Computerized Text Analysis: Assessment and Research Potentials for Promoting Learning

Hee-Sun Lee (chair), The Concord Consortium, hlee@concord.org
Danielle McNamara (discussant), Arizona State University, dsmcnama@asu.edu
Zoë Buck Bracey, BSCS Science learning, zbuck@bscs.org
Christopher Wilson, BSCS Science learning, cwilson@bscs.org
Jonathan Osborne, Stanford University, osbornej@stanford.edu
Kevin C. Haudek, Michigan State University, haudekke1@msu.edu
Ou Lydia Liu, Educational Testing Service, lliu@ets.org
Amy Pallant, The Concord Consortium, apallant@concord.org
Libby Gerard, University of California, Berkeley, libbygerard@berkeley.edu
Marcia C. Linn, University of California, Berkeley, mclinn@berkeley.edu
Bruce Sherin, Northwestern University, bsherin@northwestern.edu

Abstract: Rapid advancements in computing have enabled automatic analyses of written texts created in educational settings. The purpose of this symposium is to survey several applications of computerized text analyses used in the research and development of productive learning environments. Four featured research projects have developed or been working on (1) equitable automated scoring models for scientific argumentation for English Language Learners, (2) a real-time, adjustable formative assessment system to promote student revision of uncertainty-infused scientific arguments, (3) a web-based annotation tool to support student revision of scientific essays, and (4) a new research methodology that analyzes teacher-produced text in online professional development courses. These projects will provide unique insights towards assessment and research opportunities associated with a variety of computerized text analysis approaches.

Purpose, structure, and significance

Written texts play an important role in the process of learning as they are used for learners to read inscribed knowledge as well as to express their ideas and understanding. Many discipline-specific practices are carried out by means of texts, e.g., explanation, argumentation, and communication. Written texts are used for the purpose of assessing learners or for the purpose of generating theories of learning. For the last few decades, the analysis of written texts relied on a relatively stable, circumscribed, post hoc set of qualitative data analysis methods, e.g., Miles and Huberman (1994). But we are now entering a new era, one in which our toolkits have been expanded by a quite different set of computational techniques, rooted largely in methods drawn from machine learning. For researchers in the CSCL community, computational text analysis can play a number of roles, which fall into two main categories. First, forms of computational text analysis can be embedded within computer-based learning environments that we design. In this context, computational text analysis can allow these environments to better guide learners, as well as to provide better feedback, both to learners and their teachers. Second, computational text analysis can be employed more directly as a tool for researchers, as a means of analyzing data in support of research goals. In the context of CSCL research, the text for this latter category can come from multiple sources. It might be generated by learners as they interact with a computer-based learning environment, or it might be developed in alternative ways, such as when learners are interviewed at the end of a computer-based intervention and the interviews are transcribed.

The purpose of this symposium is to shed light on these two different ways to use computerized text analyses in the current work of four research projects as shown in Table 1. The first and fourth presentations address the research aspect of computerized text analyses while the second and third presentations address practical applications to support student learning in the classroom. The first two presentations address texts written by middle school students while the third presentation addresses those written by high school students. Texts in the fourth presentation were created by K-2 school teachers who were taking online professional development courses. Computerized text analyses were conducted by different software packages including c-rater ML™ developed by Educational Testing Service and Tactic Text. The first presentation concerns an important issue of whether automated scoring models provide equal opportunities for students with different language backgrounds. The second presentation shows a seamlessly integrated formative assessment system that can help students' revision of written arguments in real time. The third presentation tests a formative assessment system in two task conditions to optimize design features. The fourth presentation uses computerized text analysis to find emergent

patterns without the researcher's implicit bias.

The session moderator, Dr. Hee-Sun Lee, will briefly describe the purpose and organization of the symposium, followed by the introduction of the speakers (~5 minutes). All presenters will introduce their computerized text analysis methods and applications in their learning contexts (a total of ~60 minutes, ~15 minutes per presentation). Dr. Danielle McNamara at Arizona State University will lead a discussion focusing on the challenges and complexities involved in research and development efforts for collaborative learning settings (~10 minutes). Then, the audience will have opportunities to interact with presenters as well as the discussant for synthesis of ideas presented in the symposium (~15 minutes). The topic of this symposium is critical for setting the next research agenda in designing and testing productive language-intensive learning environments where computerized text analyses can add unique instructional values.

