Newcomer Integration in Online Knowledge Communities: Exploring the Role of Dialogic Textual Complexity

Nicolae Nistor, University of Munich, Germany/Walden University, USA, nic.nistor@lmu.de
Mihai Dascălu, University “Politehnica” Bucharest, Romania, mihai.dascalu@cs.pub.ro
Ștefan Trăușan-Matu, University “Politehnica” Bucharest, Romania, stefan.trausan@cs.pub.ro

Abstract: Using online knowledge communities (OKCs) as informal learning environments poses the question how likely these will integrate newcomers as peripheral participants. Previous research has identified surface characteristics of the OKC dialog as integrativity predictors. Yet, little is known about the role of dialogic textual complexity. This contribution proposes a comprehensive approach based on previously validated textual complexity indexes and applies it to predict OKC integrativity. The dialog analysis of $N = 14$ blogger communities with a total of 1937 participants identified three main components of textual complexity: dialog participation, structure and cohesion. From these, dialog cohesion was higher in integrative OKCs, thus significantly predicting OKC integrativity. This result adds to previous OKC research by uncovering the depth of OKC discourse. For educational practice, the study suggests a way of empowering learners by automatically assessing the integrativity of OKCs in which they may attempt to participate and access community knowledge.

Introduction

Learning in knowledge communities (KCs) is a significant topic of the Learning Sciences (Bereiter & Scardamalia, 2014), especially in the context of social web applications, that in recent decades dramatically extended the possibilities of communication and collaboration, giving birth to online KCs (OKCs). In this context, newcomer integration in KCs is a research topic of increasing importance (Eberle, Stegmann, & Fischer, 2014). In the relatively new domain of Learning Analytics, some attempts have been made to automatically monitor and predict newcomer integration in OKCs as a first step in an informal learning process (Nistor et al., 2015b). Such studies assume that community practice is reflected in, or even an organic component of, the community discourse (Wenger, 1998); consequently, newcomer integration in OKCs can be monitored and predicted by automated dialog analysis (Baker & Siemens, in press). Yet, little is known about the textual attributes of the OKC dialog, and specifically about its textual complexity. This contribution (1) proposes a comprehensive approach to textual complexity, and (2) applies this approach in the dialog analysis of blogger OKCs in order to (a) identify the ground dimensions of dialogic textual complexity and (b) predict how likely the OKCs generating this dialog will integrate newcomers. The results are expected to contribute to a deeper understanding of OKC discourse and practice, thus enabling the integration of OKCs in formal collaborative learning environments, and empowering learners by supporting more efficient knowledge sharing in OKCs (Nistor et al., 2015c).

Theoretical background

Dialogic textual complexity in knowledge communities

KC research (Bereiter & Scardamalia, 2014; Wenger, 1998) describes discourse as connecting community practice, participants, and their knowledge about the practice. As such, discourse appears to be the essence of experience-based knowledge construction and sharing in KCs, which comprises the interplay between participation and reification. While participating, KC members acquire experience and reify it, thus developing cultural artifacts. Artifacts, in turn, support participation at a higher level. KC discourse includes both processes. At the same time, KC discourse includes the negotiation of meaning and the collaborative construction of shared knowledge, thus supporting the transition from distributed to shared knowledge (Wenger, 1998). Regarding the KC dialog as the spoken dimension of discourse, from a linguistic perspective, the following dialog characteristics can be observed:

Initiation of, and participation in, the community dialog. KC members participating in the community practice and encountering various problems will initiate and participate in a dialog aimed at coordinating activities, negotiating meaning and constructing shared knowledge. As a measurable result, a number of words, utterances and discussion threads will be produced (Dascălu, 2014).

Dialog structure. Connecting practice and knowledge about practice requires at linguistic level a specific vocabulary. Therefore, the produced words and utterances will build upon discourse structures consisting of main notions and connectors of these notions, such as cue phrases, co-references, speech acts, adjacency pairs, and rhetorical schemas (Jurafsky & Martin, 2009).

