International Society of the Learning Sciences

Middle School Students' Engagement with Quantitative Data Representations of Fictional Texts

Raquel Coelho, Sarah Levine, Dorna Abdi, Lena Phalen, Liz Harris, Dorottya Demszky, Victor R. Lee
rcoelho@stanford.edu, srlevine@stanford.edu, dabdi@stanford.edu, lphalen@stanford.edu,
harrislz@stanford.edu, ddemszky@stanford.edu, vrlee@stanford.edu
Stanford University

Abstract: Advances in computing and digital data are changing approaches to literacy not just in STEM fields, but humanities. For instance, at the university level, students use digital humanities to enrich their literary literacy, such as understandings of authorial craft and the power of language. We hypothesized that applying digital humanities tools--especially text analytics-based visualizations, such as word clouds and line graphs--could enrich both language arts literacy of younger students, who tend to focus on surface features of literary texts. We report findings from think-aloud interviews with U.S. middle school students engaging with text analytics-based visualizations of familiar stories like *Cinderella*. Students engaged in aspects of literary literacy, including analysis of authorial craft, as well as data literacy, including analysis of frequency and change over time in relation to the stories. Integrating text analytics-based visualizations into middle school literacy instruction may support both language arts and data literacy learning goals.

Introduction

Advances in computing and digital data are reshaping approaches to literacy not only in traditionally-recognized STEM fields, but across all domains (Jiang et al, 2022; Wang & Lester, 2023; Mackey, 2019). For example, current definitions of *literary literacy* include concepts, skills, and dispositions, such as the concept of authorial craft, interpretive skills, and a disposition toward questioning, multiple meanings, and attention to cultural context (Lee et al., 2016). Until recently, the exercise of literary literacies was limited by human capacity to read, connect, and interpret. Now, computing and digital data have inspired the field of digital humanities, a discipline that uses computation to make broad and deep examinations of texts. For instance, a key tool in digital humanists' toolkit is text analytics-based visualization (Clement, 2013) such as a chart, graph, or word cloud. These visualizations can computationally represent word frequency, gendered language, shifts in sentiment, and other features (Wilkens, 2015). With such tools, readers can explore anything from patterns in translations of the Bible to the frequency of particular consonant sounds in poetry (Jänicke et al., 2017).

Much digital humanities work occurs in universities or professions. At those levels, readers draw upon not only literary literacy, but data literacy (Clement, 2013). As with literary literacy, data literacy includes not just skills but ideas and epistemologies, or, in Gebre's definition (2022), "conceptions, competencies, and contexts," including an understanding of the ubiquity of data, the ability to read things like data visualizations, and an appropriately tempered approach to drawing conclusions from data representations. This combination of literacies helps readers access an abundance of rich, challenging textual explorations (Berry, 2012).

K-12 teachers want such experiences for their students as well. However, teachers often express uncertainty about how best to teach literary literacy. And literary instruction often focuses on characters and plot events as opposed to the concepts, skills, and dispositions of literary literacy which can make literary reading so fulfilling (Janssen et al., 2012; Shanahan, 1998). This study was motivated by the hypothesis that digital humanities could also benefit emerging literary readers, such as middle school students. At the same time, using such representations of texts might also support students' data literacy.

Recent work in integrated data literacy education in arts (e.g., Matuk et al., 2021) and history (e.g., Shreiner, 2018) instruction indicates that situating data literacy in non-STEM classrooms can "create alternative entry points into data literacy by building on learners' non-STEM interests" (Matuk et al., 2021). In this study, we turned to middle school English Language Arts (ELA) classrooms to ask two questions: How might visual representations of texts support middle school students' development as literary readers? And how might engaging with textual representations support students' data literacy?