Table 1: Summary of computational text analysis method, learning context, and research focus for each presentation

Presentation	Computational Text Analysis	Research/Learning Context	Research Focus
1	Automated scoring model development	Assessment of middle school students' scientific argumentation	Examination of potential linguistic bias in the automated scoring models as compared to human scoring
2	Automated content scoring using c-rater ML; machine learning-based natural language processing	Scientific argumentation tasks embedded in an online Earth science curriculum module	Formative assessment function: student performance diagnosis using automated scoring and real-time, targeted feedback provided to students
3	c-rater ML to provide content guidance and Annotator to provide revision strategy guidance	Short essays embedded in web-based unit on plate tectonics	Design of automated guidance to strengthen student agency and ability to make constructive revisions to short essays
4	Topic modeling, simple counts of word usage	Online professional development environment for K-2 math and science	Demonstration of a computational environment designed to support the work of qualitative data analysis

Presentation 1: Using automated scoring to assess argumentation while minimizing linguistic bias

Zoë Buck Bracey, Christopher Wilson, Jonathan Osborne, and Kevin C. Haudek

The automated scoring project in this study is carried out by an interdisciplinary team of science educators, cognitive scientists, and computational linguists. The goal of the project is to develop automated scoring models and corresponding multidimensional science assessment items aligned with the scientific argumentation practice identified in the Next Generation Science Standards (NGSS Lead States, 2013) for grades 6-8. The project builds upon the learning progression-based scientific argumentation assessment work by Osborne et al. (2016). This project's automated scoring model development addresses two concerns: (1) whether we can develop automated text scoring models for students' explanation and argumentation responses that are comparable to expert human scoring and (2) whether or not the degree to which the computer-based algorithmic text scoring is more or less biased against English Language Learners (ELLs) than human scoring of the same data (relative linguistic bias). As machine scoring of open-ended responses is expected to permeate into the classrooms internationally, combined with the current trend of classrooms having more and more culturally diverse students as they bring new languages into the classroom makeup, it is imperative to ask these questions in order to monitor the potential impact of the use of automated text scoring on the assessment of students from non-dominant cultural and linguistic backgrounds who are often underserved by educational reforms.

Written assessments have the capacity to expand the ways in which participants can express their competences (Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001), but only when our interpretation "listen[s] past English fluency" to evaluate students' science ideas (Moschkovich, 2007). In other words, teachers and other educators scoring assessments can learn to consciously counteract their biases associated with the linguistic patterns of students who are learning English. Computers may not have that ability. However, computer-based models have been shown to take on the biases of the humans who scored the data that were in turn used to train them. This suggests that there may be steps we can take to reduce the bias associated with the computerized text scoring models. This study addresses the research question: Can we develop automated computer scoring models of students' explanations that are unbiased to variations in English fluency?

We are currently collecting data in school districts across Northern California. After a short lesson on argumentation in science, students answer a set of argumentation-related science items on an online platform (i.e., Qualtrics) designed by the project's interdisciplinary team. The assessment is written in English. ELL students within our sample are asked to produce textual responses in English. ELL status from our sample will be determined by the state's English language proficiency scores. The sample size of this study is approximately 1,000 students, of whom about 15% are ELLs. The data are scored by three human scorers with expertise in the science content and the scientific argumentation practice. Human-scored responses are used to develop automated text scoring models. In addition, these human-scored responses are analyzed using Facets software to determine relative bias between the biased human scorers, the less-biased human scorers, and the two computer models based on level of English proficiency. The Facets software extends the objective measurement principles of Rasch modeling (Rasch, 1966) using generalizability theory to apply to more complex areas such as judged performances (Solano-Flores, 2006). This analysis is used to determine the relative bias of the human versus the computers, as well as the relative bias of the human scorers on the overlapped data sets, and to train the machine learning algorithm on selected sets of less and more biased scorers.

