Microgenetic Learning Analytics: A Computational Approach to Research on Student Learning

By

Florence R. Sullivan
(corresponding author)
fsullivan@educ.umass.edu
College of Education
University of Massachusetts, Amherst

W. Richard Adrion
School of Computer Science
University of Massachusetts, Amherst

P. Kevin Keith
College of Education
University of Massachusetts, Amherst

A paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 16 – 20, 2015

Abstract: In this paper, we present our work-in-progress related to a new research method we term Microgenetic Learning Analytics. The goal of our work is the development of a theoretically based, computational method for performing microgenetic analysis of, co-present, collaborative problem-solving group conversations in a robotics-learning environment. Our research focuses on the development of microgenetic learning analytic methods that are built on a strong theoretical foundation based in Vygotsky's socio-cultural theories of learning, Bakhtin's (1986) notions of speech genre's and Goffman's (1974) theory of social frameworks. We theorize a problem solving in computational environments (PSCE) speech genre, which includes a bounded domain of speech featuring the relatively stable use of particular types of talk. We seek to understand the "work" utterances are doing in terms of the groups' problem solving efforts within the PSCE. We coded trigrams utilizing a natural language processing java library, developed at Stanford, featuring a parts-of-speech (POS) tagging Treebank developed at the University of Pennsylvania. Then, through a deliberative process, we mapped the POS trigram tags to PSCE speech genre codes. In this way, we developed a picture, over time, of student collaborative problem solving talk. The very preliminary results of our work, thus far, indicate the method is useful for identifying interactionally rich sections of the transcript. We have been able to detect an important shift in student understanding of the use of a sensor in the robotics environment from that of sensor as measurement device to sensor as computational device. Limitations and plans for further development of the method are discussed.

In this paper, we present our work-in-progress related to a new research method we term Microgenetic Learning Analytics. The goal of our work is the development of a theoretically based, computational method for performing microgenetic analysis of discourse data (specifically, co-present, collaborative problem-solving group talk in a robotics learning environment). Microgenetic analysis is one of the most robust forms of educational research on learning one can undertake (Kuhn, 2002). However, due to the intensive nature of data collection and analysis, microgenetic research is typically performed with very small numbers of participants. This limits the utility of findings.

In our work we seek to address known limitations of both microgenetic research techniques and those of learning analytic techniques (as currently practiced), towards the goal of: (a) expanding researchers' ability to perform microgenetic analysis of data from a larger number of participants; and (b) expanding the range of microgenetic questions that can be asked and answered. We believe that computational means can be developed and meaningfully deployed to assist researchers in microgenetic research. Here we present the current state of our method development. The organization of this paper is as follows: first, we define microgenetic analysis from a Vygotskyan perspective. Second, we discuss learning analytics and the various approaches educational researchers are currently taking to create computational means of performing educational research. Next we present learning analytic approaches to analyzing talk and we discuss the special issues that cohere when working with face-to-face conversational data. We then present our theoretical approach to developing computational means for performing microgenetic analysis, describe the method we have developed, thus far, and provide some results from a preliminary application of the method to an existing data set.

Microgenetic Analysis: A Sociocultural Approach to Understanding Learning

Our theory of learning that guides the development of this method is rooted in the work of Vygotsky (1978). Vygotsky argues that learning leads development and that all learning occurs on two planes, first inter-psychologically (through social interaction) and then intra-psychologically (through internalization); hence, the primacy of social interaction in learning. According to Vygotsky, learning creates what is known as the zone of proximal development (ZPD). The ZPD exists in contrast to a student's actual developmental level, which may be assessed with various measurements. The ZPD, however, is the developmental level a student can obtain through the help of an adult or a more capable peer. In Vygotskyan psychology, the adult or more capable peer regulates the learning of the student through providing scaffolds to understanding. At some point, the student internalizes the use of the scaffold and is thus able to regulate her own learning in a given area (Wertsch, 1979). As Vygotsky argues "...learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment and in cooperation with his peers. Once these processes are internalized, they become part of the child's independent developmental achievement (p. 90)."

A second tenet of Vygotskyan psychology that is highly relevant to the development of this research approach is the idea that both signs (language) and tools mediate learning. According to Vygotsky (1978), to build a more complete picture of the development of human understanding one must attend to both how people use language and how they use external tools in learning activities. In attending to such as a researcher, Vygotsky developed an approach that is now known as microgenetic analysis.

