| | |
|---|---|
| Title | Understanding idea flow: Applying learning analytics in discourse |
| Author(s) | Alwyn Vwen Yen Lee and Seng Chee Tan |
| Source | *Learning: Research and Practice, 3*(1), 12-29 |
| Published by | Taylor & Francis (Routledge) |



Citation: Lee, A. V. Y., & Tan, S. C. (2017). Understanding idea flow: Applying learning analytics in discourse. *Learning: Research and Practice, 3*(1), 12-29. http://dx.doi.org/10.1080/23735082.2017.1283437

# Understanding idea flow: Applying learning analytics in discourse

Alwyn Vwen Yen Lee, Seng Chee Tan

*Centre for Research and Development in Learning (CRADLE@NTU)*

*Nanyang Technological University, Singapore*

62 Nanyang Drive, North Spine, N1.2-B1-02a, Singapore 637459

alwynlee@ntu.edu.sg, sengchee.tan@ntu.edu.sg

**Alwyn Vwen Yen Lee** is a Ph.D. student and research associate at the Nanyang Technological University, Singapore. He has been researching on learning analytics related to the understanding of learning, analysis of collective responsibility within students, and using pedagogical and technological methods to sustain creative work for improving ideas through Knowledge Building theory. As part of his Ph.D. studies, he contributes to research on data mining and discourse analytics within online discussions and has also published novel work in international academic conferences related to identification and analysis of ideas in online discourse. His other works as an electrical engineer include ambient sensor networks and rule-based approaches for activity recognition. At present, he is integrating technology in the field of computer sciences with educational theory to understand further students' intentions for learning, and the applicability of skills outside of classrooms.

**Seng Chee Tan** is an associate professor and the deputy director of the Centre for Research and Development in Learning (CRADLE@NTU) at the Nanyang Technological University, Singapore. He earned his Ph.D. in Instructional Systems from the Pennsylvania State University and joined Nanyang Technological University in 2000. Prior to the appointment in CRADLE, he has taken on different roles related to advancing the use of ICT in education, including the Head of Learning Sciences and Technologies academic group in the National Institute of Education, Singapore, and an Assistant Director in the Educational Technology Division in the Ministry of Education. He has taught courses on instructional design and learning sciences at the graduate level and has conducted professional courses for organizations such as the Ministry of Education.

# Understanding idea flow: Applying learning analytics in discourse

The assessment and understanding of students' ideas in discourse have often been a difficult problem for teachers to tackle. Recent innovations and technologies such as text mining can provide a partial solution by generating an estimated count of important keywords which are representative of ideas within discourse. However, investigating idea development and flow within discourse is a much more challenging task, and requires elaborate processing and analysis. In this study, a method for analysing idea flow was proposed and tested: (a) text mining and network analysis are employed to identify and validate ideas from textual discourse; (b) identified ideas are grouped together and mapped to relevant learning objectives; (c) groups of ideas are then aggregated using a binning method and scored; (d) a flow diagram is generated using the aggregated scores to visualize idea flow within discourse. By understanding how ideas flow within discourse, discussed key ideas can be monitored and any lapses in student understanding can be identified so that teachers will have information to provide timely interventions to support and scaffold learning.

Keywords: learning analytics, discourse analysis, knowledge building discourse, idea flow, education technology

## Introduction

Educational researchers have been lamenting about the obsession of didactic teaching and lecturing just to prepare students for examination; such obsession in transmission paradigm of instruction have resulted in a culture of surface learning (Bowden, Marton & Zull, 2014). Surface learners are not concerned with deep understanding (Bain & Zimmerman, 2009). Consequently, there is little improvement and application of knowledge beyond classrooms or after examinations. Approaches such as participatory (Sfard, 1998) and knowledge creation approach (Paavola & Hakkarainen, 2005) to learning have been proposed as alternatives, which necessarily involves learning in a community. In learning communities that encourage the collective advancement of knowledge, students often need to take the initiative in identifying, analyzing and solving problems. Supported by new software and applications, this learning approach can bring learners together and provide opportunities for intellectual exploration and social interaction (Stahl, Koschmann & Suthers 2006). One type of such a learning environment is the knowledge building environment that is attuned to education in the Knowledge Age (Bereiter & Scardamalia, 2006). A critical element of knowledge building is to engage students in collaborative and constructive discourse, which entails collective contribution, co-construction, critique, and improvement of ideas. As a community, the students strive to pursue the advancement of knowledge (Scardamalia, 2002).

A key challenge, however, is in moving the community and discourse towards the right direction that leads to continuous idea improvement. Among a swathe of opinions

and information within the discourse, students have to make the balancing act between exploring diverse viewpoints that are time-consuming but can enrich the discussion, and engage in more promising ideas that could converge and lead to concrete outcomes. Tracking and monitoring ideas as discrete entities within the discourse is an ambitious task, more so when it is done qualitatively. Therefore, as the overall objective of knowledge building discourse focuses on advancing collective knowledge within the learning community (Scardamalia & Bereiter, 2003), it is crucial to investigate the flow of ideas within a discourse. In other words, to understand the trajectory or life journey of ideas, such as the discovery, flourishing, and fading of ideas. This study is guided by the overarching research question: "How could learning analytics be utilized in community discourse to support idea work within knowledge building discourse?" Specifically,

1) How can ideas within knowledge building discourse be tracked and monitored with the aid of technology?
2) How can ideas be derived from sets of keywords and contributions be measured in discourse?
3) How can major phases of idea development be visually represented?

