ICO ISS 2021

ICO International Spring School 2021

Presentation type: Round Table **Presenter:** Dennis A. Rivera, M.Ed.

A corpus linguistic approach to analysing cognitive engagement in the forums of Massive Open Online Courses (MOOCs).

Dennis A. Rivera[1], Mariane Frenay[1]
[1] Psychological Sciences Research Institute (IPSY), université catholique de Louvain, Belgium

ARTICLE INFO

Keywords:
Cognitive engagement
MOOCs
Discussion forums
SOLO Taxonomy

ABSTRACT

Discussion forums of Massive Open Online Courses (MOOCs) are a valuable resource for instructors to help engage learners cognitively. Recognising different levels of cognitive engagement in MOOC forums, however, presents several challenges to MOOC instructors. This study proposes a corpus analysis approach to identifying linguistic categories that have been consistently correlated with different levels of cognitive processes. These categories are used in the context of a learning taxonomy to portray the level of language complexity and the organisation of thoughts in MOOC forum messages. This approach can better inform MOOC instructors about the cognitive engagement of their learners in MOOC discussion forums.

1. Introduction

Discussion forums are the interactive component of Massive Open Online Courses (MOOCs). These forums are designed to engage learners from a variety of backgrounds in different types of discussions. The highest value of MOOC forums lies in their potential to promote understanding through the application of several cognitive processes (Stephens-Martinez et al., 2014). However, the identification of cognitive processes poses a challenge to MOOC instructors who lack the means to navigate through the messages and identify those where their expertise is most needed. In this study, we propose a corpus analysis approach, based on a learning taxonomy, to study the cognitive engagement of MOOC learners in discussion forums. This approach aims at analysing manifest measures of the levels of language complexity and the structure of MOOC forum messages. The results from the analysis are presented as levels of cognitive processes that MOOC instructors are familiar with. This can inform them about the quality of the discussions in their forums.

2. Literature Review

Cognitive engagement has been defined as the psychological investment that learners put in their learning (Newmann et al., 1992). In MOOC forums, this engagement is analysed through the messages posted by the learners because the level of language complexity and the organisation of ideas can portray different levels of cognitive processes (Carroll, 2007; Van Swol et al., 2016). MOOC forum messages can include a simple description of facts or a complex interpretation of the material studied. Previous studies have used learning taxonomies to examine the cognitive engagement of MOOC learners (e.g.: O'Riordan et al., 2016; Wong et al., 2015). However, such examinations usually require a manual content analysis approach. This approach for studying recorded content can be useful to identify latent variables (Krippendorff, 2004). However, its application becomes impractical in MOOC forums due to the large number of messages (Breslow et al., 2013).

2.1 Corpus analysis approach

A corpus analysis approach can address the issue of the large number of messages in MOOC forums. This approach facilitates the analysis of massive amounts of texts by recognising words and finding patterns of grammatical structure within the texts. The results from corpus analysis are usually displayed in categories such as the total number of words, lists of word frequency, and part-of-speech taggers. These categories can be useful in MOOC forums to understand how messages are structured;

however, on their own, corpus analysis categories carry little meaning to MOOC instructors. In spite of the limitations of corpus analysis, previous research has found several categories consistently correlated with different levels of cognition. Amongst these categories we have the number of words and sentences, the type of words used, and the presence of cohesive devices (Joksimovic et al., 2014; Kovanović, et al., 2019; Moore et al., 2019; O'Riordan et al. 2020). The number of words and sentences are considered a manifest measure of the complexity of MOOC forum messages (Schrire, 2006). Although forum messages with few words and sentences do not necessarily imply low levels of cognition, the length of a message is a useful quantitative metric to explore its structure.

Contrary to the quantitative nature of the length of a message, cohesive devices within the messages, such as transition words, can provide a qualitative measure of the connection of ideas expressed in MOOC forum messages. Transition words indicate a logical or semantic relation between units of discourse (Halliday & Hasan, 1976); hence, they can be used to represent the addition, clarification, comparison, or contrast of ideas. Similarly, the type of words used in the messages can also provide a qualitative measure of the level of language complexity. Previous studies have found that a high percentage of dictionary words in a MOOC forum message can provide indicators of high levels of cognitive engagement (Moore et al., 2019).

To identify nuances within different levels of cognitive processes, it is important not only to identify dictionary words but also to differentiate academic vocabulary from other type of words. Oztok et al. (2013) have previously used Coxhead's (2000) Academic Word List to explore the use of academic language in online courses. Although this list is useful to recognise words that appear frequently in academic texts, Paquot's (2010) Academic Keyword List (AKL) provides a list of 930 words (including transition words) not based on frequency but on keyness. This textual feature helps the AKL to identify words that do not only appear frequently in academic texts but also have been shown to play an essential role in structuring academic prose. Thus, the AKL can be used to identify academic vocabulary and transition words in the messages written by MOOC learners. The number of words and sentences in a MOOC forum message, together with the use of transition and vocabulary words, can display different levels of language complexity in MOOC forum messages and the way how learners have organised their thoughts. However, these measures alone do not inform MOOC instructors where their attention is most needed. They should be used in the context of a learning taxonomy to be meaningful to MOOC instructors.

