theories and methodologies

Why Distant Reading Isn't

JOHANNA DRUCKER

Language is easy to capture, but hard to read.
—John Cayley, "Terms of Reference and Vectoralist Transgressions," *Amodern 2: Network Archaeology*

IF *READING* WERE USED EXCLUSIVELY TO DESIGNATE HUMAN ENGAGEMENT WITH SYMBOLIC CODES, THEN IT WOULD BE RELATIVELY EASY TO dismiss *distant reading* as an oxymoron—unless it were referring to mystical scrying from dizzying heights or deciphering printed matter from across a room.[1] Debates about what constitutes human reading are as varied as the many hermeneutic traditions and pedagogical or cognitive approaches on which they draw (Bruns). But *reading* has been used to describe many mechanical processes and sorting techniques. Punch-card rods, slotted light triggers, Jacquard looms, and many other devices were reading encoded information long before the standard MARC (machine-readable cataloguing) records became ubiquitous in library systems in the 1970s.[2] Outmoded mechanical reading devices have a seductive, steampunk fascination. Many mimicked human actions and behaviors. In addition, these older technologies were embedded in human social systems and exchanges whose processes the machines' operators could partly read. The machines' actions were encoded and decoded by individuals' cognitive intelligence even if the machines functioned automatically.

As computational methods came to the fore after World War II, the mimetic motions of mechanical parts were replaced by the processing of electric signals and then electronic circuitry. Just as magical but far less visible, these techniques made reading into symbolic processing. Now machines exchange information with other machines without human operators. Their processes operate out of sight, invisibly, in ways that a human reader cannot read. These encoded forms are a kind of blind media. The surface of a compact disc, of a piece of magnetic tape, or of a flash drive offers nothing legible to the eye, hand, or ear—our modes of sensory input. We cannot read these media inscriptions, cannot discern any meaning

JOHANNA DRUCKER is the Breslauer Professor of Bibliographical Studies in the Department of Information Studies at the University of California, Los Angeles. A collection of her essays, *What Is?* (Cuneiform Press), was published in 2013, and *Graphesis: Visual Forms of Knowledge Production* (Harvard UP) appeared in 2014.

PMLA 132.3 (2017), published by the Modern Language Association of America

in their traces; we see them mainly as cultural artifacts, even though the devices for which they are designed have automated capacities with which to read them. In the vernacular, we refer to the processing that decodes these invisible traces as reading. But computers do not interpret; they simply find patterns.

Distant reading is the computational processing of textual information in digital form. It relies on automated procedures whose design involves strategic human decisions about what to search for, count, match, analyze, and then represent as outcomes in numeric or visual form. For texts or collections of texts to be read computationally, they have to be divided into meaningful units that an automatic process can recognize. This process is known as tokenization, the decision about how to break a string (or set) of elements (or phenomena) into meaningful units. What is considered meaningful will vary depending on the model of research. A model is a general scheme, or template, that is independent of the individual cases in the study: research on a set of novels has a different model than work on census data, for instance. The models are argument structures that represent components of the information (features of texts, images, numbers, etc.) and their relations to each other. If your project is about tracking spelling changes, then you might need to design a model of analysis that recognizes individual letters and sequences. If your research is on social relations in a narrative, then proper names and other nouns might be what you want to search for in the text. If your work focuses on sentiment analysis, then keywords in context provide a better model of how to sort and count. Or you might need to use probabilistic inference techniques. The decisions about what can be and will be counted are known as parameterization, establishing limits of what is quantitatively (or statistically) meaningful for the project. Consider the parameters used in imaging the surface of a planet for topographical features. The parameters are radically different from those used for optical character recognition (OCR). Because scientists imaging the surface of the planet are searching for features such as tonal variation rather than what are known as crossing points in OCR, their processing techniques are mathematically and conceptually distinct. In both cases, the idea of what is significant has to be modeled in terms the processor can identify explicitly (grayscale value versus place in a grid).