## Understanding ideas and flow within discourse

Though there are often ample ideas in a discourse, however, to ensure productive engagement in the discourse, it is essential for students to understand the *meaning* of talks that are embedded within a discourse. The process of analyzing constructive talks that contain ideas is, nonetheless, far from being trivial. Ideas are complex entities, and to most people, represent concepts, thoughts, and knowledge. Locke (1836) initially used the word *idea* to represent the most basic unit of thought. In comparison, modern definitions (Idea, 2016) consider ideas as "transcendent entities that are a real pattern of which existing things are imperfect representations." An idea is hence, not just a unit of thought, but more of what it can achieve, such as the provision of epistemic function to represent something else, and the ability to improve beyond itself. To students, ideas could mean something that is imagined or pictured in mind, such as evolving or indefinite concepts. These ideas are often represented as inquiries within a community, and these inquiries are attempts by students trying to understand concepts that they are working on. Inquiries often exist as forum posts or discussion threads in online discourse and consist mainly of textual data that are representative of concepts, ideas and students' understanding of concepts.

In a productive discourse, instances of isolated ideas are scarce. Instead, groups of ideas tend to co-exist and revolve around certain themes of discussion. The groups of ideas also either gain prominence and attention from the community or fade over time in discourse, thus complicating the process of tracking idea movements and understanding of idea flows within discourse. Further, as students often have difficulty grasping evolving concepts due to its tentative and experimental nature, there is likely a diversity of ideas within an online discourse. Teachers constantly struggle with the sizeable amount of ideas within the community discourse, which requires significant effort to help converge similar ideas to achieve a shared understanding among the students. Although idea diversity could be a boon for the sustenance of interests in inspirational discussions and continuous improvement of ideas, it could also create disruptions and rifts within communities because of differing views. As a facilitator, teachers need to invest significant effort and time to acknowledge ideas from individual students and to engage the class in collective monitoring and analysis of ideas that reside within communal discourse. It is a real challenge to be able to do it in an efficient and scalable manner.

## Association of ideas with keywords

In addition to the above challenge, ideas in a discourse have always been difficult to monitor and track, not to mention ideas that are considered promising to the community. The ability to identify promising ideas, that is, ideas that become consequential when worked on, is essential for creative work with ideas at all levels (Chen, Scardamalia & Resendes, 2013). Textual information within an online discourse is a representation of opinions and views of students in the digital format, comparable to speech and presentations within traditional classroom discourse in the physical and audible form. Various methods for conducting discourse analysis already exist, including usage of two common choices in Natural Language Processing (NLP) – latent semantic indexing (LSI) (Hofmann, 2001) and latent Dirichlet allocation (LDA) (Blei, Ng & Jordan, 2003). Both methods are used for topic modeling in information retrieval problems.

This study focuses on the usage of unique idea-centric keywords as a core unit of analysis. Instead of discovering keywords through LDA analysis similar to Sun et al.'s paper (Sun, Zhang, Jin, & Lyu, 2014), we seek to explore deeper than the concept of topic and themes by identifying the type and nature of ideas that the students are trying to portray. The process is done by obtaining and grouping frequently occurring keywords in discourse from different sections of discourse, and relating the identified keywords with different temporal positions of the discourse. By observing the temporal positions where the sets of keywords are introduced into discourse, keyword patterns and emerging trends can then be matched with sections of discourse to be coded, so that sets of keywords approximately represent ideas within discourse. Although specific cases of misrepresentation can still occur, these rare occasions can be qualitatively analyzed on a case-by-case basis to provide a reasonably fair discourse analysis. For example, keywords that are frequently used throughout the discourse and possess overlapping meanings will have to be qualitatively assessed by coders to decide the assignment of keywords to sets.

## Analysis of idea flow using discourse analysis

It is intriguing to discover how discourse acts as a site that provides meaningful learning and knowledge construction, using verbal, imagery or textual content. As the community works on advancing their knowledge during the discourse, students often bounce ideas off one another to obtain insight and feedback. Rather than measuring the capability of individuals or the skills of teachers, educational success and failure may also be explained by the quality of educational dialogue (Mercer, 2004). Contributions to discourse can, therefore, be considered an important indicator of measuring meaningful learning within classrooms. By measuring contributions and understanding the direction of idea flow in community discourse, conducting discourse analysis within educational settings can provide insights into students' level of understanding and the social interactions within the community.

A number of approaches to analyzing written, vocal languages or semiotic events were developed, such as analysis of textual discourse within an online learning environment that can assist teachers to identify key ideas and be able to provide deeper insights (Ferguson, 2009) for knowledge building. Discourse analysis, when used as an approach to analyzing a language beyond its literal use, can inform the design and improve the productivity of knowledge creation (Chiu & Fujita, 2014). Through the usage of discourse analysis as a means of exploring imbrications between language and social-institutional practice (Fairclough, 2013), analysts, teachers and students can gauge the level of understanding within the learning environment, and also recognize how interventions are introduced to improve discourse. The process of analyzing text has

proven to be a difficult task as it is heavily dependent on interpretations and context to ensure that the source text is wholly and correctly understood (Nord, 2005). It is also a tedious job for text analysts, even with innovative usage of generic models that do not require reference to specific characteristics of context and language use. The usage of text to investigate ideas is no trivial feat, and the process is often underrated, even though it is recognized as an important step in the understanding of learning processes within the learning community. By making use of suitable technological aids, discourse can be scrutinized to identify and analyze ideas through textual content, and this could be done succinctly in a more efficient and scalable manner. Previous research on promising judgments (Chen et al., 2013) in discourse has been explored, and we agree that neither single ideas nor a mere combination of them could constitute problem solutions.

Therefore, the choice to use discourse analysis of community discourse is suitable for deciphering and identifying promising ideas within communities. With the aid of learning analytics, technological tools such as text miners could be used to help teachers in identifying presence of keywords and ideas during discourse. Through visualizations of network graphs, teachers can understand interactions between students and gauge the interests and potential directions of discussions. Most importantly, by monitoring the overall flow of ideas throughout a discourse, teachers can recognize lapses in student understanding and therefore be able to provide timely interventions to assist and scaffold learning.