2.2 The SOLO taxonomy

The Structure of Observed Learning Outcomes (SOLO) is a learning taxonomy that was developed through the analysis of the structure of higher education students' written responses (Hattie & Brown, 2004). This hierarchical model of five levels categorised in terms of their quality lie along a continuum from simple to complex that reflect low and high levels of cognitive processes (Biggs & Collis, 1989). The lowest levels of the SOLO taxonomy deal with simple knowledge acquisition and surface understanding. In these levels, learners structure information in an oversimplified way. Conversely, as learners move to high levels, they make mental connections that allow them to analyse, evaluate, and relate information. Quantitatively, the SOLO taxonomy considers the number of concepts or ideas expressed by the learners. Qualitatively, this taxonomy analyses whether such concepts or ideas are connected or extended. This connection of ideas portrays whether the learners can "see the wood or just isolated trees" (Biggs & Tang, 2011, p. 89). We propose to use the corpus analysis categories that have been consistently correlated with different levels of cognition to match the descriptors of the levels in this taxonomy.

The first SOLO level is the pre-structural level. At this level, learners use irrelevant and unorganised pieces of information, essentially void to actual content with no relation to the topic. Cognition does not exist in this level. The second level, called unistructural, is the first level of cognition. At this level, learners deal with one single aspect of the content, perform simple tasks (name, count, paraphrase, identify, etc.), use basic terminology, and remember things but there is no connection of facts or ideas. The multi-structural is the third SOLO level. At this level, learners deal with several unconnected aspects of the content. Although learners could use cohesive devices, these are used as a form of "knowledge-telling" (Bereiter & Scardamalia, 1987); this is, expressing ideas retrieved from long-term memory without analysing or evaluating their value.

The last two levels of the SOLO taxonomy, called relational and extended abstract, are considered high levels of cognitive processes (Hattie & Purdie, 1998). At the relational level, learners understand relations between several aspects of the content. This allows them to connect ideas and compare, analyse, or explain them in terms of cause and effect. At the last level of the SOLO taxonomy, namely extended abstract, learners go beyond what is given and generalise, theorise, or transfer their ideas to new areas. Based on the description of these five levels, we put together the results of corpus analysis categories to inform MOOC instructors the extent to which their learners are engaging cognitively in the forums.

3. Methods

This study involved 9.172 forum messages posted by the learners in 3 edX MOOCs throughout different editions of the courses from 2015 to 2019. All MOOCs were self-paced courses in the area of social sciences developed by the Université catholique de Louvain (UCLouvain). Our selection criteria for the MOOCs was based on the access we received to the course data from the MOOC developers. The MOOCs investigated in this study, together with the number of forums and forum messages in each course are displayed in Table 1 (see Appendix 1).

We used corpus analysis techniques to extract the number of words and sentences from the messages posted by the learners in the forums. We tagged the words according to their part of speech (noun, verb, adverb, etc.) and used Paquot's (2010) Academic Keyword List (AKL) to identify academic vocabulary and transition words in the messages. Tagging the words helped to avoid erros when identifying homographs (e.g.: *study* (verb) and *study* (noun)). We calculated a measure of lexical sophistication based on the percentage of academic words within the forum messages. Additionally, we calculated the quartiles of the distribution of the messages to form groups that divide our dataset.

Based on the corpus data, we constructed a characterisation of MOOC forum messages that reflected the levels of cognitive processes in the SOLO taxonomy. Quantitatively, this characterisaton included the number of sentences in a message. Qualitatively, we considered the lexical sophistication of the messages and the presence of transition words. Thus, we describe the SOLO pre-structural level as a message composed of a single sentence, with low lexical sophistication (within the first quartile), and not including any connectors. Conversely, the SOLO relational level is described as a message containing three or more sentences, with high lexical sophistication (within the fourth quartile), and including one or more transition words. The complete characterisation can be seen in Table 2 (see Appendix 2).

The highest cognitive level of the SOLO taxonomy, namely extended abstract, is not captured using corpus analysis because this level is highly dependent on the context. Therefore, a manual analysis of the messages is still necessary to reveal latent content beyond the structure of the messages (De Wever et al., 2005). Despite this limitation, our approach can better inform MOOC instructors about the levels of cognitive process that their learners use in different discussion forums.

References

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Erlbaum.

Biggs, J., & Collis, K. (1989). Towards a model of school-based curriculum development and assessment using the SOLO taxonomy. *Australian Journal of Education, 33*(2), 151–163. doi:10.1177/168781408903300205

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th Ed.). Milton Keynes, UK: Open University Press.

Breslow, L., Pritchard, D., DeBoer, J., Stump, G., Ho, A., & Seaton, D. (2013). Studying Learning in the Worldwide Classroom Research into edX's First MOOC. *Research & Practice in Assessment, 8*, 13-25. Retrieved from https://eric.ed.gov/?id=ej1062850

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: a quantitative analysis. Assessing Writing, 12(3), 213-227

Coxhead, A. (2000). A new academic word list. *TES*OL Quarterly, 34, 213-238.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2005). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education, 46*(1), 6-28.

