Research Note

Doing Big Data: Considering the Consequences of Writing Analytics

Eric James Stephens, *Clemson University*

Structured Abstract

- **Aim:** This research note focuses on some of the consequences of big data as an emerging methodology. Its purpose is to provide a brief literature review of the method's development and some of the critical questions researchers should consider as they move forward. Salvo (2012) contends that big data as a form of design of communication itself "is necessarily a rhetorically-based field" (p. 38). With big data as an up and coming methodology (McNely, 2012; Salvo, 2012), using caution in its application is a necessity for scholars. Not only should researchers seek out the unseen and untapped applications of big data, but they should learn its limitations as well (Spinuzzi, 2009). You adopt a methodology, you adopt its flaws.

- **Problem Formation:** This section identifies a gap in the field as it relates to some of the consequences of applying big data as a methodology and seeing it as a rhetorical tool. As big data gains steam in the field of humanities, some are sure to question what they see as a flaw: the act of quantifying language. This argument is not new nor is its rebuttal. Harris (1954) discusses the distributional structure of language with each part of a sentence acting as co-occurents, each in a particular position, and each with a relationship to the other co-occurents (p. 146). Salvo (2012) argues that the combination of these new methodologies and technologies "knits together invention, arrangement, style, memory, and delivery in ways that challenge conceptions of print based literacy and textuality" (p. 39). While big data itself has several rhetorical methodologies embedded within, deciding which one to use depends on the amount of data and how it's aggregated.

DOI: 10.37514/JWA-J.2017.1.1.12

- **Information Collection**: As described above, this research note functions primarily as a brief review of literature. This section focuses on how writing analytics developed from content analysis in mass communications and shifted into latent semantic analysis assisted by computer technology. Riffe, Lacy, & Fico (1995) offer a clear explanation of content analysis, which was developed with comparably small data sets in mind: "Usually, but not always, content analysis involves drawing representative samples of content, training coders to use the category rules developed to measure or reflect differences in content, and measuring reliability (agreement or stability over time) of coders applying the rules" (p. 2). Finding a representative sample of content was once a more feasible methodology, but in the digital age that amount of content exponentially increases every day.

- **Conclusions:** As latent semantic analysis is an extension of quantitative content analysis (and vice versa)—and knowing that an adopted methodology carries adopted flaws—it makes sense to turn to some of the concerns voiced by mass communication scholars in order to understand limitations. While quantitative content analysis grew in popularity in mass communication, so did the refining of its methods. Reporting the reliability of a study adds credibility to the study itself, and when a human coder is involved, the reporting of this intercoder reliability becomes imperative (Hayes & Krippendorf, 2007; Krippendorf, 2008, 2011). While intercoder reliability measures the degree to which coders agree, researchers should also be keenly aware of the theory and valence informing their study, which impacts their coders, which ultimately impacts the results of the study itself.

- **Directions for Further Research:** As the field of writing studies begins to adopt big data methodologies, researchers must continue to challenge and question their applications, implementations, and implications, turning to familiar questions from our own fields. Big data is exciting and new, but it's not *the* methodology to explain it all. It's just as rhetorical as every other methodology—it's just better at hiding it.

Keywords: big data, methodology, rhetoric, writing analytics

1.0 Introduction

It seems that nearly every industry and academic discipline has at least one thing in common—stuff. Lots and lots of stuff. These industries and academic disciplines all face a similar dilemma—what to do with all that stuff? "Stuff," obviously, is too vague and a rather unhelpful term, so another term has been provided, what Lewis and Westlund (2015) dubbed "the buzzword *du jour*": big data. Unfortunately, it seems, "big data" isn't any more helpful than saying "stuff." What constitutes big data varies widely from field to field, but most realize that they need to do something with it: "A nearly ubiquitous catchphrase, big data directs attention toward a new phenomenon: production, storage, and analysis of vast quantities of data—data that may exceed the ability of available technologies and methodologies to process" (Graham, Kim, DeVasto, & Keith, 2015, p. 70). Even here, "vast quantities," doesn't help to pin stuff down, perhaps because big data is just so big. As it turns out, *how much* data is not nearly as important as the "capacity to search, aggregate, and cross-reference [it]" (boyd [sic] & Crawford, 2012, p. 663). Whether it's the NSA or Twitter working with zettabytes (one trillion gigabytes) or humanities academics working with thousands of pages of content (less than a few gigabytes), what matters more is the ability to see it and to make sense of it. Big data isn't a thing. Big data is a *verb;* it's a methodology.