Microgenetic Analysis

Wertsch (1991) presents four socioculturally-based time scales of human development as follows: (1) cultural historical time, which refers to the development of culture and society over time; (2) phylogenesis, which refers to the evolutionary development of a species; (3) ontogenesis, which refers to developmental processes that occur in an individual in the span of her lifetime; and (4) microgenesis, which refers to the development of a specific understanding or facility in an individual that occurs over a short period of time (minutes, hours, days). The development of certain types of thinking or the understanding of specific concepts happens at the level of microgenesis.

Microgenetic analysis, therefore, is an observational research technique in which the researcher attends closely to the social interactions and the use of tools within the learning environment in order to understand the genesis (or the origins) of cognitive change. Siegler (2006) has described three essential properties of microgenetic analysis: "(1) observations span the period of rapidly changing competence; (2) within this period, the density of observations is high, relative to the rate of change; and (3) observations are analyzed intensively, with the goal of inferring the representations and processes that gave rise to them" (p. 469). As Siegler points out, the robustness of microgenetic techniques for understanding cognitive growth derives from high-density observation. Collecting and analyzing all interactions over a given period of time gives the researcher the advantage of understanding the trajectory of the cognitive change, including what preceded it, what followed it, and the durability of the change. Empirical findings arrived at through microgenetic analysis are remarkably consistent (Kuhn, 2002).

In our own research, we have utilized microgenetic techniques to understand how

students develop their systems understanding while solving robotics problems (Sullivan, 2008), how students develop and extend creative ideas in collaborative robotics contexts (Sullivan, 2011), how students develop problem-solving strategies to solve robotics problems (Sullivan & Lin, 2012) and how students negotiate the social environment of collaborative interactions as they decide on the "how" of group work (Sullivan & Wilson, 2015). Our microgenetic technique involves the video-based collection of observational data including all student and teacher utterances and student interactions with tools in the environment (computers, mini-computers, sensors, Legos, student worksheets and other written material). Our data analysis methods involve close viewing of videotaped interactions in which we either apply theoretically derived codes and/or develop data driven codes that help us identify instances of the phenomena of interest. This work has allowed us to begin to build a picture of aspects of students' computational thinking (systems understanding; creativity; collaboration) in robotics contexts.

In conducting this work, we have become intimately familiar with the constraints of the microgenetic technique: primarily, the amount of time it takes to collect the data and to conduct close analysis of it (Siegler, 2006), and secondarily the lack of the generalizability of findings derived from microgenetic case studies featuring only a few participants (Pressley, 1992). Yet, as Siegler has pointed out, the level of explanatory power made possible by microgenetic analysis vis-à-vis the development of higher order processes is unparalleled by other analytic techniques. With the constraints and affordances of current microgenetic techniques in mind, we have turned to the burgeoning world of learning analytics to consider whether it may be profitably deployed to mitigate the time and generalizability constraints of microgenetic techniques.

Learning Analytics

Learning analytics is a general term used to define a number of new computational methods of educational research that developed as a result of advances in computing and the availability of "big data" sets created through student engagement with online learning environments (Atkisson & Wiley, 2011; Fournier, Kop & Sitlia, 2011; Siemens & Gašević, 2012). Learning analytics has been defined by the organizers of the first annual learning analytics conference as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (as cited by Vatrapu, 2011 p. 127). Learning analytics utilizes computational means to gather and analyze a variety of learner-generated data including (but not limited to) mouse clicks, button pushes, keystrokes, text and other keyboard or screen-based movements (Blikstein, 2011).

Learning analytics has been deployed to examine learning processes as they are related to the development of online learning social networks (deLiddo, Shum, Quinto, Bachler & Cannavacciuolo, 2011), the timing of high level and high participation discussions in online learning settings (Ferguson & Shum, 2011), and the identification of specific and varied problem-solving approaches in writing computer programs by skill level, for example, novice vs. experienced programmers (Blikstein, 2011). Findings related to these studies primarily have implications for the development of either tools (e.g., the linking and annotating social networking tool Cohere (deLiddo, Shum, Quinto, Bachler & Cannavacciuolo)) for structuring student interactions online (Ferguson & Shum) and for the development of differentiated scaffolds for learning for students of varying levels of expertise in computer programming settings (Blikstein).

Approaches to learning analytics are both behavioral and discourse-centric. For example, as regards the collection of behavioral data, Blikstein (2011) recorded all of the participant students' actions while working in a specific computer-programming environment, including their cutting and pasting activity and their accessing of provided code examples. In this way, he was able to build a timeline of student activity, which revealed student approaches to writing the computer programs. Meanwhile, as regards the analysis of discourse, Ferguson and Shum (2011) used a technique that focused on identifying specific discourse features in online discussions as well as time stamp data and counts of participants to allow them to pinpoint particularly meaningful discussions in the context of a daylong online workshop for teachers.