Dialog cohesion. Carrying out community practice and collaboratively constructing knowledge over longer periods of time (Wenger, 1998) requires cohesive dialog. Both the threads of utterances produced around single moments of practice, and different dialogs emerging in time from the community practice will be cohesive. This implies local cohesion (i.e., the dialog will be cohesive in itself, as shown by the dialog structure dimension), as well as global cohesion (i.e., between discussion threads and dialogs within the community practice) (McNamara, Graesser, & Louwerse, 2012).

These three dimensions of the collaborative KC dialog were named here in the order of increasing *textual complexity*. Initiation of, and participation in, the community dialog is supposed to reflect the surface, dialog connectedness its structure, and dialog cohesion the depth of KC discourse.

The assessment of collaborative dialog appears to be a productive method of quantitative KC research. After several decades of qualitative research, quantitative methods become more visible in the empirical approaches to knowledge communities. In Learning Analytics, particularly in Discourse Analytics, methods including social network analysis, clustering, and factor analysis were applied to identify socio-cognitive structures and predict learning in technology-based environments (Baker & Siemens, in press). Assuming that community discourse is tightly connected with socio-cognitive structures, practice and learning (Wenger, 1998), Nistor et al. (2015c) use *ReaderBench,* an automated dialog analysis tool based on Bakhtin's (1981) dialogism and on the polyphonic model of discourse (Trăușan-Matu, 2010), to assess the quality of the collaborative text-based dialog in OKCs. These dimensions were correlated with participants' expertise and centrality in the KC (Nistor et al., 2015c). Yet, little has been done to quantitatively assess the textual complexity of the KC dialog and to explore its relationship with the KC structure and processes.

Assessing textual complexity

Every dialog must be understood by the individuals involved in it. Building on this basic assumption, several categories of complexity indexes have been developed and validated ranging from surface factors to more in-depth dialog characteristics such as syntax and semantics (Dascălu et al., 2013; 2015). Firstly, the *surface* category is based on statistics of individual analysis elements (words, phrases, paragraphs) derived from classic readability formulas, as well as Page's (1966) grading technique for automated scoring covering basic structure complexity (e.g., number of words, of commas, of sentences, word length, average number of syllables per word, or of words per sentence) and word/character entropy. Secondly, the *syntax* category changes the focus to statistics applied per different parts of speech (e.g., nouns, prepositions, pronouns), as well as the complexity of the parsing tree in terms of its maximum depth and its size of the parsing structure (Dascălu, 2014). Thirdly, the *semantics and discourse analysis* category is based on cohesion graphs (Dascălu, 2014), the strength of the links between different analysis elements (intra-contribution between sentences, inter-contributions to reflect a cohesive flow), as well as named entities identification and specific discourse connectives covering coordination, subordination, conditions, contrasts or sentence linking.

Methodology

Research questions

Given the comprehensive collection of indexes named above, it is still unclear which of these are representative for the notion of OKC dialog complexity and predictive of newcomer integration. Therefore, the following research questions are examined: (1) Which independent dimensions of textual complexity can be identified in the OKC dialog? (2) Which of these can predict newcomer integration?

Data collection

The analysis was conducted on the Internet, in blogger communities publicly available on the blogspot.com and wordpress.com platforms. In a prior study, the researchers had attempted to initiate discussions in several blog communities, observing that some communities were more open to dialog and more likely to integrate newcomers than others, consequently the former were regarded as integrative ($n = 3$), the latter as non-integrative ($n = 11$) OKCs. After these $N = 14$ blogger communities with a total of 1937 participants were chosen, the entire community discourse produced within a year (ending with the day of the intervention that should have initiated new conversation threads) was downloaded and automatically analyzed. No personal data of the participants were collected.