Our university team partnered with middle school ELA teachers and coordinators from a Northern California school district for a three-year project in which we designed numerical and text analytics-based visualizations to support both language and data literacies. For instance, based on teachers' curricula and district ELA goals, we designed data visualizations to represent Poe's *The Telltale Heart* and Garcia McCall's *Summer of the Mariposas*). We engaged in professional learning with partner teachers, observed teacher instruction and

classroom discussion of visualizations, and implemented interviews and think-alouds with middle school students using familiar texts such as *Cinderella* as well as less familiar texts.

In this study, we will focus on one part of our study--an analysis of students' interviews and think-alouds in response to visual representations of two versions of *Cinderella*--one by Charles Perrault, and the other attributed to the Brothers Grimm. We focus on think-alouds here to explore *the degree to which students engaged in literary and data literacy practices when reading and responding to ELA data representations.*

Method

Context and participants

We visited a partner teacher's school to interview students from that teacher's 7th grade ELA classroom. The teacher planned to introduce units including plays and a novel, and she wanted to help her students move beyond surface features of texts to explore other aspects of literary literacy, including attention to authorial choices and the power of language. The teacher selected seven students--two females and five males-- whom she felt would be comfortable sharing ideas with visiting adults. Each interview lasted 20 - 30 minutes.

Materials

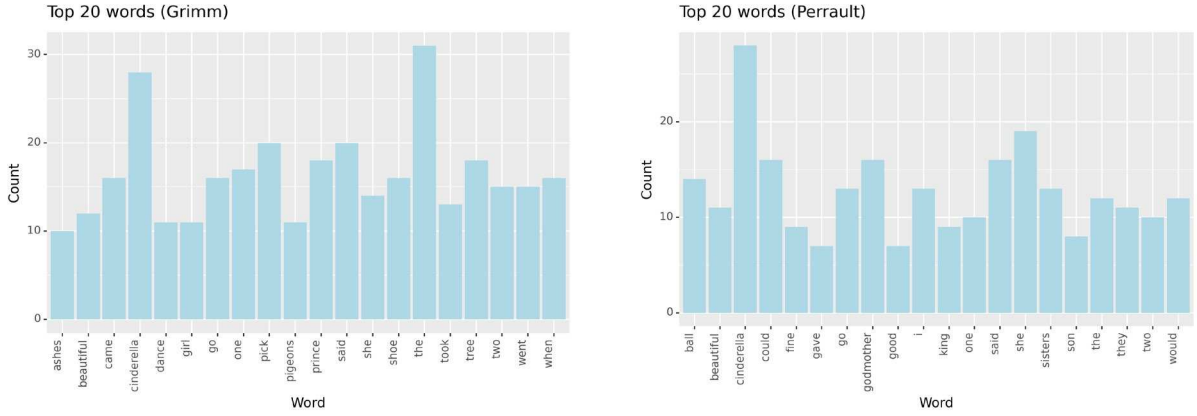
We decided to represent two versions of Cinderella so that students could compare word frequencies and computationally-generated indices of textual qualities, such as mood or sentiment. We also chose to create several representations of these two *Cinderellas* to explore how students drew connections and contrasts. First, we created word clouds using the Python *wordcloud* package. This package displayed the 100 most frequently used words in each version of *Cinderella*. Most frequent words appear in larger font sizes. (Figure 1). Second, we created bar graphs representing the 20 most frequently occurring words in each story (Figure 2). Finally, we created line graphs with VADER, a rule-based sentiment analysis tool. This tool illustrates the average sentiment scores (positive, negative, or neutral) for every 500-word chunk of the two versions (not pictured).

Figure 1

Word Cloud Representations of Two Versions of Cinderella (Grimms' version on left; Perrault on right)

Figure 2

Bar Graph Representations of Two Versions of Cinderella (Grimms' version on left; Perrault on right)



Procedure

In interviews, we presented each type of visualization separately. We prompted the students to share what they noticed and wondered, and then asked targeted follow-up questions. We then asked students to consider similarities and differences between the kinds of representations and the two versions of the story.