Learning analytics, then, can be deployed to analyze the same types of interactions that we focus on with our microgenetic techniques: behavior (tool interactions) and discourse (social interactions). Because of this, we argue that microgenetic learning analytic techniques can be developed and deployed and that such techniques will meaningfully address the two constraints of current microgenetic techniques: time and generalizability. Microgenetic learning analytic techniques have the potential to greatly expand the generalizability of empirical findings related to student learning and, in turn, to allow us to confidently design robust tools and methods of instruction to support learning. Moreover, our proposed new method of Microgenetic Learning Analysis may well open up new areas of microgenetic research that, to this point, have been too cumbersome to undertake.

While learning analytics presents a potentially powerful new approach to microgenetic analysis, there is a critical concern with the method that must be considered.

Atkisson and Wiley (2011) point out the possibility of confusing methods with meaning. They argue that the availability of large data sets and open source learning analytic tools with which to explore them may lead to unprincipled, atheoretical "poking around" in the data. Such analysis may lead to the reporting of findings that are replicable with other data sets, but tell us little about the actual process of learning. They argue that learning analytics research must proceed from strong theoretical foundations that guide and direct data gathering and data analytic procedures. Gašević, Dawson and Seimens, (2015) make a similar argument, they advocate for learning analytic approaches that are well aligned to the goals of educational research, focusing on how learners construct knowledge.

Our research focuses on the development of microgenetic learning analytic methods that are built on a strong theoretical foundation based in Vygotsky's sociocultural theories of learning, Bakhtin's (1986) notions of speech genre's and Goffman's (1974) theory of social frameworks (discussed below). Our work is in line with Shum and Ferguson's (2012) notion of social learning analytics. Though social learning analytics is still geared, primarily, towards the analysis of online data, Shum and Ferguson have a sociocultural view on learning and argue that online discourse exchanges can be meaningfully analyzed with the aid of computational means.

Analysis of co-present data. Our research extends learning analytics beyond the analysis of online learning data. We are analyzing discourse data captured on video and audio in face-to-face classrooms and informal learning environments. Sherin (2013) is also working on developing computational means for analyzing co-present discourse data; in his case, seeking to understand students' commonsense learning in science. Sherin's work uses statistical natural language processing methods (vector space

modeling and cluster analysis) to examine student models of the seasons and dynamic shifts in those models that may occur during the course of a clinical interview. This approach depends, in part, on counting the instances of particular terms in a student's speech and the proximity of such to one another. Our approach also deals with student talk in co-present situations and conceptual change over time, but our student participants are working in small collaborative groups, not responding to interviewer prompts: we are analyzing their collaborative problem solving discussions. This is a different type of data to that which Sherin analyzed, and as a result it requires a different analytic approach. Our data is highly contextualized talk, which predominately features indexical terms (this, that, here, there) and pronominal terms (consisting largely of the word "it") uttered in sentence fragments. To assist us in investigating how conceptual understanding is developed with these data we have turned to the notions of speech genres (Bakhtin, 1986) and social frameworks (Goffman, 1974).

Speech Genres and Social Frameworks

According to Bakhtin (1986), speech genres are characterized by relatively stable types of utterances occurring within a particular sphere of human activity. There are many and varied types of speech genres from everyday talk – "short rejoinders in everyday dialogue" (p. 60) - to various forms of writing (e.g., the novel, scientific reports) to verbal military commands to poetry. The social and symbiotic nature of speech genres may be regarded as tools that help us act in and make sense of the world, and as products of our acting in and making sense of the world (Varelas, Becker, Luster & Wenzel, 2002). Speech genres serve to organize a sequence of interactions in a culturally recognizable situation (Wells, 1999). This culturally recognizable situation

may best be thought of as Goffman's (1974) social interaction frame.

Goffman (1974) argues that all social interactions are framed by the socio-cultural context and an individual's understanding and interpretation of that context. Social frameworks "provide background understanding for events that incorporate the will, aim, and controlling effort of an intelligence, a live agency, the chief one being the human being" (p. 22). Varenne (1998) adds to Goffman's (1974) frame theory by discussing the "always already there" (p. 185) impact of historically situated cultural and social facts. However, both theorists stress the idea that individuals - while influenced by the cultural and social frames they are born into - have the ability to act independently to achieve their own specific goals.