This research contributes to the field by establishing not only the feasibility of creating high-quality automated scoring-based assessments for scientific argumentation, but also examining the degree to which the automated scoring models for such assessments are more or less biased against responses written by students who are learning a new language than human scoring. Researchers, both within science education and in the education community at large who are considering automated scoring technologies, need to have productive conversations about how to diagnose, monitor, and counteract bias. This study provides evidence to inform the nature and the direction of those conversations. Science teachers and curriculum designers should be aware of the potential risks associated with bringing automated text scoring into the classroom as formative or summative assessment methods. For researchers, it is important to critically examine the ways in which the risks and potentials can manifest while automated text scoring tools are in place so that these automated text scoring technologies can be leveraged to provide opportunities for multilingual students to express competence while learning science.

Presentation 2: Formative assessment of scientific argumentation practice enabled by automated scoring

Ou Lydia Liu, Hee-Sun Lee, and Amy Pallant

This presentation addresses supporting secondary school students' revision of scientific arguments when students' claims and explanations about scientific phenomena are based on imperfect data. Students need adequate support so that they formulate strong written scientific arguments, particularly when uncertainty arises due to theoretical, methodological, measurement-related, analytical, and interpretative limitations associated with investigations. We developed HASbot, a formative feedback system that (1) diagnoses students' written arguments through automated-scoring technologies, (2) provides instant feedback on student performance, and (3) offers a teacher dashboard for teachers to monitor class-level performance in real time.

HASbot is integrated in an online curriculum module that explores freshwater availability and sustainability. There are eight scientific argumentation tasks in this water module. In writing scientific arguments, students submit open-ended responses that explain how their data support claims and how limitations of their data affect the uncertainty of their explanations. Students are expected to develop scientific reasoning that explains their claims based on evidence (McNeill, Lizotte, Krajcik, & Marx, 2006), and articulate critical thinking that examines limitations of the investigations (Allchin, 2012; Lee et al., 2014). Figure 1 shows a set of four prompts that elicited a student's scientific argumentation responses. In this study, HASbot evaluated these responses in real time using c-rater-ML™, a natural language processing scoring engine that uses machine learning methods to extract and weight textual features relevant for scoring developed by Educational Testing Service. Table 2 lists Human-Human and Human-Machine agreements measured in Quadratic Weighted Kappa values for all automated scoring models associated with the eight tasks. HASbot returns scores with feedback to guide student revisions. See Figure 1 for how formative feedback was provided to a student after submission (note the colored bar and text below the bar in the figure).

Data were collected from 343 middle and high school students taught by nine teachers across seven states in the United States. Students took the uncertainty-infused scientific argumentation test developed by Lee et al. (2014) before and after the water module. Students' initial formulation and revision of eight scientific argumentation tasks in the module were logged. We analyzed these data to investigate how students' utilization of HASbot feedback impacted their ability to formulate scientific arguments related to freshwater systems. We also collected video data that captured how students worked together with HASbot. We analyzed videos of 14

groups of students working on the first scientific argumentation task to identify affordances and limitations of the current design of HASbot.

Paired t-tests indicate that students made statistically significant gains from pre-test to post-test, effect size = 1.52 SD, $p < 0.001$. Our linear regression analysis of student posttest scientific argumentation score indicates that students' interaction with the HASbot system significantly contributed to the post-test score after controlling for gender, English language learner status, and prior computer experience. HASbot helped students (1) determine what information to include and how to revise argument responses, (2) motivated to revise with feedback from a friendly, non-judgmental robot, (3) frame how to talk about uncertainty as part of argumentation, and (4) engage more deeply with the content and the data. HASbot constrained students because (1) false positive machine scores hindered students' revision efforts, (2) some students had difficulty interpreting the feedback statements, and (3) repetitive feedback statements irritated some students when their revisions did not yield improved scores. We discuss implications for supporting scientific argumentation involving uncertainty and developing a feedback system based on automated text scoring.

Figure 1. HASbot feedback example.

