Exploring Computerized Text Analysis to Predict the Validity of Students' Proof Construction

Elizabeth L. Pier, University of Wisconsin-Madison, epier@wisc.edu

Candace A. Walkington, Southern Methodist University, cwalkington@mail.smu.edu

Virginia Clinton, University of Wisconsin-Madison, vclinton@wisc.edu

Rebecca Boncoddo, Central Connecticut State University, boncoddo@ccsu.edu

Mitchell J. Nathan, University of Wisconsin-Madison, mnathan@wisc.edu

Abstract

We used two computerized text analysis tools, Coh-Metrix and the Linguistic Inquiry and Word Count (LiWC) software, to explore how language predicts undergraduates' mathematical proofs. We analyzed 240 transcripts of students' justifications of two mathematical tasks and investigated the significant correlations between various linguistic categories and incidence of correct proofs to explore critical language-based differences between valid and invalid proofs. Results showed distinct linguistic patterns for correct and incorrect proofs; multiple regression analysis showed that these categories were significantly predictive of the validity of a proof. This study contributes to our understanding of the role of language in mathematical proof production and demonstrates the vast potential for text analysis tools to elucidate linguistic aspects of students' emerging disciplinary discourse practices.

Keywords: Cognitive Processes/Develoment, Language Processes, Student Cognition

Objectives

Research has shown an important link between language and cognitive processes (e.g., Pennebaker & King, 1999; Vygotsky, 1987). While researchers and educators have emphasized the importance of language development in cognition, few analyses have explicitly explored how certain types of language use by students may predict success on specific educational achievement outcomes. Given the complexity and diversity of human language use, analyses of this kind can be onerous and time-consuming (e.g., Ericsson & Simon, 1993). The recent emergence of computerized text analysis programs has made examining language use more efficient and practical. Two text analysis tools in particular, the Linguistic Inquiry and Word Count software, or LIWC, and Coh-Metrix, have recently been used to investigate the relationship between language and educational outcomes. Coh-Metrix was developed to assess the difficulty of a text given to a student (Graesser, McNamara, Louwerse, & Cai, 2004), but has more recently been applied to investigating the nature of student *speech*, such as with think-alouds (Jeon & Azevedo, 2007). LIWC, developed for clinical and health-related research (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), has been used to analyze patterns of word use associated with inferences made while reading (Clinton, Carlson, & Seipel, 2013) and the quality of student responses when using an intelligent tutoring system (Williams & D'Mello, 2010). This study aims to apply these text analysis tools to explore the quality of student-generated mathematical arguments.

Theoretical Framework

The construction of valid mathematical arguments, or "proofs," is central to both the professional domain of mathematics and mathematics education (Schoenfield, 1994); it is the means by which mathematicians communicate key ideas and novel understandings to each other.

The Common Core Standards (2010) identify constructing viable arguments and critiquing the reasoning of others as a key component of mathematical practice. However, adopting practices of mathematical proof is difficult for students across grade levels (Dreyfus, 1999; Knuth, Choppin, & Bieda, 2009), emphasizing the need for research into the practices of successful proof generation.

One way to investigate the nature of successful proof is to examine the language students use when constructing proofs. Mathematical proofs can be thought of as a specific kind of disciplinary discourse practice (e.g., Gee, 2007; Gresalfi & Cobb, 2006) and in K-12 classrooms, proofs often take spoken—rather than formal, written—forms (Healy & Hoyles, 2000). When constructing narrative proofs, individuals often communicate a logical and persuasive chain of reasoning using descriptive language and verbal inference. However, simply providing a descriptive narrative does not necessarily result in a correct proof. Valid mathematical arguments, as conceptualized by Harel and Sowder (2005), have several necessary characteristics: They are general, meaning the argument is shown to be true for all cases; they involve operational thought with a progression through goals and subgoals; and they utilize logical inference in which conclusions are drawn from valid premises.

Examining linguistic differences amongst students' attempts at constructing valid mathematical proofs may elucidate some of the basic features of successful proof generation. We investigate the following research question: How are different linguistic measures (e.g., pronoun use, verb tense, etc.) from LIWC and Coh-Metrix associated with valid, verbal mathematical arguments across two tasks?