Therefore, specific socio-cultural contexts, such as working in a small collaborative group to solve a robotics problem in a sixth grade science class, invoke a social interaction frame for students and evince the relatively stable use of particular utterances in speech interactions. In this paper, we define a problem solving in a computational environment (PSCE) speech genre that refers to talk that occurs among middle school students within the context of solving a robotics problem in class (a particular sphere of human activity). The categories created for this analysis are reported in the methods section. In adopting this speech genre approach, we are *not* attempting to map out the entire domain of possible speech acts that may occur in the setting; rather we are seeking to identify the regularities in the speech genre that may point, over time, to the microgenetic development of conceptual understanding.

To guide us in this endeavor, we also turned to the work of Ginzburg and Fernandez (2010) who have contributed, theoretically, to the development of

computational models of dialogue for automated agents in tele-service settings; for example, booking an airline flight over the phone. From Ginzburg and Fernandez's work, we derived ideas for developing a coding scheme for sentence fragments, as the majority of the utterances in our data corpus are highly contextualized sentence fragments featuring indexical and pronominal terms. We now turn to a description of the method we have developed using this speech genre theoretical framework.

Methods

Speech Genre Analysis and Qualitative Models of Activity

Our method includes a speech genre analysis in which we seek to understand the "work" that particular types of utterances are doing in a given student interaction while solving a robotics problem. We utilize qualitative models of student activity in the problem-solving environment to help us contextualize student utterances and better understand the possible meaning of an utterance. In our previous work, we have developed a number of qualitative models of student learning activity in robotics environments – including activity related to the development of scientific habits of mind (Sullivan, 2008), the development of problem solving strategies (Sullivan & Lin, 2012) and the process of collaborative creativity (Sullivan, 2011). These qualitative models provide a framework with which to begin demarcating elements that are characteristic of a particular activity, including the types of speech that might be used in the setting.

In developing our method, thus far, we are using a data set we collected as part of a prior study. This data set consists of over 36 hours of two middle school student groups collaboratively solving robotics problems. From a text mining perspective, this is a huge data set containing well over 250,000 words and 36,000 individual utterances. For the

purposes of this method development project, we chose to work with a portion of this data set that we had previously analyzed (Sullivan, 2011). In this prior analysis, we examined student's collaborative development of a creative idea in solving a robotics problem.

As part of this prior analysis, we developed a qualitative model of student problem solving activity with robotics. This model consisted of a troubleshooting cycle, which includes the following activity: "(1) writing and testing the program, (2) diagnosing problems with the program or structure of the device, (3) proposing and arguing for specific changes to the program/structure, (4) making changes to the program/structure, and (5) testing the device again" (Sullivan, 2011, 57). In our previous analysis, we identified 17 discreet instances of students working through the troubleshooting cycle over a 30-minute period. The troubleshooting cycle is a relatively regular and stable feature of student activity while solving robotics problems. We worked with this qualitative model of the troubleshooting cycle as the basis for developing a temporal analysis of the data set, which is crucial to microgenetic analysis. Through our computational analysis, we sought to identify regularities in speech that might map to the identified regularities in student troubleshooting activity.

Computational Analysis – Parts of Speech

Based on the troubleshooting cycle qualitative model of student activity, we sought to linguistically identify student activity over time. We did this through using a natural language processing library created by colleagues at Stanford University ((Toutanova, Klein, Manning, & Singer, 2003), featuring a parts-of-speech Treebank developed at the University of Pennsylvania by Santorini (1990). Due to the relatively

stable character of troubleshooting cycle activity, we hypothesized a relatively stable character to the domain of utterances that may be offered during these times: in short, an identifiable speech genre. We reasoned that the parts-of-speech tagger would begin to help us identify types of utterance that may all be doing the same type of "work" in terms of the troubleshooting cycle; for example, we sought to linguistically identify periods of diagnostic activity, periods of argumentation, and periods of problem definition activity.

Ngram analysis. We chose to work with utterances at the level of the bigram (two words) and trigram (three words). We selected these ngrams because we reasoned these were the smallest level at which complete utterances might be made. The data used to pilot this method consisted of the interactions of three students during a 30-minute videotaped classroom activity devoted to learning how to use the light sensor. To accomplish the PSCE speech genre analysis, the transcripts of the words uttered by the students were broken down into tuple or duple (*ngrams*) word segments which were named in such a way as to retain temporal differentiation. For example, an excerpt from the transcription reads:

1	*I:*	*okay*
2	*J:*	*oh okay*
3	*I:*	*but we need a ruler to make it go far away*

Single word utterances were not considered for this analysis, therefore the utterance in line one was not included. An utterance of two words was included in the analysis if and only if the entire utterance consisted of two words. Since line two consists of only two words, it was included in the analysis. Any utterance of three or more words was then divided into multiple overlapping three-word segments and included for analysis. Line

three would have been divided into 9 segments: *but we need, we need a, need a ruler, a ruler to, ruler to make, to make it, make it go, it go far, go far away.* The line number of the original utterance was preserved along with each unique segment to retain temporality. The data set produced 2,627 unique ngram segments of text.