Table 2: Quadratic Weighted Kappa (QWK) values for human-human and human-machine agreements

Task	Students (n)	Explanation		Uncertainty Attribution	
		Human-Human	Human-Machine	Human-Human	Human-Machine
1. Trap	935	0.90	0.78	0.90	0.86
2. Bedrock	522	0.94	0.92	0.94	0.89
3. Pumice	890	0.96	0.85	0.93	0.91
4. Aquifer	717	0.95	0.90	0.95	0.88
5. Vernal	709	0.94	0.84	0.92	0.85
6. Impact	704	0.86	0.70	0.93	0.86
7. Runoff	638	0.94	0.83	0.87	0.85
8. Supply	457	0.93	0.85	0.96	0.87
Average		0.93	0.84	0.93	0.87

Presentation 3: Critique essay by peer or self to learn to revise in science

Libby Gerard and Marcia C. Linn

In prior research, we used c-rater ML™ to develop automated scoring models for students' short essays on a 0-5 knowledge integration rubric that assesses the connections among normative ideas about science (Liu, Rios, et al., 2016). To work with autoscores, knowledge integration (KI) guidance was developed to help students move up one score level in the KI rubric. Even though this autoscore-based, adaptive KI guidance was more effective in improving students' knowledge integration abilities than other types of guidance typically used in middle school

classrooms (Gerard, Ryoo, et al., 2015), many students still struggled to use the KI guidance to revise their essays. Some students tacked ideas on to the end of responses rather than thoroughly integrating the new information, leading to superficial edits while others did not revise their essays at all (Gerard, Linn, Madhok 2016). These findings were not unexpected as research shows that students tend to add disconnected ideas, fix mechanical errors, or make superfluous edits rather than modifying connections among all ideas (Fitzgerald, 1987). When confronted with contrasting evidence, students tend to ignore the evidence and restate their own perspective (Mercier & Sperber, 2011), consistent with confirmation bias (Clark & Chase, 1972).

To promote integrated revision, we developed the Annotator. The Annotator provides students with an interactive model of the revision process (see Figure 2). Students place pre-written or self-constructed labels on sections of an essay to suggest areas for change or improvement. Selecting the relevant labels and placing them in the essay encourages distinguishing of key ideas and the integration of new and prior knowledge, rather than novice practices of tacking on disconnected information. In the initial Annotator design, students critiqued a fictional peer's essay containing common ideas that required revision (Gerard et al., 2016). We added the Annotator guidance to the adaptive KI guidance to strengthen the quality of student essay revisions.

This study compared peer- and self-annotations to determine optimal design features. We compared the initial Annotator design involving peer annotation to a modified version involving self-annotation that was intended to strengthen student agency in revision. We hypothesized that instantiating the student's own essay in the Annotator would encourage students to view their essay as a scientific product and attend more carefully to each expressed idea and the connections among them. Flower and Hayes (1980) showed that when students succeeded in analyzing the structure and argument of their essay they were capable of making valuable revisions to their reasoning. When students moved to the next step, they were randomly assigned to one of two conditions: (a) annotate their essay or (b) annotate Sara's (see Figure 2). In the "annotate own" version, the student's essay was imported into the Annotator. In the "annotate Sara's" version, an essay by a fictitious peer was pre-loaded in the Annotator. In both conditions, students used labels to address gaps or inaccuracies in the essay; reviewed their essay to revise; and then had one opportunity to receive adaptive KI guidance and revise again.

The study was conducted in two schools with four teachers and their 513 students who used the WISE "*Plate Tectonics: Why are there more earthquakes, volcanoes and mountains on the West Coast?*" One school served primarily white, middle-class students (School A, N = 332 students, 37% non-White, 11% free/reduced price lunch); the other school served primarily non-White, low-income students (School B, N = 181 students, 94% non-White, 89% free/reduced price lunch). Data included students' logged initial and revised embedded and pre/post-test essays, student annotations, interviews, and classroom observations. Essays were scored using the five-point knowledge integration rubrics (Liu, Lee, et al., 2008); annotations from one teacher in each school were scored using a 0-3 rubric assessing engagement and accuracy.

The Annotator plus adaptive KI guidance supported students in both conditions to successfully critique and revise their essays during inquiry. Although there were substantial school differences, the rate of revision was the same between the two conditions (School A: 88% revised; School B: 89%). Students significantly improved their essays in both conditions during revision (School A: $t(164) = 7.57$, $p < 0.001$; School B: $t(89) = 3.50$, $p < 0.001$). In School A, there was a marginal effect for condition when controlling for initial essay scores in favor of annotating a fictional peer's response, $F(2, 162) = 12.08$, $p = 0.086$.

