

# Automating the Analysis of Collaborative Discourse: Identifying Idea Clusters

Nobuko Fujita, Christopher Teplovs, University of Toronto, 252 Bloor Street W., Toronto, ON, Canada, M5S 1V6,  
nobuko.fujita@gmail.com, christopher.teplovs@gmail.com

**Abstract:** This poster explores CSCL practices relating to the use of a tool that employs information visualization techniques and large-scale text processing and analysis to complement qualitative analysis of collaborative discourse. Results from latent semantic analysis and qualitative analysis of online discussion transcripts are compared. Findings suggest that such tools that automate analyses of large text-based data sets can offer CSCL researchers a quantitative and unbiased way of identifying a subset of data to study in depth.

## Introduction

A growing body of literature emphasizes the use of mixed methods to investigate educational processes (e.g., Johnson & Onwuegbuzie, 2004). In computer-supported collaborative learning (CSCL) research, mixed methods may be needed to understand collaborative discourse that is the primary mechanism for learning in these environments (Hmelo-Silver, 2003). A qualitative approach yields deep insights into what is happening in small segments of discourse, but it is time consuming and not practical for examining a large corpus of data (Sawyer, 2006). Yet relatively little attention has been paid to CSCL practices relating to the use of technological tools that automate analyses of large text-based data sets. Tools that employ information visualization techniques and large-scale text processing and analysis seem to hold considerable potential for advancing learning science practices. In this paper we explore the use of one such tool, the Knowledge Space Visualizer (KSV) that combines latent semantic analysis (LSA) and graph-based information visualization and network analysis (Teplovs, 2008).

The data comprise one of the iterations of a larger design-based research project that investigated how to foster the development of progressive discourse in three online graduate course contexts (Fujita, 2009). Progressive discourse is the process through which participants share, question, and revise their ideas to deepen their understanding and build knowledge (Bereiter, 2002). To answer questions about what kinds of instructional scaffolding are most effective in fostering progressive discourse for knowledge building, it is important to be able to detect when a group of students construct a new understanding in a particular computer-mediated communication (CMC) transcript. Whereas we do not propose LSA as a replacement for qualitative analysis, we suggest it might offer CSCL researchers an alternative approach to analyzing data from one iteration of a design-based research study to inform the design of interventions in a subsequent iteration. To this end, the KSV was used to investigate the following research question: “What are the major themes or ideas clusters found in the student discourse in each course discussion view?”

## Methods

This study examines one iteration of a design-based research (Collins, Joseph, & Bielaczyc, 2004) project that investigated how to foster online progressive discourse for knowledge building in three graduate education courses. The participants in this iteration were 17 students in a 13-week online graduate course surveying educational applications of computer-mediated communication. A tenure-stream faculty member taught the course entirely online using web-based Knowledge Forum, a software environment specifically designed to support knowledge building. As Scardamalia & Bereiter (2003) explain, “the basic units in Knowledge Forum are ideas, represented as *Notes*. The basic workspace for developing, sharing, organizing, and creating multiple representations of ideas is a *View*” (p. 24). Knowledge Forum differs from typical asynchronous computer conferencing systems by having advanced features such as scaffolds, co-authored notes, annotations, and “rise-above” capabilities to support knowledge building processes.

This paper focuses on analysis of 1010 notes contributed to 11 Knowledge Forum views. Each view represented one week of discussion. The first week and last week views were omitted. The view for week 1 was used to introduce the students to the course, to the database, and to each other; the week 13 view was used for course evaluation.

For qualitative coding, the text of notes in Knowledge Forum was exported to rich text files (.rtf) and imported into the NVivo qualitative data analysis software. Each note was read several times in the context of the discussion thread in which the author posted the note, then coded through a process of constant comparison (Merriam, 1998; Strauss & Corbin, 1998). Subcategories were created as more possibilities were found under each emergent theme or idea cluster category. When notes contained more than one theme, they were assigned multiple codes.

LSA is a statistical technique used to extract the deep meaning of patterns of words in specific contexts of use. The technique is performed by applying methods from linear algebra (matrix decomposition and dimension reduction) to matrices that represent usage patterns of terms in documents (Deerwester et al., 1990; Landauer et al., 1998). Documents can be projected into the resulting reduced- but still high-dimensional semantic space and semantic similarity can be determined by calculating the cosine between the vectors for any two documents. Whereas cosine values lack the statistical properties of correlation coefficients they can be interpreted in a similar fashion, with identical documents having inter-vector cosines of 1.0, unrelated documents having inter-vector cosines of 0 and somewhat related documents falling somewhere between 0 and 1.

The resulting similarity matrix for a group of documents – in this case a group of notes from a view – can be used to generate a graph. In the complete projection of such a similarity matrix, all documents are linked to each other if only at minimal cosine values. A more reasonable approach, however, is to select a threshold value for cosine similarity, above which documents are considered linked (and below which they remain unlinked). The selection of the threshold value is empirically determined and detailed in Teplovs (2009). The appropriate value for the analysis reported here was determined to be 0.6.