These trigram and bigram segments were then processed through the Java implemented Stanford log-linear parts-of-speech (POS) tagger (Toutanova, Klein, Manning, & Singer, 2003). POS taggers tokenize individual words and then utilize computational methods to assign a POS (such as noun, verb, conjunction, etc.) to each word. The Stanford POS tagger utilizes the Penn Treebank tag set (Santorini, 1990).

A report was then created of each unique POS tag string, along with the associated text segment and line number. Based on our domain expertise and the qualitative model of the troubleshooting cycle, we coded each POS tag string according to a PSCE coding system (Table 1). It is important to recognize that the same POS tag string may be assigned to different trigrams. We sought to interpret the "work" each utterance was doing by looking across the trigrams, which garnered the same POS tag string. This was a deliberative process undertaken by the first and third authors of this paper. The process included reading the trigrams in context and discussing each one – in this way, we inductively developed the coding system.

Table 1 – Problem Solving in Computational Environments Speech Codes

Diagnosis	**Query**	**Argumentation**
Evaluation	***Clarification***	***Group Regulation***
		Organization of tasks/roles
		Modal

	Activity Negotiation
Confirmation	***Content and Concepts***
	Programming Elements
	Comparative
	Explanation
	Building Elements
	Comparative
	Explanation
Puzzlement	***Problem Definition***
	Familiarization

Text segments were coded when 5 or more segments were associated with the same POS. Exceptions were made when a partial POS tag could be coded to the same code. For example, in table 2, the ngrams were all coded as *activity negotiation*. Therefore, any segments with a partial match of RB VBP were coded as *activity negotiation*. Some ngrams were not coded beyond being assigned the POS tag string, even if there were five instances or more. This was the case if it was clear that the ngram consisted of ideas belonging to two separate sentence clauses. In these instances, the ngram would be handled more appropriately in a different constellation with no break in meaning.

Table 2 – Example of POS to PSCE Coding Scheme

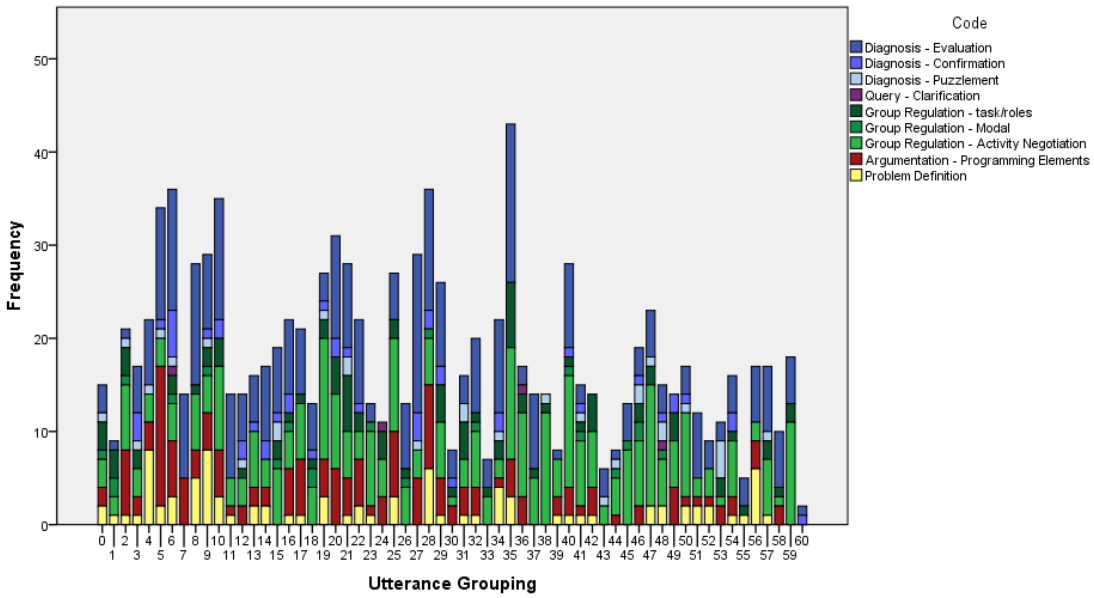
Text Segment	**POS Tag String**	**Tag Meaning**	**PSCE Code**
Now do the	RB VBP DT	Adverb, Verb, Determiner	Activity Negotiation
Now put it	RB VBP PRP	Adverb, Verb, Preposition	Activity Negotiation
Hey don't play	RB VBP RB VB	Adverb, Verb, Adverb, Verb	Activity Negotiation

In the results section, we present the preliminary results of our analysis. As noted above, we analyzed a 30-minute segment of data that we previously analyzed. Our goal was to identify, computationally, the qualitative model of the troubleshooting cycle we had previously found.