Data mining includes any activity of abstracting information to create or detect patterns. In digital processes, this activity uses algorithms that follow instructions about what to find, match, or count according to the parameters set by the model. The step-by-step instructions are simply ways of making explicit what the computer is to find—just as if a horde of readers were hired to search the entire corpus of Anthony Trollope novels to count the instances of feminine and masculine nouns and pronouns. The results can be listed in a table, spreadsheet, graph, chart, or other mode of display. This data aggregation and display can be embedded in an interactive environment, where filters are applied to search selectively—by date, title, or other relevant information. The difference between a computational version of this process and a human one is largely scale. The mathematical and computational complexity of algorithms might make them hard for an individual to imitate, but, conversely, intuitive decisions that are part of human reading are difficult to specify algorithmically. When these decision-making processes are automated, they can be done at speeds, volumes, and scales (nano-, micro-, macro-) impossible for a human reader to match. Using algorithms to make decisions is, in essence, how the automated text analysis known as distant reading works. Still, no matter how sophisticated the algorithms, they are all based on models designed as interpretative acts.

The distinction between human engagements with symbolic codes and human engagements with machine ones need not be based on a romantic view of humanity but can be made on the terms the machines embody. Automated processes are different from those of a human reader, whose fallibility and vulnerability are crucial to the production of meaning. The machines are fallible too, of course—bugs, errors, and processing mistakes abound—but they are mechanical failures, not the inflected expression of individual thought projected onto and entwined in a work that is produced anew through every interpretative act. The distinction between mechanical and hermeneutic reading, between machine processing and cognitive engagement, between the automatic and the interpretative, between unmotivated and motivated encounters with texts, is essential. Processing is not reading. It is literal, automatic, and repetitive. Reading is ideational, hermeneutic, generative, and productive. Processing strives for accuracy, reading for leniency or transformation. No text-analysis program weeps when it reads the passages in Felix Salten's *Bambi* in which Bambi's mother dies.

We build elaborate systems for natural language processing: vast inventories and vocabularies, keywords in contexts, parsers, and part-of-speech detection.[3] As escalation of automation continues apace, we can model topics according to latent and explicit semantic analysis. We can use sentiment detection in corpus linguistics and high-level training of automated systems to detect shifts in communication patterns across vast collections of textual records through feature recognition and the calculation of maximum likelihood of meaning. Long lateral studies provide views into the shifting values of the culture across time and location. These studies make use of computational language processing (and, increasingly, image processing, though the differences in the remediation of image and of language are not trivial in their implications for computational work). The sense we make of the results of such processes depends on many factors, and these factors are linked to often-unacknowledged aspects of the programs by which textual analyses work.

The phrase *distant reading* produces a nice frisson in initial encounter. To students it suggests that they might skip the labor of passing their eyes over literary or historical texts. The popular media love studies of big trends and global statements, such as research summarizing sentiment across decades of literature in America or in large batches of *Twitter* feeds.[4] But these studies are not just novelty acts that amuse us. They are ways of viewing the cultural record that are as useful as seeing the earth from space when the planet is replaced by a digitally remediated file of features. Some of these approaches depend on texts that are preprocessed using markup schemes or metadata, such as the texts in the Old Bailey archives project, with its centuries of records documenting criminal accusations, judgments, and punishments, or the texts in the Provenance Index of the Getty Research Institute.[5] Enormous amounts of human labor were involved in inserting tags that, in the case of the Old Bailey project, identify plaintiffs and accusers, distinguish charges from sentences, or add elaborate descriptions using controlled vocabularies or text fields that support the imagined research in which the files will be used. The most complex projects use a series of probabilistic and statistical procedures.

But, as described at the outset, much text analysis is fully (or nearly) automatic, so long as *automatic* is understood correctly as the implementation of decisions and designs. The oft-cited founding project of the digital humanities, Roberto Busa's automated production of a concordance of Thomas Aquinas's work, has spurred generations of offspring that demonstrate daily the ways such efforts pay off.[6] But some scholars mistakenly assume that these computational methods of analysis are objective in contrast to the individuated and situated practices of human reading and interpretation. This objective fallacy is problematic. Designing a text-analysis program is necessarily an interpretative act, not a mechanical one, even if running the program becomes mechanistic. The contrast between machine processing and human cognition remains, but the automated methods are also fraught with cultural, historical, and other prejudices built into their design. The methods are based on epistemological assumptions that get translated into metrics and always operate only within the limits of current technical capabilities. One problem with these methods was exemplified in early attempts at mining printed material that used the long form of the *s*—often read as an *f* by OCR programs. Troubleshooting this mix-up was a nontrivial problem.