Analysis

The interviewers video recorded each interview, transcribed each with an AI transcription service, human-cleaned and edited transcriptions, and de-identified them. Then, two pairs of coders from the research team engaged in top-down and bottom-up coding, using students' turns of talk as the unit of analysis, (N = 119).

First, to explore what students noticed and wondered, we categorized turns of talk as ***observations*** ("I see that the word 'godmother' shows up 14 times"), ***inferences*** ("Maybe this line goes down because Cinderella's stepsisters were unhappy or something"), or both. During subsequent passes through the data, we developed a general code related to ***literary literacy***, which included students' references to authors and literary features. We developed subcodes to identify types of language arts literacy, such as *tone* or *feeling, narrator* or *speaker*, *verb tense*, and *style*. We also developed codes related to ***data literacy***, such as references to *frequency* (e.g., *common, a lot)* or references to *changes over time* (e.g., *trend, fluctuation*), as well as *minimums and maximums*, (e.g., *drop off).* Throughout the coding process, the coders shared their analyses with the whole research team for feedback and revision.

Findings and discussion

We wanted to know if and how students might engage literary and data literacies in analyzing quantified textual data. We analyzed students' observations and inferences when thinking aloud in response to different representations of *Cinderella*. The seven students made a total of 90 observations and 47 inferences. Nineteen inferences were associated with bar graphs, 15 with word clouds, and 13 with line graphs.

We will highlight two trends in students' responses. First, we found that in responding to the visualizations, each student engaged in aspects of literary literacy, including making inferences about authorial craft or literary elements, or offering multiple interpretations of data. When doing so, they pointed to specific elements of the visualizations (e.g., increases on the line graph, small or large words in the word cloud). While we did not include a comparison group in this study, we can say that much classroom research on middle or high school students does not report uptake of such practices. Second, each student also engaged in data literacy practices by exploring aspects of data and by adopting skepticism toward data representations. Both of these findings suggest that these representations might be useful tools for expanding students' ELA and data literacy.

Literary literacy: Authorial craft

During students' interviews and think-alouds, all students considered and questioned the authors of the two versions of *Cinderella* at least three times. They discussed authors' choices regarding point of view, diction, plot, tone, and style. For instance, when looking at the two word clouds, Student 4 noticed more words related to feeling, saying, "Grimm is not just more emotions. It's a lot more." Student 1 inferred that Perrault's style was "less morbid" than Grimm. This student incorporated her knowledge of the Grimm Brothers as "dark" to interpret the frequent references to "ashes" as symbolizing "somebody's darkness or death…that what that makes me think." Student 6 pointed out that, based on the differences in line graphs, "the authors have different techniques -- just small things that the authors wanted to add." Another student considered two possibilities at once: "Maybe the Grimm brothers were using more detail. Or maybe it was a longer story." All of these aspects of literary literacy move beyond students' typical attention to plot and summary.

Data literacy: Frequencies and trends

Students explored foundational features of data, such as frequency and trend. For instance, each student used words like *a lot* or *common* several times to comment on trends in language and authorial choice. Student 1 noted the frequency of the pronoun "I" in one story, saying, "This [version of *Cinderella*] uses 'I' like, almost 15 times. And that's interesting, because maybe it's in first-person …maybe it's like, 'I went to the ball.' That would be interesting, because I've never seen that before…it usually gives like a third-person point of view." In this case, noting the frequency of the pronoun "I" led the student to consider narrative choices typical to a genre, and ways that an author might diverge from those choices--both are big-picture, sophisticated literary concepts.

Students also expressed skepticism at some of the representations. For instance, Student 3 questioned the line graph's representation of sentiment in one version of *Cinderella*, saying "When the prince comes and like, found [Cinderella], I feel like [the line graph] should go up in the mood." Another said, "I don't know if I can

totally trust it because computers don't really have feelings." Such questioning represents an important aspect of data literacy: epistemic caution in response to data representations and quantifications.