## Methods

### *Study settings*

An online knowledge building discourse was investigated in this study. There were 13 graduate students and two teaching staff who participated in the discourse community, as part of an introductory course to Learning Sciences in a Master of Education program, taught over a span of 13 weeks. The course instructors adopted a knowledge building approach that engaged the participants in co-constructing their understanding about the learning sciences. For the first two sessions, the instructors modeled the method to facilitate a seminar discussion and students, in pairs or trios, took turn to lead seminar discussion the seminar discussion on particular chosen topics. The topics discussed include "Paradigms and metaphors of learning", "constructivism", "knowledge building", "situated cognition", "computer-supported collaborative learning", and "new learning environment". Knowledge Forum (Scardamalia, 2004), an online discussion tool, was used to record discussion in class and to continue the discussion after class. The participants were encouraged to participate as a knowledge builder, by contributing actively and positively to the discourse, and sharing their resources and personal experience about teaching and learning. Each student was asked to write a position paper on a specific learning sciences topic and to justify their positions. Students can opt to design and implement new ideas of instruction. For example, some topics chosen by students include enhancing direct instruction, implementing "flipped classrooms", "learning through authentic activities and assessment", and "using meaningful learning to overcome misconception in science". Consequently, the discourse among the participants contains three categories of content, with content belonging to 1) *Facilitating understanding of Learning Sciences*, 2) *Conceptualization and Design of Learning Innovations*, and 3) *Application and Validity in Natural Settings*. For more concise presentation and from this point, these three categories will be referred to as *phases* and are labeled as the *Understanding*, *Design,* and *Application* phases respectively. Content

in these three phases reflected students' progress in gaining a foundational knowledge of the field of Learning Sciences, engaging in conceptualizing and designing of innovations using technology, applying innovations on authentic problems in natural settings, and re-evaluating their designs to improve further on their ideas through knowledge building discourse.

These phases appeared within discourse in the sequence of Understanding, Design, and Application with some overlap, that is, at some point in time, the phases co-existed and ran parallel to each other during discourse. Students had, for example, acquired theories without proper understanding during the Understanding phase and begun designing their innovations in the Design phase, but due to the lack of theoretical understanding, reverted to the first phase of understanding, to concretize their knowledge of learning theories. According to the mentioned scenario, as a result, the proportion of ideas that belong to the Understanding phase decreased accordingly when students started to discuss their designs, but the proportion of ideas related to the Understanding phase increased again when students restarted discussions on a deeper understanding of theories.

During the whole discourse, students used Knowledge Forum as the primary online discourse platform, and teaching staff was also present to assist in co-creation of knowledge. Information shared among the community members are archived as knowledge artifacts for building on and future referencing, thus allowing the community to continuously improve on ideas and hence be able to advance communal knowledge. At the end of the course, the entire discourse was anonymized and analyzed by the researchers.

### *Analysis procedures*

This study used analytical and visualization toolkits to discover the presence of ideas obtained from textual content and monitor the movements of these ideas across the whole discourse. During pre-processing, the discourse was broken down into conversation turns, with each turn representing an individual participant's contribution at a particular moment in the discourse. For example, from the time that a student S1 started to pose a question and before student S2 began to reply, the content communicated by student S1 was recorded as a conversation turn. This method of talk segmentation was applied throughout the whole discourse. After which, we proceeded to identify ideas through textual content, validate the ideas, and map the ideas to relevant phases of the course so that the ideas can be visualized for easier understanding.

#### *Identification of keywords from discourse*

To identify ideas within discourse, we first used a text miner named SOBEK (Reategui, Epstein, Lorenzatti & Klemann, 2011) to identify conceptually relevant keywords and inter-keyword relationships from unstructured text data. This was done through frequency analysis of textual material, together with an inbuilt thesaurus to filter out common words. Some common words are noun markers, also known as determiners (*a, an, the, this*), pronouns (*his, her, him*), and groups of words belonging to similar concepts and meanings (*student, students, pupil*) that can be grouped under a single word (*student*). In this study, the online knowledge building discourse was mined to determine relevant conceptual keywords which are indicative of students' perceptions and ideas within discourse. Using SOBEK to text-mine discourse is considered a more neutral alternative that can provide additional insights into a discourse, as compared to methods previously used in other studies. Prior methods used before this study, seeks to extract topical keywords from a

list of keywords agreed by field experts with high (>80%) inter-rater agreement. The differences between the experts were resolved after discussions, with the remaining agreed keywords consolidated into the final list of keywords. Hence, the text mining method is preferred in this study, considering that pre-determined keywords from field experts and teachers might be influenced by their perspectives and preferences, and may not reflect ideas and thoughts which the students were engaging in discourse.

*Mapping of keywords to ideas and verification*

The identified keywords obtained through text mining were entered into the Knowledge Building Discourse Explorer (KBDeX; Oshima, Oshima & Matsuzawa, 2012), a discourse analysis program that can conduct a step-wise analysis of discourse over time. KBDeX can determine social network metrics for multiple entities within discourse, such as individual agents that initiate the conversation, the content of conversation turns, and individual keywords within discourse. Discourse data that was segmented into conversation turns during the pre-processing phase can also be collectively or individually analyzed. KBDeX offers a playback function, which allows analysts to move forward or backward in conversation turns so that certain sections of discourse can be analyzed by time, or retrospectively observed to detect the development of patterns or trends. In this study, we focused on analyzing the relationships of keywords and the emergence of keywords as discourse progressed over time during the course. Network graphs from various temporal segments of discourse were generated to verify that extracted ideas from respective sections of discourse are representative of context and content from the discourse. Once verified, ideas were then grouped and mapped according to the relevant phases of the course.