2.0 Aim: Explicating Assumptions of Big Data

Since big data is a methodology, it is not bound by a single discipline, which allows scholars from a variety of academic interests to draw upon this tool. This borrowing, adapting, and hijacking of methodologies carries certain risks. Spinuzzi (2009) argues that research techniques are often viewed as the "atoms or essential building blocks of research projects" (p. 411), which tempts researchers to view these methodologies as "arhetorical," believing them to be just methodologies, not rhetorical arguments. Spinuzzi expands on this notion:

> In terms of conducting research, I suggest that technical communicators should view other fields' research approaches with all the caution they would apply to political agreements or legislative settlements. . . . Technical communication is its own field with its own orientations, problems, and environments; if technical communicators deploy another field's negotiated settlement, without alteration, they cede their own status as negotiating partners, and they may find that the settlement does not work nearly as well as they thought it would! (p. 440)

Spinuzzi's caution is one to be considered by any scholar looking to apply an unfamiliar methodology, approaching it with a rhetorical understanding as well. All methods have an underlying rhetoric; in fact, Salvo (2012) contends that big data as a form of communication design itself "is necessarily a rhetorically-based field" (p. 38). Many scholars have begun questioning the use of big data itself. McNely (2012) suggests that

> This data is most often useful in the aggregate—depersonalized, decontextualized, and pasted together with millions of others. There is a real need for communication design researchers and practitioners to formulate approaches that distinguish between the usefulness of big data in the aggregate . . . and big data applied to situated, local, human users. (p. 27)

While big data is certainly a beneficial method for understanding things in the aggregate, as scholars, we should begin thinking about how we can make that data applicable to human users. If this ideal is pursued—and ideally achieved—then research can work towards undermining the utilitarian view of writing assessment as outlined by Elliot (2016), "the greatest net balance of satisfaction summed over individuals, [ignoring] the fate of the individual." This aggregated information always already ignores the individual, but by making the results of the data available and applicable to human users, we can mitigate the consequences. If the data is given by users, then they should have access to how it is used. With big data as an emerging methodology, using caution in its application is a necessity for scholars. Not only should researchers seek out the unseen and untapped applications of big data, but they should learn its limitations as well. You adopt a methodology, you adopt its flaws.

3.0 Problem Formation: Language Quantification

As big data gains steam in the field of humanities, some are sure to question what they see as a flaw: the act of quantifying language. This argument is not new nor is its rebuttal. Harris (1954) discusses the distributional structure of language with each part of a sentence acting as co-occurents, each in a particular position, and each with a relationship to the other co-occurents (p. 146). Once scholars, like linguists, have coded language (i.e., verbs, nouns, gerunds, etc.), it becomes a matter of counting them. Counting, however, is not where the meaning-making stops. Harris argues that the "correlation between language and meaning is much greater when we consider connected discourse. To the extent that formal (distributional) structure can be discovered in discourse, it correlates in some way with the substance of what is being said" (p. 152). Counting parts of

language in a context helps to see the substance of what is being said. In 1954, counting words and their relationships to each other and to discourses would be time-consuming to say the least, even for a single document. Fortunately, new technologies have eased this burden. As Grimmer & Stewart (2013) explain, "automated content methods can make possible the previously impossible in [a discipline]: the systematic analysis of large-scale text collections without massive funding support" (p. 268). Salvo (2012) argues that the combination of these new methodologies and technologies "knits together invention, arrangement, style, memory, and delivery in ways that challenge conceptions of print based literacy and textuality" (p. 39). While big data itself has several rhetorical methodologies embedded within, deciding which one to use depends on the amount of data and how it's aggregated. For writing researchers, this methodology most closely resembles content analysis, just on a larger scale.