Students in both conditions, on a novel item calling for students to write and revise an essay on the formation of Mt. Hood, made significant pre- to post-test gains showing that students gained integrated understanding of plate tectonics (School A: $t(304) = 8.67$, $p < 0.001$; School B: $t(116) = 5.90$, $p < 0.001$). Another post-test item measured students' ability to use guidance to revise essays by giving students one round of guidance and the opportunity to revise their initial response. Students made significant improvements (School A: $t(315) = 11.04$, $p < 0.001$; School B: $t(142) = 3.15$, $p < 0.001$). There was no significant effect of the condition on pre/post gains or post-test revisions.

Annotating a peer's essay supported deeper engagement in critiquing than annotating one's own essay. Students were more likely to identify weaknesses accurately when annotating a peer's essay. The difference between conditions was significant in School A, $F(1,88) = 8.18$, $p < 0.01$, but not in School B, $F(1, 99) = 1.87$, $p = 0.175$. In School B, this may be because a large percentage of students in both conditions did not place labels. In both schools, a greater percentage of students created their own labels when annotating a peer's essay (20% of students), compared to when annotating their own (12%). When annotating a peer's essay students created new labels that called for incorporation of mechanistic evidence such as, "*why does the blob rise to the top*" or "*what does heat have to do with it*?" When annotating their own essays, student-constructed labels often paraphrased an idea already expressed in their initial essay or corrected spelling.

In sum, the Annotator plus adaptive KI guidance engaged students in critique and revision of their science essays. Annotating a peer's essay showed advantages in the depth of analysis of an essay, possibly by reducing

the effect of confirmation bias. The different school outcomes suggest that automated guidance for annotation may benefit some students by helping them to create and place labels. These results suggest the value of further study of peer collaboration and confirmation bias.

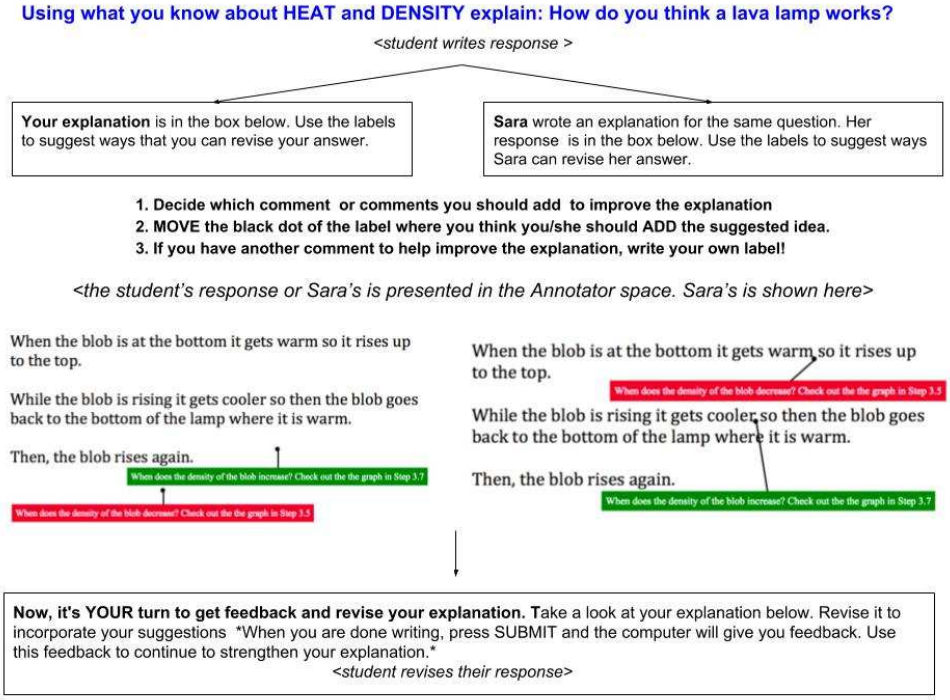

Figure 2. Comparison between "Annotate own" and "Annotate Sarah's" versions.

Presentation 4: Tactic Text - A new platform for computational text analysis

Bruce Sherin

The work reported on here is primarily concerned with the application of computational text analysis to research data. The potential benefits here are great. But the work is still very much in its early days, and I believe we still do not have a full picture of how the tools of computational text analysis should be woven into our research workflows. The purpose of this presentation is to introduce a new platform for computational text analysis, Tactic Text, designed for social scientists engaged in the study of learning. It is worth introducing Tactic Text in this venue not solely because it exists and is a new tool for researchers, but also because it embodies a proposed model for how we can incorporate text analysis into our research workflows. Tactic Text has been mentioned in earlier conference presentations and talks. However, to date, no presentation or research paper has laid out the design of the technology, and the argument for that design, in any detail.