The field needs to engage in more research – especially comparative research – to better understand the role of text-based data visualizations in an ELA classroom. We speculate that data visualizations may afford literary literacy because such visualizations literally invite students to see texts differently. For instance, it is unusual for students to engage with a text as a whole, represented on one page or screen. That big picture representation might support increased analysis of craft, language, and literary reasoning.

Digital humanities offer a marriage of literary and data literacy which we think might enrich emerging readers' literary literacy and "create alternative entry points into data literacy by building on learners' non-STEM interests" (Matuk et al., 2021). At the same time, computational approaches to textual analysis invite challenges. One is that quantitative and automated text analysis is inherently biased and incomplete. Such approaches to textual analysis often do not account for either semantic or cultural context. In such cases, reliance on data representation could be counterproductive for literary literacy. Another challenge is that to use these text-based data visualizations, humanities teachers must learn and teach data literacy. However, this challenge is worth facing, as all teachers and learners must account for the shifting state of literacy education.

References

Berry, D. M. (2012). Introduction: Understanding the digital humanities. In *Understanding digital humanities* (pp. 1-20). London: Palgrave Macmillan UK.

Clement T. (2013). Text analysis, data mining, and visualizations in literary scholarship. In *Literary Studies in the Digital Age: An Evolving Anthology*, ed. KM Price, R Siemens. New York: Mod. Lang. Assoc. Am. https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literaryscholarship/

Gebre, E. (2022). Conceptions and perspectives of data literacy in secondary education. *British Journal of Educational Technology*, *53*(5), 1080-1095.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2017). Visual text analysis in digital humanities. In *Computer Graphics Forum* (Vol. 36, No. 6, pp. 226-250).

Janssen, T., Braaksma, M., Rijlaarsdam, G., & van den Bergh, H. (2012). Flexibility in reading literature: Differences between good and poor adolescent readers. *Scientific Study of Literature,* 2, 83–107.

Jiang, S., Lee, V. R., & Rosenberg, J. M. (2022). Data science education across the disciplines: Underexamined opportunities for K-12 innovation. *British Journal of Educational Technology*, *53*(5), 1073-1079.=

Lee, C. D., Goldman, S., Levine, S., & Magliano, J. P. (2016). Epistemic cognition in literary reasoning. In W. A. Sandoval (Ed.), *Handbook of Epistemic Cognition*. Routledge.

Levine, S., Trepper, K., Chung, R. H., & Coelho, R. (2021). How feeling supports students' interpretive discussions about literature. *Journal of Literacy Research*, *53*(4), 491-515.

Mackey, M. (2019). Literacy constants in a context of contemporary change. *English in Education*, *53*(2), 116-128.

Matuk, C., DesPortes, K., Amato, A., Silander, M., Vacca, R., Vasudevan, V., & Woods, P. J. (2021). Challenges and opportunities in teaching and learning data literacy through art. In *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021*.

Shanahan, T. (1998). Readers' awareness of author. In N. Nelson & R. C. Calfee (Eds.), *The reading-writing connection* [Ninety-seventh yearbook of the National Society for the Study of Education] (pp. 88–111). Chicago: University of Chicago Press.

Wang, N., & Lester, J. (2023). K-12 Education in the Age of AI: A Call to Action for K-12 AI Literacy. *International Journal of Artificial Intelligence in Educatio*n, 1-5.

Wilkens, M. (2015). Digital humanities and its application in the study of literature and culture. *Comparative Literature*, *67*(1), 11-20.

Xiong, C., Lee-Robbins, E., Zhang, I., Gaba, A., & Franconeri, S. (2022). Reasoning affordances with tables and bar charts. *IEEE transactions on visualization and computer graphics*.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, *27*, 1073-1079.

Acknowledgments

The authors thank participating teachers and students. This work was supported in part by funding from the National Science Foundation under Grant No. 2241483. The opinions expressed herein are those of the authors and do not necessarily reflect those of the National Science Foundation.