4.0 Information Collection: The Role of Perception

While content analysis has been a popular methodology in mass communications and political science for quite some time (Berelson, 1952; Grimmer & Stewart, 2013; Lewis, Zamith, & Hermida, 2013), only recently has "mass" taken on a new meaning as well. Riffe, Lacy, & Fico (1995) offer a clear explanation of content analysis: "Usually, but not always, content analysis involves drawing representative samples of content, training coders to use the category rules developed to measure or reflect differences in content, and measuring reliability (agreement or stability over time) of coders applying the rules" (p. 2). Krippendorf (1989) wrote that "Content analysis is indigenous to communication research and is potentially one of the most important research techniques in social sciences" (p. 403). He goes on to explain that this method allows researchers to "analyze data within a specific context in view of the meanings someone—a group or culture—attributes to them" (p. 403). Initially, mass communication meant communication to the masses, but in this digital age that data has amassed quite a bit. Finding a representative sample of content was once a more feasible methodology, but in the digital age that amount of content exponentially increases every day.

With the development of the internet, researchers found themselves looking for ways to cope with the increase of data available for study (McMillan, 2000; Weare & Lin, 2000). Some looked for more creative ways to search and understand the data as time moved forward (Herring, 2010), describing the amount of "ever-changing, user-influenced, and border-crossing" data with "terms such as liquid, dynamic, and fluid" (Karlsson, 2012, p. 387). It is clear that quantitative content analysis "enable[s] a broader investigation of texts over an

extended period of time" (Boettger and Palmer, 2010, p. 346), and big data as a methodology enables an even broader investigation of these texts.

With qualitative content analysis relying on researchers and coders to make meaning of texts, new methods had to be introduced. In an effort to access these large data sets, researchers began turning to latent semantic analysis, which "is a statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information" (Foltz, 1996, p. 198). When it comes to textual analysis, it's easy to see that not every author or every document uses the same words. Language is complex. Words are intricate. Linguistics is multifarious. Referencing the "classic joke 'Time flies like an arrow. Fruit flies like a banana,'" Gimmer & Stewart (2013) issue a caution: "The complexity of language implies that all methods *necessarily* fail to provide an accurate account of the data-generating process used to produce texts. Automated content analysis methods are insightful, but wrong, models to help researchers make inferences from their data" (p. 270, original emphasis). In an effort to mitigate this necessary failure, researchers rely on operationalized terms and coders (Keyton, 2006) to make sense of how words connect—their semantic meanings. The act of reading a text relies on this method of meaning making, of connecting like-terms to make sense: "To comprehend a text, a reader must create a well-connected representation of information in it. This connected representation is based on linking related pieces of information that occur throughout the text. The linking of information is a process of determining and maintaining coherence" (Foltz, Kintsch, & Landauer, 1998). When one reads, one comprehends, and when one comprehends, one uses latent semantic meaning-making. Latent semantic analysis is a methodology that conducts a quantitative content analysis using a computer to find meaning and correlations in a large corpus of texts that researchers and coders would miss due to the sheer size of the data. Landauer, Foltz, & Laham (1998) explain:

> [Latent semantic analysis] is closely related to neural net models but it is based on a singular value decomposition—a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people. (p. 260)

Essentially, researchers and coders train a computer to make meaning within the corpus of the text to the near similarity of a person—and technology is only getting better. Landauer et al. continue that this methodology "induces its representations of meaning of words and passages from analysis of text alone. None of its knowledge comes directly from perceptual information about the

world; from instinct; or from experiential intercourse with bodily functions, feelings, and intentions" (p. 261). A key word in their explanation is rather telling: "directly." While the sense making certainly comes from the corpus of text itself, the computations and algorithms did not spontaneously appear—someone designed them. As O'Neil (2016) warns, "many of these models encoded human prejudice, misunderstanding, and bias into [these] software systems" (p. 3). Despite this false objectivity, people believe these methods were "fair and objective" because they "didn't involve prejudiced humans digging through reams of paper, just machines processing cold numbers" (O'Neil, 2016, p. 3). The human programmers and human researchers use these machines and algorithms under their "perceptual information about the world" dismissed by Landauer et al.

