the pattern observed for high performing learners. On the contrary, linguistic profiles of high performing learners are characterized by formal discourse that uses expository style language (i.e. negative relationship with narrativity), and more surface and deep level cohesive integration (i.e. positive relationship with referential and deep cohesion) (Table 1).

Table 2. All Learner Mixed-Effects Model Coefficients for Predicting Social Network Centrality with Language

Measure	Closeness		Betweenness		Degree	
	β	***SE***	**β**	***SE***	**β**	***SE***
Narrativity	.070*	.03	.03	.03	.07**	.02
Deep Cohesion	.008	.02	.01	.02	-.02	.02
Referential Cohesion	-.15**	.03	-.02	.03	-.06**	.02
Syntax Simplicity	.13**	.03	.09*	. 03	.06*	.02
Word Concreteness	-.09**	.03	-.03	. 02	-.05*	.02

Note: * $p < .05$; ** $p < .001$. Mean (***M***). Standard deviation (***SD***). Fixed effect coefficient (**β**). Standard error (*SE*). *N*= 1754.

Table 3. Active Learner Mixed-Effects Model Coefficients for Predicting Social Network Centrality with Language

Measure	Closeness		Betweenness		Degree	
	β	***SE***	**β**	***SE***	**β**	***SE***
Narrativity	.32**	.07	.17*	.07	.21**	.06
Deep Cohesion	-.06	.08	.02	.08	.05	.08
Referential Cohesion	-.33**	.07	.11	.07	.09	.07
Syntax Simplicity	.07	.07	.42**	.07	.47**	.07
Word Concreteness	.14	.09	-.07	.09	-.06	.09

Note: * $p < .05$; ** $p < .001$. Mean (***M***). Standard deviation (***SD***). Fixed effect coefficient (**β**). Standard error (*SE*). *N*= 471.

5. GENERAL DISCUSSION

This paper adopted a novel approach, which uses language and discourse as a tool to explore its association with two established measures of learning, namely traditional academic performance and social centrality. Specifically, we explored the extent to which characteristics of discourse diagnostically reveal the performance and social position of learners as they interact in a MOOC. The findings present some methodological, theoretical, and practical implications for the educational data mining and learning analytics communities. First, as a methodological contribution, we have highlighted the rich contextual information that can be gleaned from combing deeper level linguistic analysis and SNA. Particularly, discourse features add a significant improvement in predicting both the performance and social network positioning in MOOC forums.

Secondly, the results pose some important theoretical and practical implications for transferring analytic approaches to scaled environments without careful consideration. The results indicate that learners 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 communication network engaged using a more narrative style discourse with less overlap between words and ideas, simpler syntactic structures and abstract words. In other words, high performers and those with central

positions in the network are not necessarily the same individuals. The misalignment between the linguistic features associated with improved performance and more centrally located network positions is captured by the discrepant pattern for narrative, referential and deep cohesion. These three discourse features are inversely related with high performance and centrality in networks. This difference has important implications because these linguistic dimensions are strongly associated with comprehension according to construction-integration and constructivist theories.

The study also suggests that in open online environments two established measures of learning: traditional academic performance and social centrality reflect different learning outcomes. Academic performance represents a snapshot of students' mastery of the subject, and is one way of accessing the state of subject comprehension. Positioning in social network represents a snapshot of the participation processes and social learning activities. In this study, we demonstrate that the skills associated with these two learning-related outcomes differ.

It could be speculated that the observed misalignment between linguistic performance and social network position in the analyzed open online course, shows the difference in communication patterns of formal and informal learning environments. Formal learning environments have a clearer start and end, and often require participation related to the subject matter, as embedded in tasks, or course design. In open learning environments, adult learners can opt in and opt out of the learning situations. The issue is further complicated by the discussions being held by the learners on MOOC forums on various topics: from subject matter, to technical troubleshooting, or clarification of administrative issues. Centralized forums of MOOCs are more than a social learning space; they are also a communication space. As a result, learners' high activity on a number of issues during one or two weeks of the course may result in a more central position in the network of learners, but may not necessarily indicate that the learners engaged with the content, or demonstrated the required understanding of the subject at the end of the course.