*Visualization of idea flow*

Last, we want to provide a way of visualizing the constituents and proportion of ideas within discourse in a simple yet informative manner. Ideas that are mentioned at a certain point of discourse should be recognizable, and important content should also be easily interpreted. The use of a flow diagram is an appropriate method to represent the movement of ideas and inner workings of discourse. We chose to use a variant of the Sankey diagrams to indicate idea flows within discourse for the following reasons.

First, the Sankey diagram can be used to visualize the proportion and variety of ideas within the discourse at any point in a discourse. The width of arrows used within the diagram is proportionate to the flow quantity, which in this case refers to the volume of ideas within the discourse. Our variant of the Sankey diagram focuses on maintaining a net flow of the proportion of ideas within discourse, as opposed to keeping the apparent volume of ideas consistent throughout the discourse. In this manner, we can allow generative and expansive discourse that aligns with knowledge building theory and still be able to track idea flow within the discourse. Second, Sankey diagrams place a visual emphasis on the major flows within a system. Dominant contributions within a system would be easily recognized, and in this study, the dominant group of ideas is often mainstream ideas. These would then be easily identified by teachers, and enable teachers to decide how to guide the development of ideas throughout the discourse. Also, there is flexibility for the teacher to identify other minority idea flows that are deemed essential for understanding, which can be highlighted as part of the teacher's co-contribution to the communal discourse.

Considering that the keywords and ideas at this point were already grouped into the respective phases of the course using data binning methods, we can then visualize

how the key ideas from different parts of the course pan out throughout the discourse. The scores for groups of ideas at different temporal positions of the discourse could then be used as data for fitting into the Sankey diagram. To determine scores within the binning process, the process scored conversation turns based on the presence of keywords in the individual turns, and the discrete scores provided a quantitative amount of idea presence within sections of the discourse. To avoid excessive analysis at every turn, we provided reasonable fidelity of idea flow by analyzing blocks of ten conversation turns. The scoring cycle began with the detection of keywords in turn. If keywords that belong to one of the course phases were mentioned in the turn, the turn scored a point for the particular phase. It is possible for a turn to score a point for every phase if the turn consisted of ideas from different phases. This cycle of scoring was then conducted for the whole block of ten conversation turns, and the points were accumulated for each phase. Each block consists of three scores that were awarded for the three different phases, to indicate the proportionate presence of ideas in discourse within the block of ten conversation turns. As the number of ideas at any point in time of the discourse was not definite, the proportionality of ideas was therefore calculated relatively to the total amount of ideas currently present in the discourse. When the proportionality of ideas was calculated for all blocks, complete information can be provided to generate the Sankey diagram that visualizes the idea flow within discourse.

Using the above sequence of processes, we can visually show the flow of different groups of ideas as discourse progresses, and be able to provide a general feel of how groups of ideas are moving between the various phases within the discourse. This information can also be used as feedback for teachers in understanding how content and ideas within discourse are evolving, persisting or fading within the discourse community.

## Results

The discourse was segmented into 139 conversation turns. The latter part of the discourse was filled with collaborative discussions that were used for preparing group and individual assignments, and discussion about conceptual understanding and growth. This portion of the discourse was not relevant to the goals of this study. Hence, this analysis was only conducted on the first 60 conversation turns, where the community focused on achieving the learning objectives and was engaging in constructive discourse towards understanding, designing, and application of Learning Sciences. This section consists of four sets of results that are relevant to the respective processes of identification, validation, mapping, and visualization of ideas. To achieve the final goal of monitoring idea flow within discourse, we first identify keywords within discourse.

### *Identification of keywords through text mining*

A total of 10 keywords with sufficient mention (set at more than 20 counts throughout the discourse to indicate significant usage) were discovered in the discourse, and it was not surprising that *learning* was the most mentioned keyword, due to the content of the course being heavily linked to the Learning Sciences. The list of keywords (Table 1) is an indication of key ideas that participants generate and discuss throughout the discourse. From a cursory inspection, the community was interested in discussing a few categories of topics, namely, increasing the understanding of learning sciences, ways of designing innovations using technology, and how the application of innovations creates an impact on teaching and conceptual understanding of students.

Table 1. Frequency of related keywords text mined from discourse

| Keywords | Frequency | Keywords | Frequency |
|---|---|---|---|
| Learn/Learning | 272 | Teach/Teaching | 42 |
| Knowledge | 171 | Concept/Conceptual | 40 |
| Understanding | 60 | Time | 29 |
| Students | 50 | Information | 28 |
| Process | 43 | Technology | 27 |

It is, however, important to note that it is also possible for keywords to be used in different contexts and there can be dual meanings or misinterpretation of keywords. Therefore, we delved further into the discourse to validate that the identified keywords are representative of intended ideas in the discourse, by scrutinizing the usage of language around the identified keywords itself. This process was conducted with the help of the discourse analyzer KBDeX, which generates bipartite graphs that associate keywords, discourse participants, and conversation turns together on a single platform. Network graphs of the respective agents within discourse were then analyzed over multiple sequential conversation turns, so as to obtain an understanding of how discourse occurred in the community over time.

***Determining keywords in course phases to form semblance of ideas***

Through our analysis, groups of keywords were found to surface and cluster mainly at certain sections of the discourse, and the usage of keywords was often triggered by different phases of the course. The different phases of the course (Understanding, Design, and Application) normally encompass general broad ideas, and these are driven mainly by learning objectives that are determined by the teaching staff. Our observations show that within each phase of the course, only certain keywords have a distinctively higher frequency count than other keywords within the discourse. That is to say, the majority of identified keywords surfaced within certain sections of the discourse more frequently than others. However, it also does not mean that they cannot be found in other parts of discourse. We discovered that although students were increasingly using identified keywords in more parts of the discourse, they were also not neglecting other ideas within the discourse. Besides, our analysis of such occurrences shows that the presence of lesser known keywords among other more important identified keywords in discourse does not significantly affect the way students present their ideas during the respective phases. Therefore, we conclude that the presence of related keywords in discourse shows that diverse ideas are present within discourse, but by focusing on the main ideas, students can improve their understanding and be more confident in achieving learning goals. Ideas are also explicitly represented, more so visually, based on the group of identified keywords, providing students with a clearer direction of what discussions they should be focusing on.