Methods

Participants (*N* = 120) were undergraduates (*M* age = 19.2 years; 51% female) enrolled in a psychology course at a large Midwestern university. Two tasks were presented that prompted students to verbalize proofs from two distinctly different mathematical domains: one relating to planar geometry, the other to an inference about parity (i.e., an odd/even pattern) in a physical system. Students were instructed to read the following prompts (in randomized order) aloud and to think aloud as they provided their answer:

1. *Mary came up with the following conjecture: "For any triangle, the sum of the lengths of any two sides must be greater than the length of the remaining side." Provide a justification as to* why *Mary's conjecture is true or false.*
2. *An unknown number of gears are connected together in a chain. If you know what direction the first gear turns, how could you figure out what direction the last gear turns? Provide a justification as to* why *your answer is true.*

Student responses were videotaped and uploaded into Transana, a software program that allows for transcribing and analyzing video data (Woods & Fassnacht, 2012). Undergraduate research assistants transcribed each participant's verbal responses to both prompts, and our team segmented them into 240 separate transcripts. We coded each transcript as "correct" or "incorrect" based on Harel and Sowder's (2005) definition of valid proofs. Inter-rater reliability for two coders was 85%. Overall 40.8% and 50% of participants provided valid proofs to the gear and triangle prompts, respectively.

Data Sources

We analyzed the 240 transcripts using LIWC (Pennebaker, Booth, & Francis, 2007) and Coh-Metrix (Graesser et al., 2004). LIWC comprises various "dictionaries" of words in different categories, such as auxiliary verbs or “leisure” words; its output consists of the percentage of words used from each dictionary (Tausczik & Pennebaker, 2010). Coh-Metrix analyzes cohesion relations and measures of language, text, and readability, such as word frequency, parts of speech, and syntactic complexity (Graesser et al., 2004). Using both tools together allows us to analyze the technical aspects of the language and readability gathered from Coh-Metrix, as well as the content of the language participants use from LIWC.

For each participant’s proof, we calculated Pearson’s correlation coefficient between their score in each category from LIWC and Coh-Metrix, and whether the participant generated a valid proof (coded as 0/1). To identify components of valid proof production across the two tasks, we selected the categories that showed significant correlations that were in the same direction for both tasks. In order to interpret the significant correlations, we systematically investigated the twenty transcripts that scored highest on each category and the twenty that scored lowest to determine what aspects of the transcripts resulted in high or low scores on these components. Finally, to assess the predictive validity of the identified linguistic categories for our entire data set, we entered the selected categories as predictors into a multiple logistic regression model for all 240 transcripts, with the correctness of proof as our outcome variable.

Results

Correlational Analysis

Eleven categories from LIWC and Coh-Metrix were significantly correlated in the same direction with valid proofs for both tasks. Table 1 lists and describes each of these categories and provides the Pearson correlation coefficient for each. Table 2 provides examples from transcripts that scored high and low on each of these categories. Our exploration of the transcripts allowed us to determine the patterns of words and phrases that were associated with high or low scores on a LIWC or Coh-Metrix measure. We found that there were three major linguistic areas associated with correct or incorrect proofs: the use of conditional statements (i.e., "if… then"), word specificity, and the use of self-conscious phrases such as "I don't know." The following section explains how each significant correlation fits into these three categories.