Results

Figure 1 is a stacked bar chart that indicates the number of PSCE codes generated per groups of ten utterances. For this analysis, the utterances were grouped temporally by their original location in the transcript. The number and type of coded ngrams that occurred in each group determined the height of the bar. This graph was then analyzed to identify clusters of high occurrences of coded ngrams, which might indicate sections within the transcription that were potentially generative to the research question.

Figure 1 – Stacked Bar Chart of Coded Ngrams



From this graphical representation, we identified seven clusters of utterances, which appear to be doing a lot of "work" in the troubleshooting cycle. We returned to each of these clusters within the context of the transcripts to closely analyze the discussion. While we did not specifically identify the 17 troubleshooting cycles found in the previous analysis, we were able to identify student conceptual development related to their understanding of the functioning of the light sensor. In the following, temporally sequential excerpts (tables 3 – 7), we demonstrate, with the data, the change in student's way of thinking and talking about the light sensor over a 30-minute period. The challenge students are trying to solve is to move the robot forward until it senses a black line on the floor, which will then trigger the robot to make a 90-degree turn and then move in reverse for one foot. The utterances that were most relevant to our interpretation are presented in bold text.

In the first excerpt (Table 3), students discuss programming the robot to move

forward with a timing element, which is not required. The students needed to program the motors to switch from forward motion to a turning motion, once the light sensor detected the black line on the floor; the light sensor should be programmed, but no timing code is required. The fact that the students were discussing how to program the forward movement of the robot, indicates they did not, yet, understand how the light sensor functions.

Table 3 – Cluster #1: Programming with Timing Elements

Line #	**Speaker**	**Utterance**
50	J	well I don't know why it's only doing the first of three that step let's do one cause it probably **has to be on time first**
51	S	**yeah that has to be on time**
52	J	then this has to do the light sensor
53	S	no do it
54	J	**this has to be this has to be time again**

A few minutes later, the students were still using the timing element, but they were starting to discuss the idea that, in order for the program to function properly, the sensor must see the black line, which they erroneously call the “black light”. This excerpt is presented in table 4.

Table 4 – Cluster #2: Students Realize Need to Program Sensor

Line #	**Speaker**	**Utterance**

95	S	**it's going forward for a time it's going to step one**
96	J	**going forward right for one second**
97	S	***yeah but it has to step for one second right but it has to go and touch the black line right yeah cause then the sensor***
98	S8	anybody lose a ring yeah I know
99	S	yeah that's good now let's send it
100	J	wait wait you have to wait now **I have to put this is time** now I have to put this back onto this stupid line
101	S	no not that that doesn't really have to oh yea that does yeah no
102	J	yes it does yes wait
103	S	oh look at this
104	J	it has to go backwards
105	S	oh it does
106	J	yeah it has to go back it has **to hit the light** then go backwards

In table 5, students continue to think they need to program both the light sensor and provide a timing element in order to move the robot forward.

Table 5 – Cluster #3: Continuing to Think Two Elements are Necessary

Line #	Speaker	Utterance
199	J	watch **I think I know what the problem is the light let's put it at thirty five** cause it's on so now send wait hey hey

		hey
200	I	this is going to be over a foot
201	S	it's going too slow
202	J	oh my god **no it doesn't want to go forwards cause there's no time limit**

In this next excerpt (Table 6), a few minutes later, the students articulate the idea that the robot will turn when it "sees" the "black light" and so, one cannot start the program from the "black light." This is the first indication in their discussion that the students understand that the sensor, if programmed correctly, may trigger another event.

Table 6 – Cluster #4: Sensor as triggering device.