Data analysis

The textual complexity analysis tool provides a wide range of indexes out of which 89 dialog indexes were used: 18 *surface indexes* (e.g., average sentence length in characters, average number of commas per sentence, average number and standard deviation of words in sentence, word entropy), 25 *syntactic indexes* (e.g., average number of nouns/ pronouns/ prepositions/ adjectives/ adverbs/ verbs per sentence/ paragraph, average parsing tree depth, average parsing tree size), and 46 *semantic indexes* (e.g., average number of named entities per paragraph, average number of connector type per paragraph, average paragraph/contribution score, average sentence-paragraph/inter-paragraph/intra-paragraph/paragraph adjacency/ transition cohesion in terms of Wu-Palmer, Latent Semantic Analysis and Latent Dirichlet Allocation semantic distances) (see overview in Dascălu, 2014). Because these were strongly correlated with each other, a main component analysis was performed to determine the independent dimensions of textual complexity. Afterwards, two subsamples of $n = 3$ integrative (with a total of 270 participants) and $n = 4$ non-integrative (460 participants) blogger OKCs with the same discussion topic (politics and economy) were chosen, and compared with respect to the main components of the textual complexity. Subsequently, a discriminant analysis was performed in order to assess the adequacy of the classification.

Findings

Principal component analysis

The 89 selected textual complexity indexes were reduced to three factors accounting for 91% of the total variance. Three dimensions resulted after eliminating 49 indexes with eigenvalues smaller than 1 and cross loadings over .4 on more than one factor, performing varimax rotation and saving the components according to the Anderson-Rubin method. Factor 1, interpreted as *Dialog Structure,* includes 28 factors, classified as discourse connectors (e.g., number of conjuncts per paragraph), syntactic indexes (e.g., number of verbs per paragraph), and indexes of basic structure and word diversity (e.g., number of words per paragraph). Factor 2, interpreted as *Dialog Cohesion,* includes 16 indexes, classified as local cohesion indexes (e.g., sentence-paragraph cohesion) and global cohesion indexes (e.g., inter-contribution and transition cohesion). Factor 3, interpreted as *Dialog Participation,* includes 6 indexes (e.g., number of initiated discussion threads).

Predicting newcomer integration

Dialog Cohesion was higher in integrative OKCs (z values $M = .21$, $SD = 1.11$) than in non-integrative OKCs ($M = -.02$, $SD = .88$), and the difference was statistically significant with $F(1, 728) = 9.924$, $p = .002$. Dialog Participation was also higher in integrative OKCs ($M = .11$, $SD = 2.49$) than in non-integrative OKCs ($M = -.07$, $SD = .47$), however this difference failed to reach statistical significance with $F(1, 728) = 2.219$, $p = .14$. Discourse Structure was roughly the same (.24-.25) in both subsamples. A discriminant analysis confirmed Dialog Cohesion as a significant predictor of OKC integrativity, with Wilks $\lambda = .987$, $p = .002$.

Discussion

Aiming to explore the role of dialogic textual complexity, and to predict how likely OKCs integrate newcomers, this study analyzed the dialog produced in blogger OKCs and identified three complexity dimensions: Dialog Participation, Dialog Structure, and Dialog Cohesion. These synthesize a large number of indexes from previous research literature describing textual complexity (Dascălu, 2014).

From these, the surface factor Dialog Participation was somewhat higher in integrative OKCs, which is in line with the differences reported by Nistor and colleagues (2015b). As a possible explanation, integrative OKCs are more open to dialog, and more active, more "talkative", which opens newcomers more opportunities to participate and access specific OKC knowledge (Eberle et al., 2014).

More interestingly, the Dialog Cohesion, the most in-depth complexity factor, was significantly higher in integrative than in non-integrative OKCs, thus predicting OKC integrativity. This result adds to previous OKC research, and suggests that integrativity may be an aspect of the previously established community discourse. Nistor et al. (2015a) measured dialog quality as the percentage of social knowledge building per utterance, a participation centered, thus a surface indicator in terms of textual complexity. This study goes beyond the surface and uncovers the depth of OKC dialog. In non-integrative OKCs, the socio-cognitive component (mainly knowledge construction and sharing between active members) may dominate the community discourse, while in integrative OKCs this component may be balanced with the social component (member identity negotiation and development, new member recruitment, monitoring and training – Eberle et al., 2014). Therefore, the more balanced and complex nature of practice in integrative OKCs may result in a more complex discourse and, respectively, dialog.

Conclusions

For the practice of computer-supported collaborative learning, this study suggests a way of empowering informal learners who attempt to access OKC knowledge. Automated analysis tools can indicate their chances of success with a particular OKC, and thus enable them to be more efficient in their search of a responsive community. Thus, OKCs can also be integrated in formal learning environments (Nistor et al., 2015c).