Table 1

List and Descriptions of Categories Significantly Correlated with Correct Proofs

Source	Category Code	Category Name	Category Description	Pearson Correlation Coefficient	
				Triangle	Gear
LIWC	Pronoun	Pronoun use	Incidence rate of pronouns	–0.25	–0.29
Coh-Metrix	WRDPRO	Pronoun incidence	Incidence rate of pronouns	–0.20	–0.32
LIWC	I	I	Incidence rate of the first person singular pronoun "I"	–0.26	–0.50
Coh-Metrix	WRDPRP 1s	First person singular pronoun incidence	Incidence rate of the first person singular pronoun (i.e., "I")	–0.26	–0.40
LIWC	Present	Present tense	Incidence rate of the present tense	–0.28	–0.33
LIWC	Insight	Insight words	Incidence rate of words in the "Insight" dictionary (e.g., think, know, understand, question)	–0.30	–0.41
LIWC	Discrep	Discrepancy words	Incidence rate of words in the "Discrepancy" dictionary (e.g., could, should, would)	0.26	0.20
Coh-Metrix	CNC Temp	Temporal connectives	Incidence rate of temporal connectives (e.g., before, after, next, then)	0.26	0.24
Coh-Metrix	SYNNP	Mean number of modifiers per noun phrase	Mean number of adjectives and adverbs for each noun phrase	0.25	0.22
Coh-Metrix	LDTTRc	Lexical diversity, type-token ratio, content word lemmas	Ratio of the number of unique words ("types") divided by the number of times that word occurs ("token"); higher ratio indicates more unique words; counts only content words	–0.28	–0.50
Coh-Metrix	LDTTRa	Lexical diversity, type-token ratio, all words	Same measure as LDTTRc, but for all words in the text	–0.25	–0.48

Table 2

Examples of Transcripts With High and Low Scores for Each Significantly Correlated Category

Category	Examples: High Scores	Examples: Low Scores
Pronoun & WRDPRO	"Um, you could draw out the gears and follow the rotation of them, um, to find out which, which it turns, and it's true because um, because you can see it um, on the, if you, if you draw it out as a diagram."	"Um okay so if one gear is turning uh one way then the next gear would be turning the other way so if there's an even number of gears, the last gear'll be turning the same way the first gear is and if there's an odd number of gears, then the last gear will be turning the opposite way of the first gear."
I & WRDPRP1s	"Well, I know for sure it's true 'cause I've learned that before, so for any triangle, the sum of the lengths of any two sides–it's like the hypotenuse always has to be the longest side. I don't really know why though."	"Mary's conjecture is true, because if the one side is long–is longer than the sum of the other two sides, then the other two sides won't be able to touch at the top. And it won't be a triangle."
Present	"What direction the first gear turns. I think the last gear should turn in the same direction as the first one. I don't know. Yeah, probably I guess. I guess the last direction should turn in the same direction. So."	"Okay so every other one is gonna turn the opposite way, so if one's going forward, one's going down, next one's going forward, next one's down, so the last one will be going depending on if it's odd or even forward or backwards."
Insight	"Because it is... Um... I'm just thinking that it- just because it is. I don't really know how to explain it. Because it's true for the triangle? I don't know."	"Um this is true and I forget the name of the actual conjecture but in geometry two sides of the triangle have to be added up to be longer than the third side because if they're shorter, then the triangle would not connect at all. It would be an open triangle."
Discrep	"Ok. Uh. Could find the pattern, it should go every other so you go with–you would count the number of gears and if it was a even number it would turn the same way as the first gear. If it was odd it would turn the same as the last gear."	"Um, you–I guess gears have to work the same way. Wait, no, gears have to work opposite ways per link so, depending on which one–how many are in the line. But it doesn't show how many are in the line, so. I don't know the answer, sorry."
CNCTemp	"Um, if you have many gears connected together and the first gear turns one way, then it would cause the other gear to turn the opposite way…So if it's an odd number, it'd be the same as the first one, but if it's even then it'd be different than the first one."	"So if you think about a gear. If you have a gear turning to the right, I guess the last gear would turn to the right too, 'cause they kind of interact with each other? I don't really–Yeah, I guess so. I would just say the gears must turn the same direction."
SYNNP	"Um the triangle has, has three lengths and if two sides, if the sum of the length of two sides equal to the remaining side, then these three lengths uh form a straight line which is nott a triangle anymore so the two sides must be greater than the remaining side."	"Well, it doesn't necessarily have to be shorter than the remaining side. Oh, the base you mean? Oh, okay. Greater than the lengths of the remaining side. I think it's false. It can be just– the sides can be equal to each other."
LDTTRc & LDTTRa	"Um, well I would assume- I don't know anything about gears and chains but, um, the–I don't know if it's on a wheel either, okay. If you know the direction of the first gear, I would assume they would all go in the same direction because in order for the chain to be linked and in a complete, um, like smooth circle, or something."	"So if the first gear is turning that way, then it's turning the next gear the opposite way, so if it's an even number of gears, it's turning the opposite direction as the first one, if it's an odd number of gears, it's turning the same direction as the first one? First one going that way, so next one would go that way, third one would go that way. So yeah. Odd number of gears, it's same way, even number it's the opposite way."