## Results and Discussion

This study used the KSV to investigate the question, “What are the major themes or ideas found in the student discourse?” Of the 1010 notes in week 2 to week 12 views, were members of idea clusters. A grouping of notes was considered to be a valid cluster if it was represented by at least 3 notes within any view. The KSV-based analysis suggests that there are 43 unique idea clusters in the notes.

For example, during week 3 the students discussed various types of CMC environments, their purposes and uses. This week was chosen for closer examination because it was the first week in which pairs of students not the instructor lead the discussion. It provided an opportunity to study how students might engage in progressive discourse as they considered synchronous environments such as chat, moo’s, and mud’s, and compared them to asynchronous environments such as conferencing, listservs, and bulletin boards. The most frequently occurring themes in week 3 identified through semantic analysis were “knowledge building,” “skype,” “moo,” “discussion,” “scaffolds,” “software,” “communication,” and “available.”

A comparison of idea clusters that were identified through LSA and through qualitative content analysis in week 3 is shown in Table 1.

Table 1: Frequencies of notes in idea clusters identified through LSA and qualitative analysis in week 3

LSA			Qualitative Analysis	
Idea Cluster	Notes	Density	Idea Cluster	Notes
Knowledge, building	13	0.230769	Chat scheduling	29
Skype, chat, MSN	12	0.212121	Emotions	19
Moos	11	0.490909	Knowledge building	17
Discussion, synchronous	7	0.476190	Moos	17
Software	5	1.00000	Technical issues	15
Scaffolds	5	0.476190	Scaffolds	9
Communication, synchronous	4	0.833333	Synchronous chat	7
Available	4	0.666667	Skype	7

The idea clusters identified through LSA are remarkably similar to categories that emerged through qualitative coding of the student notes. This lends face validity to the findings from LSA, and suggests that LSA might offer CSCL researchers a quantitative and unbiased way of identifying a subset of data to study in depth using mixed methods (Chi, 1997).

Some differences, presented more fully in the poster, are noteworthy and highlight some of the problems associated with fully automating the analytic technique. However, the LSA-based analysis provides some additional

metrics that help identify potentially problematic classifications. Table 1 includes values for the network density for each of the clusters. Network density is the quotient of links (edges) and maximum number of links for a given cluster. In practical terms, the network density of a semantic cluster is a measure of the idea diversity within that cluster. Thus, a semantic cluster with a high network density consists of notes that are highly homogeneous in terms of content. Semantic clusters with low network density are characterized by sub-clusters that may be about highly disparate topics but share a higher-level commonality. In the case of clusters from week 3, we see that the two largest clusters (“Knowledge building” and “Skype”) have relatively low density compared to the other clusters. This suggests these clusters may benefit from being described more fully using additional key terms, or that there may be confounding, difficult-to-detect themes that overrode obvious content-based commonalities.

## References

- Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6(3), 217-315.
- Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research : Theoretical and methodological issues. *The Journal of the Learning Sciences*, 13(1), 15-42.
- Creswell, J. W. & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice*, 39(3), 124-130.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.,K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41. 391:407.
- Fujita, N. (2009). *Group processes supporting the development of progressive discourse in online graduate courses*. Unpublished doctoral dissertation, Ontario Institute for Studies in Education of the University of Toronto, Toronto, ON.
- Hmelo-Silver, C. E. (2003). Analyzing collaborative knowledge construction: Multiple methods for integrated understanding. *Computers and Education*, 41(4), 397-420.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26.
- Landauer, T. K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Merriam, S.B. (1998). *Qualitative research and case study applications in education* (2<sup>nd</sup> ed.) San Francisco, CA: Jossey-Bass.
- Sawyer, R. K. (2006). Analyzing collaborative discourse. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 187-204). New York: Cambridge University Press.
- Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.
- Scardamalia, M., & Bereiter, C. (2003). Knowledge Building. In *Encyclopedia of Education* (2nd ed.). New York: Macmillan Reference, USA.
- Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2<sup>nd</sup> ed.), Thousand Oaks, CA: Sage.
- Teplovs, C. (2008). The Knowledge Space Visualizer: A tool for visualizing online discourse. In G. Kanselaar, V. Jonker, P. A. Kirschner & F. Prins (Eds.), *Proceedings of the International Conference of the Learning Sciences 2008: Cre8 a learning world*. Utrecht, NL: International Society of the Learning Sciences.
- Teplovs, C. (2009). *Visualizing idea diversity*. Unpublished doctoral dissertation, Ontario Institute for Studies in Education of the University of Toronto, Toronto, ON.

## Acknowledgments

This research was supported by grants from the Social Sciences and Humanities Council of Canada and the Canadian Foundation for Innovation awarded to Clare Brett.