Line #	Speaker	Utterance
280	J	**you're not supposed to put it on the black light that's why**
281	S	there you go
282	J	**it has to go away from the black light**
283	I	it's gonna follow your
284	S	hello black shoe
285	J	**it has to start from the you can't from the black light that's why it's not doing it**
286	S	black shoe
287	J	no let me watch now that's to get so far then **it has to see the black light and turn** but if it it doesn't do that well then it has to go turn and go backwards

After a few more minutes one of the students has a breakthrough in understanding how the light sensor functions. In table 7, the student explains the breakthrough to the other students – pointing to the displayed readings on the sensor, the student calls attention to how fluctuations in that number affect the movement of the robot. This is the essence of the sensor as a triggering device. At this point, the students are well on their way to solving the robotics challenge, having constructed an understanding of the light sensor as a computational device.

Table 7 – Cluster #6: Light sensor as computational device

Line #	Speaker	Utterance
400	J	turn it off turn it off turn it off turn it off **watch see this goes the lower the number it goes straight right and then when it changes to forty three it doesn't try to** (?) forty one see if it does it connect (?) just hold it in the air see it's the light it's it's (too big)
401	S	it's going forward
402	J	check the back line check that black line on the on the white paper the one in the middle
403	S	(?) something
404	I	right here
405	J	let's see
406	I	this one Sara
407	J	this one that one Sara not that one

408	I	this one this one
409	J	**not that one this one turn it off turn it off now do the reading of the black light do the reading of the black light what does it say?**

In these clusters, we see the students shift their thinking about the use of the sensor as a tool that controls the movement of the robotic device. In this way, the students are deepening their understanding through interaction with the tool itself.

In addition to looking at the clusters of utterances that generated a number of PSCE speech genre codes, we also looked at the segments of the data that generated fewer codes. In general the meaning of these segments was less clear and/or the utterances consisted of vocalizations that were not meaningful such as "huh, huh, huh" or "ha, ha, ha", or the discussion consisted of repetitions such as "wait, watch, watch, watch, watch."

Discussion

In this paper, we have presented our work-in-progress towards the development of computational methods for performing microgenetic analysis of discourse data. At this point in our work, we have made solid progress in addressing two issues that arose during development of this method. First, we addressed the issue related to the analysis of highly contextual and referential talk (featuring the heavy use of indexical terms, such as, this, there, here, etc.). Second, we addressed the issue of temporal analysis, which in this case refers to the need to meaningfully examine data over time.

We solved the first issue by drawing on our previous work (Sullivan, 2008, 2011;

Sullivan & Lin, 2012), in which we constructed qualitative models of student problem solving activity in computational environments. Based on these models of activity, we used a speech genre analysis approach to examine the type of "work" that an utterance may be doing in the activity. We then sought to map the verbalizations to activity. While we have not, yet, found a way to explicitly identify the 17 troubleshooting cycles we initially sought to computationally discover, we were able to identify analytically meaningful segments in the transcript from this method.

In terms of the segments of the transcript that generated fewer codes, we found the transcript to be replete with such sections. The highly contextualized nature of student talk in the face-to-face problem solving setting created situations where the conversation was frequently highly fragmented. It is important that the method be able to distinguish between moments when discoveries are made, and moments that lead up to the discovery.

Limitations and Next Steps

As this is a report on a work-in-progress, we acknowledge a number of limitations of this work that we are actively working to address. First and foremost, our decisions regarding the unit of analysis were relatively arbitrary (as opposed to theoretically derived). We may have just as profitably used any poly-gram as opposed to a trigram or bigram in analyzing the data. Our next step in this research is to adopt a systemic functional linguistics approach (SFL) (Halliday, 1993). In this approach, meaning making is analyzed at the level of the clause. SFL is a well-defined, language-based theory of learning. As such, it is a strong fit for the next phase of development of this work.

Another limitation is that we have only performed analysis on a small segment of a very large data set. While our results are hopeful, we, as yet, do not have proof that this

analysis will be useful on a larger data set. In that regard, we will perform this analysis on a longer transcript that covers a programming activity that spanned a two-day (three hour) period. Moreover, we will explore a new research question. Performing this analysis will assist us in evaluating the greater utility of the method.

Conclusion

In this paper we have presented our work-in-progress on the development of a computational method for performing microgenetic analysis of discourse data (co-present, collaborative problem-solving group talk). Our work is theoretically rooted in sociocultural theories of learning (Vygotsky, 1978), which emphasize the social nature of learning. We have built on Bakhtin's (1986) notion of speech genres in developing a method to investigate talk in the context of problem solving in computational environments. While still in a nascent period, this work has the potential to result in the development of a powerful educational research method that can be meaningfully adopted in a number of educational research settings.

References

Atkisson, M. & Wiley, D. (2011). Learning analytics as interpretive practice: Applying Westerman to educational intervention. *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge,* 117-121.