Characteristics of Valid/Invalid Proof

Conditional statements. Two of the significant correlations were driven by the use of conditional statements: discrepancy words and temporal connectives. Discrepancy word incidence, such as “if,” “would,” and “should,” was positively correlated with correct proofs, in that participants who gave correct proofs made more statements about what would or should happen in a variety of examples related to the conjecture (see row 5 in Table 2). Temporal connective incidence, such as “before,” “after,” and “then,” also was positively associated with correct proofs, in that participants who gave correct proofs made more statements with the word “then,” as part of the “if… then” format of conditional statements (see row 6 of Table 2). Thus, both of these findings were driven by the increased use of conditional statements by participants who gave valid proofs. The use of conditional statements reflects the logical inference characteristic of valid proofs from Harel and Sowder (2005).

Word specificity. Five of the significant correlations were related to participants’ word specificity: pronoun use from LIWC and Coh-Metrix, the mean number of modifiers per noun phrase, and the type-token ratio for content word lemmas and all words. Pronoun use was negatively correlated with correct proofs; using the words “it,” “they,” and “this” instead of more precise nouns is associated with incorrect proofs (see row 1 of Table 2). The number of noun modifiers, however, was positively correlated with correct proofs, in that participants who gave valid proofs used more adjectives and adverbs (see row 7 of Table 2). In addition, using fewer unique words was associated with correct proofs. We found that a higher type-token ratio, which indicates more unique words in a text, was negatively associated with correct proofs, both for content word lemmas[1] and all words. The final row in Table 2 shows that transcripts with high ratios tended to include many different, unrelated words, lacking continuity of ideas. Proofs with low ratios, however, included the repetition and subsequent transformation of key mathematical terms that comes with goal-directed inference (Harel & Sowder, 2005). Variety in word use has also previously been found to be positively associated with poor quality responses in intelligent tutoring systems (Jackson & McNamara, 2012). Thus, overall, correct proofs were associated with more specific and descriptive words, as indicated by using fewer pronouns and more adjectives and adverbs, as well as by repeating key mathematical terminology related to the conjecture.

Self-conscious statements. The remaining findings were driven by the presence of self-conscious statements participants made during the think-aloud process—specifically, verbalizing phrases such as “I don’t know,” or “I’m not sure.” The correlations in this category include first-person singular pronouns from LIWC and Coh-Metrix, the incidence of “insight” words, and the incidence of the present tense. Using first-person singular pronouns (i.e., “I”) was negatively correlated with correct proofs, because referring to oneself appeared to indicate an expression of self-doubt or confusion. The incidence of “insight” words, such as “know,” “understand,” and “think” was negatively correlated with valid proofs, since participants frequently used these terms when expressing self-doubt, as in “I think…” or “I don’t know.” Finally, the present tense was negatively correlated with correct proofs, because participants spoke in the present tense most when expressing these self-conscious statements, and not in the future or conditional tense. Thus, participants’ self-aware verbalizations of doubt or confusion manifested themselves in these four categories.

---

[1]Lemmas include nouns, most verbs, adjectives, and adverbs, but not function words such as articles, prepositions, and pronouns.

Multiple Regression Analysis

To test the predictive validity of these findings, we entered the 11 significant correlations into a multiple regression model for all 240 transcripts, with correctness of proof as our outcome variable and participant identity as a random effect. Three categories that were significant when included in the model at $\alpha = 0.05$ were discrepancy words, type-token ratio for content word lemmas, and present tense. Table 3 provides the output from the regression analysis. Each of these categories relates to one of the three characteristics of correct versus incorrect proofs: discrepancy words indicate more conditional statements, type-token ratio for content words indicates more repetitive, purposeful word choice, and present tense indicates more self-conscious statements. Thus, using more conditional statements and more repetitive, purposeful words predicts a correct proof, while using more self-conscious statements predicts an incorrect proof.