For research and development in the Learning Sciences, this study contributes to the relatively new domain of Learning Analytics with more accurate procedures and tools for monitoring and predicting learning behaviors (Baker & Siemens, in press) and, more generally, with a deeper understanding of OKC discourse and practice. Further research will propose and evaluate formal learning scenarios based on OKC integration and participation.

References

Baker, R., & Siemens, G. (in press). Educational data mining and learning analytics. *Cambridge handbook of the learning sciences* (2nd edition). Cambridge, UK: Cambridge University Press. http://www.columbia.edu/~rsb2162/BakerSiemensHandbook2013.pdf

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. London: The University of Texas Press.

Bereiter, C. & Scardamalia, M. (2014). Knowledge building and knowledge creation: One concept, two hills to climb. In S. C. Tan, H. J. So, & J. Yeo (Eds.), *Knowledge creation in education* (pp. 35-52). New York: Springer.

Dascălu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating. Studies in Computational Intelligence* (Vol. 534). New York: Springer.

Dascălu, M., Stavarache, L. L., Dessus, P., Trăușan-Matu, S., McNamara, D. S., & Bianco, M. (2015). Predicting comprehension from students' summaries. In *17th International Conference on Artificial Intelligence in Education (AIED 2015)* (pp. 95–104). New York: Springer.

Dascălu, M., Trăușan-Matu, S., & Dessus, P. (2013). Cohesion-based analysis of CSCL conversations: Holistic and individual perspectives. In N. Rummel, M. Kapur, M. Nathan, & S. Puntambekar (Eds.), *10th International Conference on Computer-Supported Collaborative Learning 2013* (pp. 145–152). Madison, USA: International Society of the Learning Sciences.

Eberle, J., Stegmann, K., & Fischer, F. (2014). Legitimate peripheral participation in communities of practice: Participation support structures for newcomers in faculty student councils. *Journal of the Learning Sciences, 23*(2), 216-244.

Jurafsky, D., & Martin, J. H. (2009). *An introduction to natural language processing. Computational linguistics, and speech recognition* (2nd ed.). London: Pearson Prentice Hall.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R&L Education.

Nistor, N., Dascălu, M., Stavarache, L. L., Serafin, Y., & Trăușan-Matu, Ș. (2015a). Informal learning in online knowledge communities: Predicting community response to visitor inquiries. In G. Conole, T. Klobucar, C. Rensing, J. Konert, & E. Lavoué (Eds.), *Design for teaching and learning in a networked world. 10th European conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15-18, 2015*, Proceedings (pp. 447-452). New York: Springer.

Nistor, N., Dascălu, M., Stavarache, L. L., Tarnai, C., & Trăușan-Matu, Ș. (2015b). Predicting newcomer integration in online knowledge communities by automated dialog analysis. In Y. Li, M. Chang, M. Kravcik, E. Popescu, R. Huang, Kinshuk, & N. S. Chen (Eds.), *State-of-the art and future directions of smart learning* (pp. 13-17). New York: Springer.

Nistor, N., Trăușan-Matu, Ș., Dascălu, M., Duttweiler, H., Chiru, C., Baltes, B., & Smeaton, G. (2015c). Finding open-ended learning environments on the Internet: Automated dialogue assessment in academic virtual communities of practice. *Computers in Human Behavior, 47*(1), 119-127.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Trăușan-Matu, Ș. (2010). The polyphonic model of hybrid and collaborative learning. In F. Wang, L. J. Fong, & R. C. Kwan (Eds.), *Handbook of research on hybrid learning models: Advanced tools, technologies, and applications* (pp. 466–486). Hershey, NY: Information Science Publishing.

Wenger, E. (1998). *Communities of practice. Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Acknowledgments

This work has been partially funded by the EC H2020 project RAGE (Realising and Applied Gaming Eco-System, http://www.rageproject.eu, grant agreement No 644187) and the Sectorial Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of European Funds according to the Financial Agreement POSDRU/159/1.5/S/134398.