Bakhtin, M.M. (1986). The problem of speech genres. In V.W. McGee, Trans., C. Emerson & M. Holquist (Eds.). *Speech genres and other late essays* (pp. 60-102). Austin, TX: University of Texas Press.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge,* 110-116.

deLiddo, A., Shum, S.B., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics. *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge*, 23-33.

Ferguson, R. & Shum, S.B. (2011). Learning analytics to identify exploratory dialogue within synchronous text chat. *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge,* 99 – 103.

Fournier, J., Kop, R., & Sitlia, H. (2011). The value of learning analytics to networked learning on a personal learning environment, *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge,* 104-109.

Gašević, D., Dawson, S., & Seimens, G. (2015). Let's not forget: Learning analytics are about learning, TechTrends, 59(1), 64-71

Ginzburg, J. & Fernandez, R. (2010). Computational models of dialogue. In A. Clark, C. Fox & S. Lappin (Eds.), *Computational linguistics and natural language*

processing handbook. Oxford, UK: Blackwell.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* New York, NY: Harper & Row.

Halliday, M.A.K. (1993). Towards a language-based theory of learning. *Linguistics & Education*, 5, 93-116.

Kuhn, D. (2002). A multi-component system that constructs knowledge: Insights from microgenetic study. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 109-130). Cambridge, England: Cambridge University Press.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F.J. Damerau (Eds.) *Handbook of natural language processing,* 2nd ed., (627-666). Boca Raton, FL: Taylor & Francis Group.

Pressley, M. (1992). How *not* to study strategy discovery. American Psychologist, 47, 1240-1241.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision). Retrieved, January 21, 2015 University of Pennsylvania, Penn Engineering, Scholarly Commons Web site: http://repository.upenn.edu/cis_reports/570/?utm_source=repository.upenn.edu%2Fcis_reports%2F570&utm_medium=PDF&utm_campaign=PDFCoverPages

Sherin, B. (2013) A Computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences, 22*(4), 600-638, DOI: 10.1080/10508406.2013.836654

Shum, S.B., & Ferguson, R. (2012). Social Learning Analytics. *Educational Technology & Society*, *15* (3), 3–26.

Siegler, R.S., (2006). Microgenetic analyses of learning. In W. Damon & R. Lerner (Eds.). *Handbook of child psychology, 6th ed.,* (pp. 464-510). Hoboken, NJ: John Wiley & Sons.

Siemens, G., & Gašević, D. (2012). Guest editorial - learning and knowledge analytics. *Educational Technology & Society*, *15* (3), 1–2.

Sullivan, F.R. (2008). Robotics and science literacy: Thinking skills, science process skills, and systems understanding. *Journal of Research in Science Teaching, 45*(3), 373-394.

Sullivan, F.R. (2011). Serious and playful inquiry: Epistemological aspects of collaborative creativity. *Journal of Educational Technology and Society, 14*(1), 55-65.

Sullivan, F.R., & Lin, X.D. (2012). The ideal science student survey: Exploring the relationship of students' perceptions to their problem solving activity in a robotics context. *Journal of Interactive Learning Research, 23*(3), 273-308.

Sullivan F.R. & Wilson, N. (2015). Playful talk: Negotiating opportunities to learn in collaborative groups. *Journal of the Learning Sciences, 24*(1), 5-52. DOI: 10.1080/10508406.2013.839945

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL*, (pp. 252-259).

Varelas, M., Becker, J., Luster, B., & Wenzel, S., (2002). When genres meet: inquiry into

a sixth-grade urban science class. *Journal of Research in Science Teaching, 39(7),* 579- 605.

Varenne, H. (1998). Local construction and educational facts. In H. Varenne & R. McDermott (Eds.). *Successful failure: The school America builds.* Boulder, CO: Westview Press.

Vatrapu, R. (2011). Cultural considerations in learning analytics. *Proceedings of the LAK 2011: 1st International Conference on Learning Analytics and Knowledge,* 127-133.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes. Cambridge: Harvard University Press.

Wells, G.C. (1999). *Dialogic inquiry: towards a sociocultural practice and theory of education.* New York: Cambridge University Press.

Wertsch, J.V. (1979). From social interaction to higher psychological processes: A clarification and application of Vygotsky's theory. *Human Development, 22*, 1-22.

Wertsch, J.V. (1991). *Voices of the mind: A sociocultural approach to mediated action.* Cambridge, MA: Harvard University Press.

Acknowledgements: The research reported in this manuscript was supported by a grant from the National Science Foundation DRL #1252350.