It is unclear from this study what relationship should be deduced between learning and social centrality measures within in the open online environments. At the minimum, the findings suggest that the social positioning in a network of learners in a MOOC may not be equivalent with measured academic performance. Further research is needed to understanding what analytical approaches, such as SNA, are reflecting in emerging educational environments.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, HCC 0834847) and (DRK-12-0918409), the Institute of Education Sciences (R305G020018, R305A080589), The Gates Foundation, U.S. Department of Homeland Security (Z934002/UTAA08-063), Natural Sciences and Engineering Research Council of Canada (356029), Social Sciences and Humanities Research Council of Canada (435-2013-1708), and Canada Research Chairs Program. Any opinions, findings, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

7. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., et al. *lme4: Linear mixed-effects models using Eigen and S4.* 2014.

[2] Biglan, A. The characteristics of subject matter in different academic areas. *Journal of Applied Psychology 57*, (1973), 195–203.

[3] Cade, W.L., Dowell, N.M., Graesser, A.C., Tausczik, Y.R., and Pennebaker, J.W. Modeling student socioaffective responses to group interactions in a collaborative online chat environment. In J. Stamper, Z. Pardos, M. Mavrikis and B.M. McLaren, eds., *Proceedings of the 7th International Conference on Educational Data Mining.* Springer, Berlin, 2014, 399–400.

[4] Carolan, B.V. *Social Network Analysis Education: Theory, Methods & Applications.* SAGE Publications, Inc. SAGE Publications, Inc., 2014.

[5] Cho, H., Gay, G., Davidson, B., and Ingraffea, A. Social networks, communication styles, and learning performance in a CSCL community. *Computers & Education 49*, 2 (2007), 309–329.

[6] Chung, C.K. and Pennebaker, J.W. Using Computerized Text Analysis to Track Social Processes. In T. Holtgraves, ed., *Oxford Handbooks Online*. Oxford, 2014.

[7] Csardi, G. and Nepusz, T. The igraph Software Package for Complex Network Research. *InterJournal Complex Systems*, (2006), 1695.

[8] Dawson, S. Online forum discussion interactions as an indicator of student community. *Australasian Journal of Educational Technology 22*, 4 (2006), 495–510.

[9] Dawson, S. A study of the relationship between student social networks and sense of community. *Educational Technology & Society 11*, 3 (2008), 224–238.

[10] D'Mello, S. and Graesser, A.C. Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. *IEEE Transactions on Learning Technologies 5*, 4 (2012), 304–317.

[11] Dowell, N.M., Cade, W.L., Tausczik, Y.R., Pennebaker, J.W., and Graesser, A.C. What works: Creating adaptive and intelligent systems for collaborative learning support. In S. Trausan-Matu, K.E. Boyer, M. Crosby and K. Panourgia, eds., *Twelfth International Conference on Intelligent Tutoring Systems.* Springer, Berlin, 2014, 124–133.

[12] Freeman, L.C. Centrality in social networks conceptual clarification. *Social networks 1*, 3 (1979), 215–239.

[13] Gasevic, D., Kovanovic, V., Joksimovic, S., and Siemens, G. Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *The International Review of Research in Open and Distributed Learning 15*, 5 (2014).

[14] Gašević, D., Zouaq, A., and Janzen, R. "Choose Your Classmates, Your GPA Is at Stake!": The Association of Cross-Class Social Ties and Academic Performance. *American Behavioral Scientist*, (2013).

[15] Gillani, N., Yasseri, T., Eynon, R., and Hjorth, I. Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs. *Scientific reports 4*, (2014).

[16] Graesser, A.C. and McNamara, D.S. Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science 3*, 2 (2011), 371–398.

[17] Graesser, A.C., McNamara, D.S., and Kulikowich, J.M. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher 40*, 5 (2011), 223–234.

[18] Graesser, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z. Coh-metrix: analysis of text on cohesion and language. *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society, Inc 36*, 2 (2004), 193–202.

[19] Graesser, A.C., Singer, M., and Trabasso, T. Constructing Inferences during Narrative Text Comprehension. *Psychological Review 101*, 3 (1994), 371–95.

[20] Graesser, A.C. and Wiemer-Hastings, K. Situation models and concepts in story comprehension. In S.R. Goldman, A.C. Graesser and P. van den Broek, eds., *Narrative comprehension, causality, and coherence*. Mahwah, NJ, 1999, 77–92.

[21] Jiang, S., Fitzhugh, S.M., and Warschauer, M. Social Positioning and Performance in MOOCs. Proceedings of the Workshops held at Educational Data Mining 2014, co-located with 7th International Conference on Educational Data Mining (EDM 2014), (2014), 14.