We sought to detect the groupings of keywords that would be broadly representative of students' ideas during the different phases. The analysis involved the use of keyword graphs to provide us with three different graphs (Figures 1 to 3) that reflect the growing number of keywords that were continuously being added to the discourse, and the building on of current ideas to what was already present within discourse. As groups of keywords were introduced at different temporal positions within the discourse, we determined the usage of the introduced keywords as being related to the particular phase of course content. The edges (straight lines between nodes) in the network graphs, between nodes that contain keywords, are representative of relationships between the keywords due to one of the following reasons. The nodes are either linked due to 1) usage of the identified keywords within the same note in Knowledge Forum, or 2) expression of the identified keywords by the same student, but in different notes or at different temporal positions within the period of discourse.

In Figure 1, the keyword graph was sampled at the 8th conversation turn near the beginning of discourse, and the identified keywords are *Learn* (inclusive of *Learning*), *Knowledge*, *Understanding,* and *Information*. These keywords have a dominant presence in the early stages of discourse, representative of the limited understanding students possess at this early stage in discourse regarding the topic of the Learning Sciences. Students had also started linking keywords together and were forming initial concepts about the field of Learning Sciences. The keyword graph in Figure 1 is, therefore, determined to be related to phase one of the course, which is *Facilitating Understanding of Learning Sciences*.

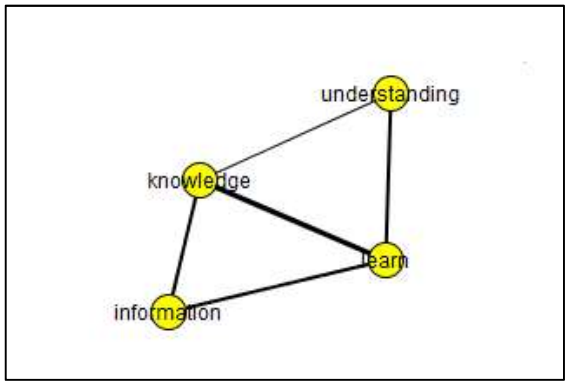


Figure 1. Keywords with high frequency in the Understanding phase of the course.

Further downstream, we found that students had begun conducting more in-depth research regarding the field of Learning Sciences and about learning processes. The students were posting more "how" questions and were also questioning one another at a deeper level. The increased interactions are clearly represented by the thicker edges seen in Figure 2, with heavier usage of the previously mentioned keywords of *Learn*, *Understanding*, *Knowledge,* and *Information*. The slightly expanded keyword graph was sampled after the first phase of the course was completed; Figure 2 shows that newer keywords were added to the top flanks of the keyword graph. The newly added keywords are *Process*, *Time* and C*oncept*, signifying the expanded usage of vocabulary and students' engagement in trying to understand topics further. More complex issues were discussed, as it included discussion on the underlying learning processes and resources required for understanding conceptual learning, such as time and effort. The keyword

graph in Figure 2 is, therefore, representative of ideas built on the first phase of the course and is also visibly present in phase two of the course, which is *Conceptualization and Design of Learning Innovations*.

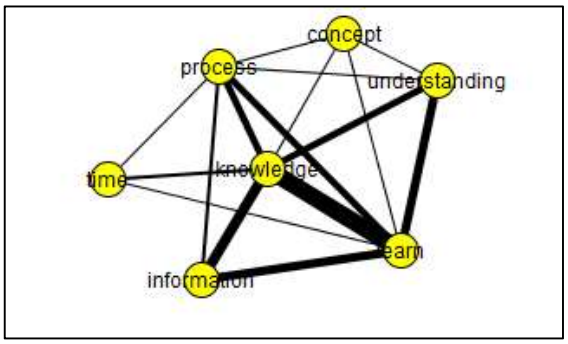


Figure 2. Additional keywords with high frequency (concept, process, time) identified in the Design phase.

Moving on to the middle of the discourse, students were taking more time in digesting the intricacies and complexities of learning processes and were also stumped at the initial part of the design phase. As compared to traditional classroom teaching where explicitly detailed instructions were provided, and students were following to the letter to completing tasks, teachers had a more challenging task of ensuring students understand what they have learned, and be able to conceptualize and design innovations that might not follow standard teaching materials. This mode of instruction and learning is notably different from just planning for achievement of learning goals and objectives. With the aid of technology, the task was made easier but remains challenging. Tech-savvy students quickly picked up on this point during discussions and suggested the usage of technology to provide solutions that could aid in improving teaching methods. As a result, Figure 3 reflects the newer inclusion of keywords such as *Technology*, *Teach* (inclusive of *Teaching*) and *Student* to the keyword graph. Overall, the third keyword graph is representative of ideas that are present from the beginning to the middle of the communal discourse and also shows a final snapshot of relevant keywords that students frequently used within the discourse.

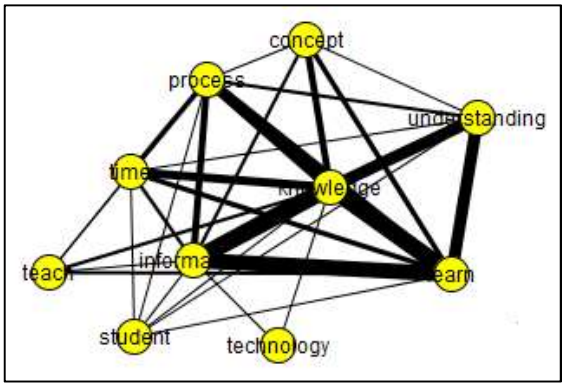


Figure 3. A newer inclusion of keywords with high frequency (technology, teach, student) in the Application phase.