The central tenet of the philosophy behind Tactic Text is this: Computational text analysis should not be seen as a replacement for, or even separate from, forms of qualitative text analysis. Rather, it should be integrated with traditional forms of qualitative text analysis in a manner that amplifies both. The two forms of analysis should be interactive. This core philosophy has a number of implications for the design of Tactic Text, and for the community of users. If the two forms of analysis are to be integrated, then there must be a population of researchers capable in both sorts of work. Furthermore, the tool must support an interactive style of analysis. In contrast, existing tools for computational text analysis tend to hide the data once it is loaded into the system.

Tactic is a fully web-based environment built to embody this philosophy. In some respects, it is akin to existing GUI-based tools for computational analysis, including Weka (Hall et al., 2009), RapidMiner (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006), and LightSide (Mayfield & Rosé, 2013). However, it is unlike these other tools in a number of respects, including the manner in which it is tuned for an interactive style of work. A more complete description of Tactic will be given in the talk, along with a contrast to other platforms. Figure 3 provides an example of a Tactic Text workspace, with an analysis in process. The data are visible on the left, filling a large table. All data in Tactic are accomplished with tiles, which can be added to the environment from menus. The workspace in Figure 3 has two tiles at the right. Tiles have access to the data, and can communicate

with each other.

Each user's library begins with a default set of tiles, and more are available from a shared repository. However, a core belief underlying the design of Tactic is that the vast majority of research projects require at least some programming by users. Thus, the ability to program tiles is fully integrated into Tactic. All programming in Tactic is done in the Python language, and Tactic provides an integrated editor, where tiles can be directly programmed (refer to Figure 4). Tiles have access both to a Tactic-specific API, as well as a wide range of libraries that are useful for computational text analysis. Although Tactic requires users to engage in Python programming, it does take steps to make this programming more accessible. For example, because Tactic is entirely web based, there is nothing for end users to install. Furthermore, all computational work is performed on the server, so Tactic can be used on any machine with a web browser.

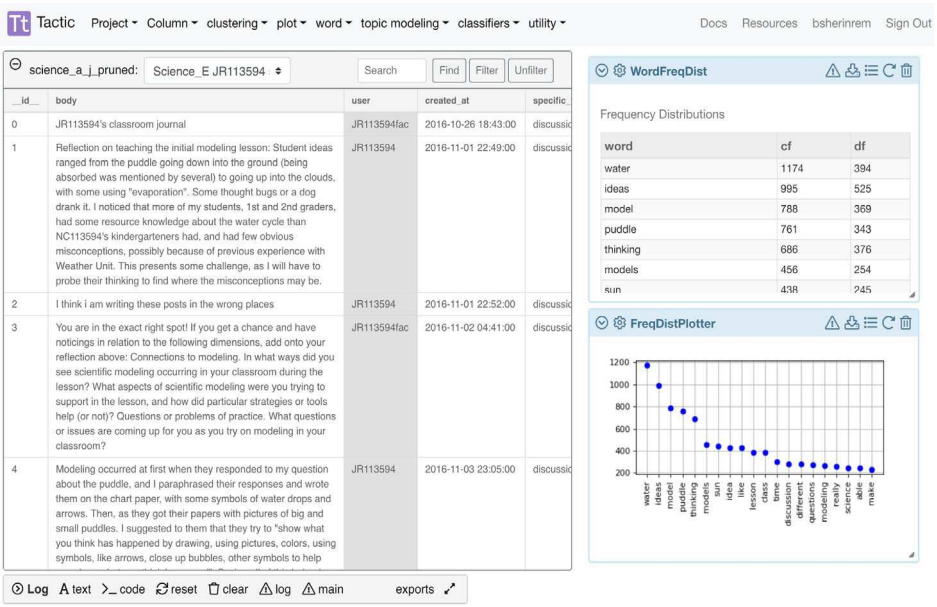

Figure 3. A sample Tactic workspace.