[22] Joksimović, S., Dowell, N.M., Skrypnyk, O., et al. How do you connect? Analysis of Social Capital Accumulation in connectivist MOOCs. In *Proceedings from the 5th International Learning Analytics and Knowledge (LAK) Conference*. Poughkeepsie, New York, 2015.

[23] Kintsch, W. *Comprehension: A Paradigm for Cognition.* Cambridge University Press, Cambridge, U.K., 1998.

[24] Kovanović, V., Joksimović, S., Gašević, D., and Hatala, M. What is the Source of Social Capital? The Association between Social Network Position and Social Presence in Communities of Inquiry. *Proceedings of the Workshops held at Educational Data Mining 2014, (EDM 2014)*, (2014), 1–8.

[25] De Laat, M., Lally, V., Lipponen, L., and Simons, R.-J. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning 2*, 1 (2007), 87–103.

[26] McNamara, D.S., Graesser, A.C., McCarthy, P.M., and Cai, Z. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press., Cambridge, M.A., 2014.

[27] McNamara, D.S., Kintsch, E., Songer, N.B., and Kintsch, W. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and Instruction 14*, 1 (1996), 1–43.

[28] Nakagawa, S. and Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4*, 2 (2013), 133–142.

[29] Petty, R.E., Cacioppo, J.T., Strathman, A.J., and Priester, J.R. To Think or Not to Think: Exploring Two Routes to Persuasion. In T.C. Brock and M.C. Green, eds., *Persuasion: Psychological insights and perspectives, 2nd ed*. Sage Publications, Inc, Thousand Oaks, CA, US, 2005, 81–116.

[30] Pinheiro, J.C. and Bates, D.M. *Mixed-effects models in S and S-Plus.* Springer, 2000.

[31] Rabbany k., R., Takaffoli, M., and Zaïane, O.R. Social Network Analysis and Mining to Support the Assessment of On-line Student Participation. *SIGKDD Explor. Newsl. 13*, 2 (2012), 20–29.

[32] Rizzuto, T., LeDoux, J., and Hatala, J. It's not just what you know, it's who you know: Testing a model of the relative importance of social networks to academic performance. *Social Psychology of Education 12*, 2 (2009), 175–189.

[33] Romero, C., López, M.-I., Luna, J.-M., and Ventura, S. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education 68*, (2013), 458–472.

[34] Russo, T.C. and Koesten, J. Prestige, centrality, and learning: A social network analysis of an online class. *Communication Education 54*, 3 (2005), 254–261.

[35] Sanford, A.J. and Emmott, C. *Mind, Brain and Narrative.* Cambridge University Press, Cambridge, 2012.

[36] Scholand, A.J., Tausczik, Y.R., and Pennebaker, J.W. Assessing Group Interaction with Social Language Network Analysis. In S.-K. Chai, J.J. Salerno and P.L. Mabry, eds., *Advances in Social Computing*. Springer Berlin Heidelberg, 2010, 248–255.

[37] Siemens, G. and Baker, R.S. Learning analytics and educational data mining: towards communication and collaboration. *Proceedings of the 2nd international conference on learning analytics and knowledge*, ACM (2012), 252–254.

[38] Siemens, G. and Gašević, D. Special Issue on Learning and Knowledge Analytics. *Educ Technol Soc 15*, 3, 1–2.

[39] Singer, M. and O'Connell, G. Robust inference processes in expository text comprehension. *European Journal of Cognitive Psychology 15*, 4 (2003), 607–631.

[40] Snow, C.E. Reading for Understanding: Toward a Research and Development Program in Reading Comprehension. Rand Corporation, Santa Monica, CA, 2002.

[41] Walsh, T. and Bowen, W.G. Unlocking the Gates: How and Why Leading Universities Are Opening Up Access to Their Courses. Princeton University Press, Princeton, 2011.

[42] Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge ; New York, 1994.

[43] Wen, M., Yang, D., and Rose, C. Linguistic Reflections of Student Engagement in Massive Open Online Courses. In *Proceedings 14th International Conference on Web and Social Media*. AAAI, Ann Arbor, MI, 2014, 525–534.

[44] Yang, D., Wen, M., Kumar, A., Xing, E., and Rose, C. Towards an Integration of Text and Graph Clustering Methods as a Lens for Studying Social Interaction in MOOCs. *The International Review of Research in Open and Distributed Learning 15*, 5 (2014).

[45] Zwaan, R.A. and Radvansky, G.A. Situation models in language comprehension and memory. *Psychological Bulletin 123*, 2 (1998), 162–185.