In addition to text mining being an alternative and more neutral method of searching for keywords, using text mining to identify keywords within a discourse can also alleviate the need to manually search for relevant keywords and make the process more efficient and less tedious. To validate the process, we viewed the discourse using conversation turns and used keyword graphs to visually relate different phases of the course with the emergence of ideas in discourse over time. These analyses were conducted to ensure that there were minimal misunderstanding and deviation in the intention of keyword usage in the discourse. To further clarify the relationships between ideas within discourse and phases of the course, we mapped these entities together in Table 2, and also extracted relevant quotes from the discourse to highlight the actual intentions of students when keywords were being mentioned in discourse.

### *Mapping of course phases and related keywords to approximate ideas*

Table 2 relates course phases, identified keywords, important ideas, backed up by relevant quotes mentioned in discourse, to justify that groups of keywords can represent ideas within discourse, and also be linked to different phases of the course. The course phases were identified from discourse and matched with relevant parts of the course syllabus, while the keywords were identified from the text mining processes. Selected short quotes were obtained from the discourse to provide a context of the discussed content, and ideas that are important to the course were mapped to the relevant respective phases, keyword groups, and discourse quotes. By relating the various entities together, we determined keywords that can adequately represent particular groups of ideas within each phase of the course. Also, we verified the relationship between keywords, ideas and course phases, using keyword graphs (Figures 1 to 3) that show the emergence of keywords and ideas over time.

On a side note, we acknowledge that it is possible that certain keywords can be present in multiple phases of discourse. We addressed this issue through our method, by providing varying weightage to certain keywords within different sets of keywords, based on the temporal usage, frequency, and positioning of the keywords in the discourse. For example, the usage of the keyword “learn” has been identified as an action in Phase 1, a process in Phase 2, and a mixture of both in Phase 3. The keyword “learn” is however emphasized and preferred as a more important keyword to be understood in Phase 1 due to the discretion of the educators, and is, therefore, prioritized to be understood before making further progress through the course.

The relation between sets of keywords and phases of discourse is, however, not visually enticing to users and can be difficult to use and understand at a glance. To further aid teachers in understanding the flow of ideas within discourse, a more tangible and visible form of idea flow should be presented to enable faster and easier recognition of ideas within communities.

Table 2. Relationship between course phases, related keywords, quotes and mapped ideas from discourse

| Course phases | Identified and related keywords | Related quotes obtained from discourse | Relevant mapped ideas |
|---|---|---|---|
| Course Phase 1: Facilitating understanding of Learning Sciences<br>• How do people learn?<br>• How do I teach to help learners acquire knowledge and knowledge building process? | Learn, Understanding, Knowledge, Information, | • "What does it mean to learn something?"<br>• "To learn is more than attaining knowledge…"<br>• "…knowledge is beyond attaining information, but to have a deeper understanding…" | • Theories of learning sciences<br>• Understanding learners and learning<br>• Deep learning versus surface learning<br>• Difference between knowledge and information |
| Course Phase 2: Conceptualization and design of learning innovations<br>• Design, analyze and discuss various theoretical principles and learning models of how learning takes place | Process, Concept, Time | • "…learning as a continuous process…"<br>• "…influence and bring about conceptual growth and change…"<br>• "…insight into this natural process…<br>• "Learner progress over time…" | • Understanding learning as a process<br>• Processes related to knowledge acquisition and building<br>• Design and conceptualization of innovations for learning<br>• Analysis of conceptual growth and change in students over time |
| Course Phase 3: Application and validity in natural settings<br>• Using technology, discuss how principles and models can be applied to design of instruction, teaching and learning | Technology, Teach, Students | • "Technology as a tool…"<br>• "…enhance teaching…"<br>• "…how technology has impacted the way individuals and communities learn…"<br>• "…model is adopted in my teaching…"<br>• "…used as a tool to evaluate students learning…" | • Usage of technology for supporting teaching and scaffolding for learning<br>• Enhance efficiency and adoption of methods and tools for authentic application and practice<br>• Evaluation and validation of learning |

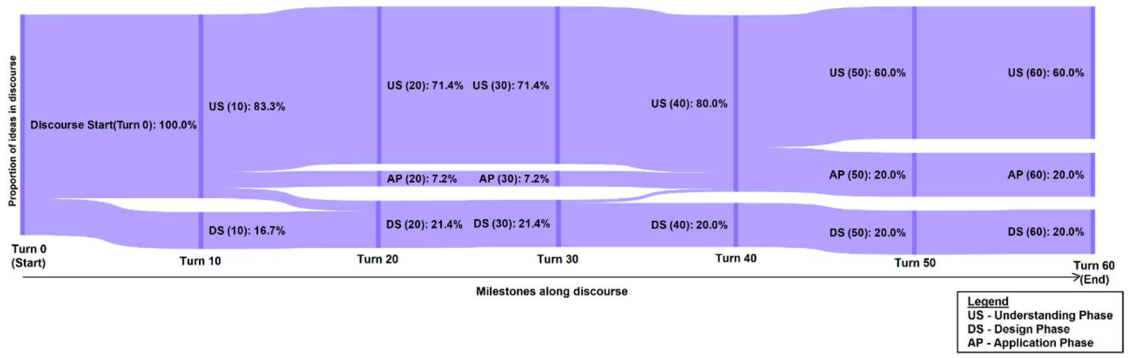


Figure 4. Sankey diagram of idea flow within discourse for the period of analysis (Turn 0 to Turn 60).