Halliday, M., & Hasan, R. (1976). *Cohesion in English.* New York: Longman

Hattie, J., & Brown, G. (2004). *Cognitive processes in asTTle: The SOLO taxonomy.* University of Auckland. Wellington, N.Z.: Ministry of Education.

Hattie, J., & Purdie, N. (1998). The SOLO model: Addressing fundamental measurement issues. In B. Dart, & G. Boulton-Lewis, *Teaching and learning in higher education* (pp. 145-176). Victoria, Australia: ACER Press.

Joksimovic, S., Gasevic, D., Kovanov, V., Adesope, O., & Hatala, M. (2014). Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *The Internet and Higher Education, 22*, 1-10. doi:10.1016/j.iheduc.2014.03.001

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2019). Towards automated content analysis of discussion transcripts: a cognitive presence case. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)* (pp. 15–24). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2883851.2883950

Krippendorff, K. (2004). Content analysis: an introduction to its methodology. Thousand Oaks, California: Sage Publications Inc.

Moore, R., Oliver, K., & Wang, C. (2019). Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments, 27*(5-6), 655-669. doi:10.1080/10494820.2019.1610453

Newmann, F., Wehlage, G., & Lamborn, S. (1992). The significance and sources of student engagement. In F. Newmann, *Student engagement and achievement in American secondary schools* (pp. 11-39). New York, NY: Teachers College Press.

O'Riordan, T., Millard, D., & Schulz, J. (2016). How should we measure online learning activity? *Research in Learning Technology, 24*, 1-28. doi:10.3402/rlt.v24.30088

O'Riordan, T., Millard, D., & Schulz, J. (2020). Is critical thinking happening? Testing content analysis schemes applied to MOOC discussion forums. *Comput Appl Eng Educ*, 1– 20. doi:10.1002/cae.22314

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis.* London & New-York: Continuum.

Oztok, M., Zingaro, D., Brett, C., & Hewitt, J. (2013). Exploring asynchronous and synchronous tool use in online courses. *Computers & Education, 60*(1), 87-94. doi:10.1016/j.compedu.2012.08.007

Schrire, S. (2006). Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis. *Computers & Education, 46*(1), 49-70. doi:10.1016/j.compedu.2005.04.006

Stephens-Martinez, K., Hearst, M., & Fox, A. (2014). Monitoring MOOCs: Which information sources do instructors value? *Proceedings of the 1st ACM conference on learning @ scale* (pp. 79–88). New York, NY, USA: ACM. doi:10.1145/2556325.2566246

Van Swol, L., Prahl, A., Kolb, M., Lewis, E., & Carlson, C. (2016). The language of extremity: The language of extreme members and how the presence of extremity affects group discussion. *Journal of Language and Social Psychology, 35*(6), 603–627. doi:10.1177/0261927X16629788

Wong, J., Pursel, B., Divinsky, A., & Jansen, B. (2015). Analyzing MOOC discussion forum messages to identify cognitive learning information exchanges. *Proc. Assoc. Info. Sci. Tech., 52*, 1-10. doi:10.1002/pra2.2015.145052010023

Appendix 1

Table 1
MOOCs and number of forums and forum messages throughout different editions in several years.

Massive Open Online Courses (MOOCs)	Years running	Editions	Forums	Messages
MOOC 1: Communicating CSR	2015 – 2018	5	44	4.913
MOOC 2: Oriental Beliefs	2016 – 2019	3	24	2.423
MOOC 3: Discover political science	2018 – 2019	2	23	1.836

Note: The number of forums represents the total forums analysed in this study throughout all editions of the MOOCs. This number may vary in some editions of the courses, especially in editions without forums created for live sessions.

Appendix 2

Table 2
Linguistic features of MOOC forum messages that relate to the levels of the SOLO taxonomy.

SOLO Levels of cognitive processes	Characterisation			
Pre-structural				
The message misses the point, does not show any kind of understanding or uses irrelevant information to cover the lack of knowledge. E.g.: *"I don't know anything about this."*	Sentences: Lex. Sophist.: Trans. words:	1 Q_1 No		
Uni-structural				
The message deals with a single aspect of the content but misses other important parts. Understanding is nominal. E.g.: *"I only know that… I would like to learn more."*	Sentences: Lex. Sophist.: Trans. words:	2 or + Q_1 No	2 or - Q_2 No	
Multi-structural				
The message deals with several aspects of the content but they are not connected. This collection of unstructured facts can include transition words as a form of knowledge-telling. E.g.: *"Therefore, this is…"*	Sentences: Lex. Sophist.: Trans. words:	3 or + Q_2 No	1 or + Q_3 Yes	3 or - Q_4 Yes
Relational				
The message deals with several aspects of the content and includes explanations (in terms of cause and effect), comparisons, and analyses that connects the aspects together. E.g.: *"On the one hand, this…"*	Sentences: Lex. Sophist.: Trans. words:	3 or + Q_4 Yes		

Note: Q_1, Q_2, Q_3, and Q_4 represent the quartiles of the lexical sophistication (Lex. Sophist.)