5.0 Conclusions: The Place of Caution

As latent semantic analysis is an extension of quantitative content analysis (and vice versa)—and knowing that an adopted methodology carries adopted flaws—it makes sense to turn to some of the concerns voiced by mass communication scholars in order to understand limitations. While quantitative content analysis grew in popularity in mass communication, so did the refining of its methods. Krippendorf (2011) makes the clear—and much appreciated—claim that "[t]he need for research to be reliable requires no justification. Testing the reliability of the coding process is a common requirement, especially in content analysis and similar research techniques that make use of human coders to generate data from texts or observations" (p. 93). This is known as inter-coder reliability, which is the degree to which coders agree on what is being coded (Keyton, 2006). Reporting the reliability of a study adds credibility to the study itself, and when a human coder is involved, the reporting of this intercoder reliability becomes imperative (Hayes & Krippendorf, 2007; Krippendorf, 2008, 2011). In one study of content analysis articles over a 26-year period, Lovejoy, Watson, Lacy, & Riffe (2014) found that not every study published its intercoder reliability consistently or uniformly, and they offer several guidelines for ensuring its proper reporting. Researchers argue that given the fundamental nature of content analysis in communication, "it would be logical to expect researchers in communication to be among the most, if not the most, proficient and rigorous in their use of this method" (Lombard, Snyder-Duch, & Bracken, 2002, p. 587). Not only should researchers report their intercoder reliability, but some "argue that the challenge of designing a content analysis can only be adequately met if researchers begin by making decisions about the nature of the content they want to analyze and the role of the theory in their study. Once these decisions are made, it becomes much clearer what the role of the coders is to be" (Potter & Levine-Donnerstein, 1999, p. 259). The content itself and the theory informing the study

impact *how* coders code. Ensuring content from the same genre holds its own benefits to a study (Lauderdale & Herzog, 2016), but the *theory* driving the study has an incredible impact. While intercoder reliability measures the degree to which coders agree, researchers should also be keenly aware of the theory and valence informing their study, which impacts their coders, which ultimately impacts the results of the study itself.

An example of a conscious use of rhetorical theory in combination with big data methodology is the previously mentioned study from Teston & Graham (2012) where "content categories [were] determined by preexisting theory and research. In this case, [they] used stasis theory as an analytic lens in subsequent analyses because each researcher consistently identified disagreement" (Hybrid Methodological Approach section, para. 3). Recalling Spinuzzi's warning, researchers should approach new and old methods with caution. Some have already articulated several ethical concerns with big data, including "privacy, informed consent, and protection from harm, [that raises] wider questions of what kinds of data should be combined and analyzed, and the purposes to which this should be put" (Eynon, 2013, p. 238). According to Fairfield & Shtein (2014):

> [S]ocial scientists are undergoing a fundamental shift in the ethical structure that has defined the moral use of these techniques. Much of social science ethics focuses on rights and responsibilities toward the individual human participant. Big data as a technique does not accommodate this well. There can be millions of research subjects, yet none of them has given traditional informed consent. (pp. 38-39)

What should researchers do with these new technologies and methodologies? Eynon (2013) declares that "[a]s a community we need to shape the agenda rather than simply respond to the one offered by others" (p. 238). This shaping occurs by being aware of the advantages and flaws of particular methods. In this case, the need to understand that intercoder reliability in latent semantic analysis is more than reporting the degree to which researchers and coders agree. It must also be informed by theory.

6.0 Directions for Further Research: Humanities-Based Perspectives

Kelly-Riley and Whithaus (2016) call for "adding—or returning to—humanities-based concerns about fairness and ethics." As we begin to adopt big data methodologies, we must continue to challenge and question their applications, implementations, and implications, turning to familiar questions from our own fields:

- Are humans at the center of our research (Miller, 1979; Katz, 1992)?
- Which humans are at the center of our research (Walton, 2016)?
- Which narratives are being forwarded? Which ones are being left out (Jones, Moore, & Walton, 2016)?
- In which rhetorical ecologies are researchers and coders operating (Inoue, 2016)?
- To what extent will results be used (Graham, Kim, DeVasto, & Keith, 2015; McNely, 2012)?

As these fields move forward, researchers should proceed with caution. Big data is exciting and new, but it's not *the* methodology to explain it all. It's just as rhetorical as every other methodology—it's just better at hiding it.