Table 3

Output from Multiple Regression Analysis

Fixed effects	Coefficient	Std. Error	Z	Sig
(Intercept)	6.90316	1.70345	4.052	***
Discrep	0.30741	0.09890	3.108	**
Present	-0.15890	0.06480	-2.452	*
LDTTRc	-10.95285	2.56814	-4.265	***
CNCTemp	-0.02669	0.01000	-2.668	**
Percent reduction in model deviance due to fixed effects: 13.98%				

Note. $^*p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

Implications

Here we have demonstrated that recently developed text analysis tools applied to speech can have powerful and innovative applications in educational research (Baker & Yacef, 2009). These findings are preliminary, but promising. To that end, we are investigating how language structures differ for ESL participants, and whether prompting students to use certain word categories may improve their proof practices. These applications demonstrate the flexibility of these computer-based text analysis tools for addressing both practical questions (e.g., How can a teacher support English language learners as they engage in mathematical discourse?) and theoretical questions (e.g., How does language mediate mathematical cognition?). In conclusion, text analysis tools can provide powerful insights into the linguistic patterns of mathematical discourse during proof activities, as well as various other types of academic discourse, and we share our initial foray in order to highlight what is possible with programs such as LIWC and Coh-Metrix.

References

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Clinton, V., Carlson, S., & Seipel, B. (2013, July). *Patterns of Word Use Associated with Valid Inference Generation.* Paper presented at the Annual Meeting of the Society for Text and Discourse in Valencia, Spain.

Dreyfus, T. (1999). Why Johnny can't prove. *Educational Studies in Mathematics, 38*(1-3), 85-109.

Gee, J. P. (2007). *Social linguistics and literacies: Ideology in discourses*. Routledge.

Graesser, A., McNamara, D., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, (36), 193–202. Retrieved from http://link.springer.com/article/10.3758/BF03195564

Gresalfi, M. S., & Cobb, P. (2006). Cultivating students' discipline-specific dispositions as a critical goal for pedagogy and equity. *Pedagogies*, *1*(1), 49-57.

Harel, G., & Sowder, L. (2005). Toward comprehensive perspectives on the learning and teaching of proof. In F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning*, NCTM.

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, *31*(4), 396–428.

Jackson, G. T., & McNamara, D. S. (2012). Applying NLP metrics to students' self explanations. In P.M. McCarthy & C. Boonthum (Eds.), Applied natural language processing: Identification, investigation, and resolution (pp. 261-264). Hershey, PA: IGI Global.

Jeon, M., & Azevedo, R. (2007). Analyzing human tutorial dialogues for cohesion and coherence during hypermedia learning of a complex science topic. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1127–1132). Austin, TX: Cognitive Science Society.

Knuth, E., Choppin, J., & Bieda, K. (2009). Middle school students' production of mathematical justifications. In D. Stylianou, M. Blanton, & E. Knuth (Eds.), *Teaching and learning proof across the grades: A K-16 perspective* (pp. 153-170). New York, NY: Routledge.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005). Coh-Metrix version 1.4. Retrieved July 9, 2013, from http//:cohmetrix.memphis.edu.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). Common Core State Standards: Mathematics. Washington D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers. Retrieved from http://www.corestandards.org/the-standards/mathematics.

Pennebaker, J. W., Booth, R. J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007* [LIWC manual]. Austin, TX: LIWC.net.

Pennebaker, J., & King, L. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social …*, *77*(6), 1296–1312. Retrieved from http://psycnet.apa.org/journals/psp/77/6/1296/

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting Final Course Performance From Students' Written Self-Introductions: A LIWC Analysis. *Journal of Language and Social Psychology*, (February). doi:10.1177/0261927X13476869

Schoenfeld, A. (1994). What do we know about mathematics curricula? *Journal of Mathematical Behavior*, *13*(1), 55–80.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi:10.1177/0261927X09351676

Vygotsky, L. S. (1987). *The collected works of LS Vygotsky: Volume 1: Problems of general psychology, including the volume Thinking and Speech* (Vol. 1). Springer.

Williams, C., & D'Mello, S. (2010). Predicting student knowledge level from domain-independent function and content words. *Intelligent Tutoring Systems*, 62–71. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-13437-1_7

Woods, D., and Fassnacht, C., (2012). Transana v2.52. http://transana.org. Madison, WI: The Board of Regents of the University of Wisconsin System.