### *Construction of flow diagram to visualize idea flow*

The group of ideas belonging to each phase of the course was scored, providing us with a quantitative value of ideas that are present at any point within discourse. The scores allow us to understand further the volume and dominance of ideas present within various sections of discourse. The Sankey diagram was generated for the first 60 conversation turns, as the discourse till that point in discourse was focused mainly on the understanding, design and application phases of the discourse, which we are interested in investigating within this study. We would then be able to obtain a clearer picture of idea trajectories and flows within discourse by focusing on this section of discourse.

Figure 4 shows the Sankey diagram of idea flow within the discourse, with the blocks of ten conversation turns indicated by milestones. The leftmost milestone indicates the original starting point in which the discourse started on a clean slate with no ideas being generated. As discourse progresses, the labels in the other milestones, especially those with separate flows, represent the different types of ideas that are flowing throughout the discourse. The label on the milestones should be read in the following manner:

<Phase which content belongs to> (Turns from the start): <Relative proportion of ideas in %>
Example: US (10): 83.3%

The example can be interpreted as such: Content and ideas related to the understanding of Learning Sciences are present in discourse by turn 10 of the discourse and represents 83.3% of all ideas at the point in time.

### Discussions and significance of analysis

After the start of discourse, the discourse nearing turn 10 started splitting up into ideas about the understanding of Learning Sciences and design of innovations. Expectedly, the dominant ideas revolve around understanding and learning about the field, and it is also understandably too early in discourse to expect the application of innovations when it has not even been discussed. In this portion of the discourse, students were still trying to understand theories of Learning Sciences, and making sense of how to design innovative theories and models of learning. For example, student S1 was trying to explain views about knowledge and learning in the following extract from Turn 4 of the discourse. The

explanation, as it may be, is what we would expect from students who are new to the field.

> It is how people acquire knowledge or skills and synthesize that knowledge or skills into various form. It takes time. — S1

Subsequently, the milestone at turn 20 shows some deviation of ideas related to the design and application phases. The deviation is reflected by the slight drop in the proportion of ideas on Understanding, and discussions start to move into phases of Design and Application of innovations. At this point of the discourse, the discourse community seems to have achieved a basic grasp of theoretical frameworks and are ready to devote more time to the understanding of innovative designs, before committing significant time to application of the designs. The proportion of ideas remained consistent for the next ten turns of discourse. By turn 40, however, it seems that efforts to discuss the application of designs have been abandoned temporarily. It was suspected that students required some time to converge on diverse ideas and come to a consensus on what kind of designs are workable in an authentic environment. This change is reflected in the Sankey flow diagram, as the proportion of ideas from the Design and Application phases converged back to the Understanding phase, as students strive to understand further and deepen their knowledge, and are returning to theories to affirm their discoveries. Moreover, as students one another's comments and also share their experiences and views, ideas and information continue to flow between the Understanding and Design phases, but there is still little indication of ideas and content in the Application phase at this point of the discourse by Turn 40. The result of continuous revision and sharing of ideas by students can be observed in some of the notes, such as those posted by student S9 and S1, where we noted a deeper sense of understanding regarding knowledge building and learning.

> Knowledge is not just simply facts and procedures which are not as applicable should anyone face a myriad of situations in the real world. Knowledge can be representations in the mind, that should be linked from one representation to another representation. It is being able to apply what you know in any situation and being able to adapt and reflect therefore being able to transit from a novice to an expert. Scaffolding, social interaction, articulating ideas, constructivism that is building ideas based on prior knowledge, reflection are some ways for building knowledge. — S9

> We understand "knowledge of" as a "deeper understanding" (beyond the problem of understanding) and "knowledge about" is factual understanding (knowing what). In the traditional classroom it is purely on delivery and acquisition of factual knowledge unlike the KB environment it advocates idea improvement and really understanding beyond the problem. — S1

The quotations from the students were shared with the community and were noticeably longer and expressed more clearly than the shorter explanation by S1 earlier in the discourse. In the meantime, students were also resuming their effort and dedicating more resources to design and application of innovative designs in the Design and Application phases.

The Sankey flow diagram analyzed the first 60 turns from the beginning of the discourse, with an estimated 60/20/20 split of ideas and content related to understanding, design, and application respectively. The split is reflective of the participants' mood within the discourse, where the majority of the community was more concerned with

understanding topics regarding learning but students were willing also to move ahead and were pondering about the design and application aspects of their proposed innovative designs in authentic and natural settings, such as classrooms.

Overall, we observed there is a dominant flow of ideas (>60%) related to understanding and learning, as students continually seek to understand the field of Learning Sciences. Given a class of students with unknown capability, there is always a possibility that some students might struggle to understand the field throughout the whole discourse. These students regularly inquire and question to seek clarification, while other students are also continually reinforcing their conceptual knowledge through knowledge-sharing and discourse. The above are just a few of the possible reasons that might have led to the dominant flow of ideas belonging to Understanding phase, as seen in Figure 4.

Also, we noted that at the end of the analysis, ideas related to the Design and Application share an equal amount of attention from the community. We describe this phenomenon as part of the design cycle that is highly encouraged in higher education classes. This phenomenon is prominently visible within this study's discourse, as the graduate students are keen to revise their designs continuously through constant redesigns and re-evaluations, in the hope of creating or modifying a successful working learning model. Further, as most of these graduate students are also in-service teachers, they know that their innovative designs would eventually be applied in authentic situations and therefore impact students that they teach. Hence it explains the significant amount of effort that is used to discuss ideas about the design and application of innovations.

In retrospect, the tool KBDeX which was used to assist in temporal analysis keyword networks also has its limitations which we intend to address. Although it was successfully used for this study, the lack of consideration of dominant terms is something that we will include in future research. By considering term frequency-inverse document frequency (tf-idf), the numerical statistic would allow us to further improve our accuracy in both text mining and network analysis processes, by offsetting frequency of words that appear more often in discourse.