Author Biography

Eric James Stephens is a PhD student in the Rhetorics, Communication, and Information Design (RCID) program at Clemson University. In addition to pedagogy, his research interests include issues of social justice, schizoanalysis, popular culture, and big data to better understand power relationships in society.

Acknowledgments

Thank you to my wife, the editors, my cohort, and Dr. Steven B. Katz for helping me clarify my thoughts while writing this article.

References

Berelson, B. (1952). *Content analysis in communication research.* Glencoe, Ill.: Free Press.

Boettger, R. K., & Palmer, L. A. (2010). Quantitative content analysis: Its use in technical communication. *IEEE Transactions on Professional Communication*, *53*(4), 346–357.

boyd, danah, & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment, 9*(1).

Eynon, R. (2013). The rise of big data: What does it mean for education, technology, and media research? *Learning, Media and Technology*, *38*(3), 237–240.

Fairfield, J., & Shtein, H. (2014). Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics*, *29*(1), 38–51.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197–202.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307.

Graham, S. S., Kim, S.-Y., DeVasto, D. M., & Keith, W. (2015). Statistical genre analysis: Toward big data methodologies in technical communication. *Technical Communication Quarterly*, *24*(1), 70–104.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2–3), 146–162.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 233–249). Netherlands: Springer.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. Anderson, SC: WAC Clearinghouse; Parlor Press.

Jones, N. N., Moore, K. R., & Walton, R. (2016). Disrupting the past to disrupt the future: An antenarrative of technical communication. *Technical Communication Quarterly*, *25*(4), 211-229.

Karlsson, M. (2012). Charting the liquidity of online news: Moving towards a method for content analysis of online news. *International Communication Gazette*, *74*(4), 385–402.

Katz, S. B. (1992). The ethic of expediency: Classical rhetoric, technology, and the Holocaust. *College English*, *54*(3), 255–275.

Kelly-Riley, D. & Whithuas, C. (2016). Introduction to a special issue on a theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1).

Keyton, J. (2006). *Communication research: Asking questions, finding answers* (2nd ed). Boston, MA; Burr Ridge, IL: McGraw-Hill.

Krippendorff, K. (1989). Content analysis. In *International encyclopedia of communications.* (Vol. 1, pp. 403-407). New York, NY: Oxford University Press.

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures*, *2*(4), 323–338.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*(2), 93–112.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Lauderdale, B. E., & Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, *24*(3), 374–394.

Lewis, S. C., & Westlund, O. (2015). Big data and journalism. *Digital Journalism*, *3*(3), 447–466.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the reporting of reliability in published content analyses: 1985–2010. *Communication Methods and Measures*, *8*(3), 207–221.

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77*(1), 80–98.

McNely, B. (2012). Big data, situated people: Humane approaches to communication design. *Communication Design Quarterly Review, 1*(1), 27–30.

Miller, C. R. (1979). A humanistic rationale for technical writing. *College English*, *40*(6), 610-617.

Moran, M. G. (1985). The history of technical and scientific writing. In *Research in technical communication. A bibliographic sourcebook.* (pp. 25–38). Westport, CT: Greenwood Press,

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*(3), 258–284.

Riffe, D., Lacy, S., & Fico, F. (1998). *Analyzing media messages: Using quantitative content analysis in research*. Mahwah, N.J: Erlbaum.

Salvo, M. J. (2012). Visual rhetoric and big data: Design of future communication. *Communication Design Quarterly Review*, *1*(1), 37–40.

Spinuzzi, C. (2009). Lost in the translation: Shifting claims in the migration of a research technique. *Technical Communication Quarterly*, *14*(4), 411–446.

Teston, C., & Graham, S. S. (2012). Stasis theory and meaningful public participation in pharmaceutical policy. *Present Tense: A Journal of Rhetoric in Society*, *2*(2).

Walton, R. (2016). Supporting human dignity and human rights: A call to adopt the first principle of human-centered design. *Journal of Technical Writing and Communication*, *46*(4), 402–426.

Weare, C., & Lin, W.Y. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review*, *18*(3), 272–292.