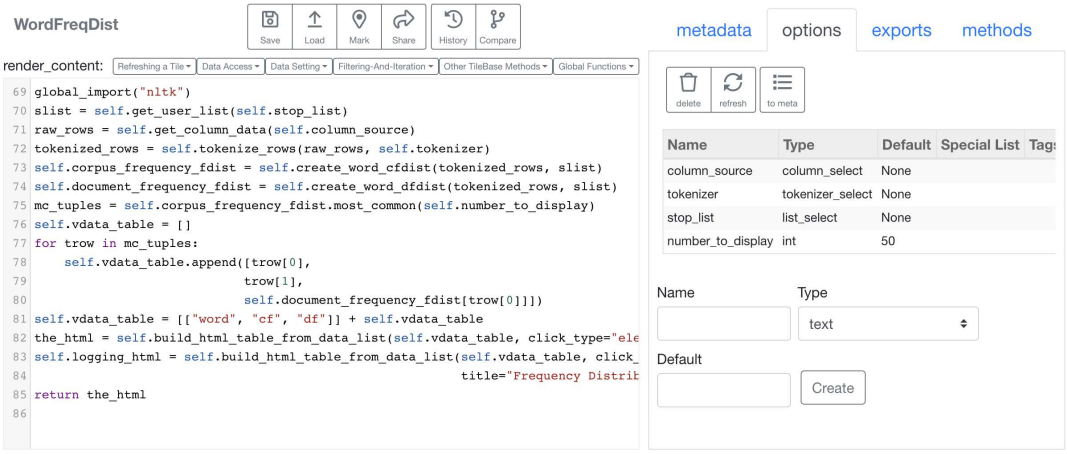

Figure 4. The tile creator in Tactic.

In the talk, I will illustrate the use of Tactic with data from the Learning Labs project (Lomax & Kazemi, 2016; Richards, Thompson, & Shim, 2016). As part of the Learning Labs project, two online courses were developed for in-service K-2 teachers, designed to guide teachers in introducing modeling-based activities into their mathematics and science lessons. The courses each span multiple weeks, and include a wide range of online activities. Many of these activities required them to enter text. Each participating teacher had to upload, watch, and comment on videos, and they could respond to the comments of other teachers, as well as answer questions presented to them as part of the course. In the presentation, I will illustrate how Tactic can be used to capture the

changing way in which teachers understood the nature of modeling as the courses unfolded.

References

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education*, *96*(5), 904-926.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*(3), 472-517.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481-506.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gerard, L. F., Linn, M. C., & Madhok, J. J. (2016). Examining the impacts of annotation and automated guidance on essay revision and science learning. In C-K. Looi, J. Polman, U. Cress, & P. Reimann (Eds.), *International Conference of the Learning Sciences* (Vol. 1, pp. 394-401). Singapore: International Society of the Learning Sciences.

Gerard, L. F., Ryoo, K., McElhaney, K., Liu, L., Rafferty, A. N., & Linn, M. C. (2015). Automated guidance for student inquiry. *Journal of Educational Psychology, 108*(1), 60-81.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581-605.

Liu, O. L., Lee, H. -S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures and evidence. *Educational Assessment, 13*(1), 33-55.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L. F., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215-233.

Lomax, K., Fox, A., & Kazemi, E. (2016). Modeling with mathematics: An online course for K-2 teachers. University of Washington.

Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text. M. D. Shermis & J. Burstein (Eds.) *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124-135). New York, NY: Routledge.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Jrl of the Learning Sciences*, *15*, 153-191.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*, 57-111.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 935-940). ACM.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis.* Thousand Oaks, CA: Sage Publications.

Moschkovich, J. N. (2007). Beyond words to mathematical content: Assessing English learners in the mathematics classroom. In A. H. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 345-352). New York, NY: Cambridge University Press.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821-846.

Rasch, G. (1966). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Richards, J. Thompson, J., & Shim, S.-Y. (2016). *Scientific modeling with young students: A blended course for K-2 teachers.* University of Washington.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching, 38*(5), 529-552.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. DRL-1561149 (Presentation 1), DRL-1220756 & DRL-1418019 (Presentation 2), DRL-1451604 & DRL-1418423 (Presentation 3), and DRL-1417757 (Presentation 4). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.