Overall, our analysis and results have shown that we can track and monitor the idea flow within knowledge building discourse to support idea work. We approximate ideas using textual data, by representing the ideas through sets of keywords and measuring contributions to the community discourse at different sections of discourse. Text mining techniques were used to extract keywords, network analysis identified sets of keywords to be matched to ideas, and the proportion of ideas for various temporal positions in discourse can be presented in a variant of the Sankey diagram. By bringing forward a tangible and visually understandable diagram representing idea flow within discourse, teachers can grasp and use the diagrams as feedback for improving teaching and learning. The usage of learning analytics can help teachers and analysts to improve further educational processes of teaching through discourse, such as devoting teaching resources to students who require more attention and rendering further support. It also offers other benefits such as revealing deeper insights into writing processes and inter-student communicative skills.

## Conclusions

There are many benefits to learning analytics, most distinctly, when deployed in the field of discourse. There are large troves of data stored within discourse databases, which are awaiting analysis to provide insights into how students learn and to inform ways to help students succeed. Our methods in this study propose a capability to understand the flow of ideas within discourse, by investigating a sequence of processes that are required for

transforming the physical presence of textual data in different parts of the discourse, into movements of ideas that can be visualized in a flow diagram. We used a text miner to identify keywords that are representative of ideas within discourse. Network graphs, using keywords as the core unit of analysis, were subsequently used to validate the groups of ideas that procedurally emerge within the period of discourse. The course objectives were then mapped correspondingly to the groups of ideas to understand the relation between groups of ideas and course themes. To understand further the movement of ideas within discourse, a flow diagram was constructed to investigate the development and flow of ideas within discourse. The flow diagram can be interpreted by both teachers and students as snapshots of ideas at any point in discourse, and also provides a visible form of idea representation within discourse.

Through the monitoring and tracking of idea movements using the Sankey diagram, teachers can uncover possible underlying reasons that result in the direction of discussions. For example, by using data that are provided by students, teachers can gauge levels of understanding, the types of ideas present within discourse without the need to constantly monitor discourse and scrutinize individual discussion threads or posts. Further, the intentions of students within discourse can also be tracked and understood, thus forming an idea trajectory, showing how ideas originated, evolved or faded out throughout the discourse. By using the flow diagram, teachers and students would be able to pinpoint ideas with disrupted flow throughout discourse, which represent lapses in understanding and possible oversights by teachers during teaching. Teachers can also pay attention to specific parts of the discourse that might have overly dominant idea flows, by tailoring the course according to student needs, in part of efforts to improve the overall teaching and learning for future courses.

## Acknowledgements

The research reported here is supported by the Centre for Research and Development in Learning, Nanyang Technological University (CRADLE@NTU). The research team would also like to thank the teaching staff and students who participated in this study.

## Disclosure Statement

No potential conflict of interest was reported by the authors.

## References

Bain, K., & Zimmerman, J. (2009). Understanding great teaching. *Peer Review*, *11*(2), 9-12.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Bowden, J., Marton, F., & Zull, J. (2014). The deep learning process and the construction of knowledge. In *Facilitating Deep Learning: Pathways to Success for University and College Teachers*, (pp. 15-48). New Jersey: Apple Academic Press.

Chen, B., Scardamalia, M., Resendes, M. (2013). Dynamics of Promisingness Judgements in Knowledge Building Work of 8- to 10-Years-Olds. Poster presented at the 2013 AERA Annual Meeting, San Francisco, CA.

Chiu, M. M., and Fujita, N. (2014). Statistical Discourse Analysis of Online Discussions: Informal Cognition, Social Metacognition, and Knowledge Creation. In S. C. Tan, H. J. So, & J. Yeo (Eds.), *Knowledge creation in education* (pp.97-112). Singapore: Springer Singapore.

Fairclough, N., 2013. *Critical discourse analysis: The critical study of language*. New York, Routledge.

Ferguson, R. (2009). *The Construction of Shared Knowledge through Asynchronous Dialogue*. (Doctoral dissertation). Retrieved from The Open University. (http://oro.open.ac.uk)

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1-2), 177-196.

Idea [Def. 1a]. (2016). In *Merriam-Webster Online*. Retrieved November 2, 2016, from http://www.merriam-webster.com/dictionary/idea.

Locke, J. (1836). *An essay concerning human understanding*. London: Balne, Printer, Gracechurch Street.

Mercer, N. (2004). Sociocultural discourse analysis: Analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics*, *1*(2), 137-168.

Nord, C. (2005). *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis*. New York: Rodopi.

Oshima, J., Oshima, R., & Matsuzawa, Y. (2012). Knowledge Building Discourse Explorer: a social network analysis application for knowledge building discourse. *Educational technology research and development*, *60*(5), 903-921.

Paavola, S. & Hakkarainen, K. (2005). The knowledge creation metaphor - An emergent epistemological approach to learning. *Science & Education, 14*, 535-557.

Reategui, E., Epstein, D., Lorenzatti, A., & Klemann, M. (2011). Sobek: A text mining tool for educational applications. In R. Stahlbock (Ed.), *International conference on data mining* (pp. 59–64). CSREA Press. Retrieved from http://cobweb.cs.uga.edu/~hra/2011-proceedings/dmin/papers.pdf

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. *Liberal education in a knowledge society*, *97*, 67-98.

Scardamalia, M., & Bereiter, C. (2003). Knowledge building. In J. W. Guthrie (Ed.), *Encyclopedia of education* (pp. 1370-1373). New York: Macmillan Reference Books.

Scardamalia, M. (2004). CSILE/Knowledge Forum. *Education and technology: An encyclopedia*, 183-192. Santa Barbara: ABC-CLIO.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 97-115). New York: Cambridge University Press.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 409-426). New York: Cambridge University Press.

Sun, W., Zhang, J., Jin, H., Lyu, S. (2014). Analyzing Online Knowledge-Building Discourse Using Probabilistic Topic Models. *International Conference of the Learning Sciences*, 2, 823-830.