Machine reading only performs on the literal text, even if inferential, probabilistic, and other techniques are used to create association among the orthographically inscribed textual elements. The issue is not whether a computer can read as well as a human being can but whether a computer will ever be able to read as badly as a human being can, with all the flawed dynamism that makes texts anew in each encounter. Not all texts, to be sure, are aesthetic objects with multivalent meanings. Poetic works and other pieces constructed to produce a rich field for interpretation are different from technical manuals, for instance. But the vast array of texts produced by humans falls in a middle ground in which the interpretative dimensions are many and varied.

At its simplest level of operation—one far exceeded by the sophistication of current algorithms for natural language processing—computational reading matches and sorts strings of characters. Imagine you want to see how often and in what context *mother* appears in literature in English between 1800 and 1900 through such techniques. The first challenge is to define and then locate what you consider the corpus of literature in English in that period. Canonical novels, popular ones, pulp literature, genre fiction, religious and faith-based works, street literature, and transcriptions of song lyrics might be obvious choices. Is there a digital collection that aggregates these works so they can all be searched? Should the corpus be all global anglophone literature of the period? And what about works in translation?

The ASCII string *mother* will vary in meaning depending on use, as in this fabricated passage:

> The mother sheltered in place with her children. The mother of all storms was approaching, and outside she heard the kids in the street teasing each other, "Yo mother is a ho, you mother f---er," as the winds increased. But in the closet in the dark, the image of the virgin mother appeared before her, shivering as only a mother will do in the face of threats to her child, and she wondered if she could really mother her hapless brood.

A word is, of course, never simply a word. Even the use of a definite or indefinite article changes a noun's value. Meaning is what language *does*, and any sense that a word simply *is* goes against a century of linguistic theory, at the least. So the technique that counts a word as a matched letter string has nothing to do with meaning or reading as an interpretative act. Text analysis conflates reading with sorting, counting, and matching. The bucket of words produced by a string search in text analysis is alive and well and freely available in online tools that let you discover the frequencies, contexts, and patterns of use in any corpus you can access as a digital file. While these tools are blunt instruments of analysis, they can be helpful as departure points for research. But they are processing, not reading. Natural-language-processing systems are incredibly fallible—by virtue of their design, which requires complex analytics without understanding. They cannot be programmed to engage with the complexities of language in its common use.[7]

Though string searching is the most basic form of text analysis, probabilistic and context-sensitive programs produce more refined analyses. Such processing still depends on modeling the concepts on which a search will take place. Suppose you want to examine the corpus of English literature to see if there are measurable differences in the stylistic features of authors according to gender?[8] If my gender model is binary, based in biologically determined (actually, assumed) sex characteristics of authors, then my parameters are also binary—male and female. If I complicate this simply by imagining that male authors write in female voices and vice versa and that narrators and authors are not identical in the discursive realm of literary production, then the model of gender is at least complicated by the variations of a woman writing as a man and a man writing as a woman. What if I want to model performative gender? Those constructs change the model from a single binary opposition to a set of categories that blur biological and narratological gender. If the gender model is grounded in recognition of trans, bi, and postbinary understandings, then the question of how it can be implemented to sort literary texts is further complicated. The markers of gender identity may be not in the corpus but in the biographical details, expressed or suppressed, of authors. The point is that the model determines what will emerge from analysis of the texts. Looking at stylistic features and asking how they come to represent or construct conventions of gender as an effect of textual practice would lead in directions other than those that presume to sort according to gender as a given category. Both modeling and parameterization shape the terms by which a text is analyzed to produce quantitative data. Once instrumentalized in processing, models and parameters are rendered almost invisible by the forms in which results are expressed.

The use of quantitative methods requires other caveats. Critical statisticians pose fundamental questions about what it means to take a sample. The sample size, the margin of error, the statistical norms, and other basic qualifications of outcome are essential components of this work. Counting is not statistics, and the difference between the two marks the line between simplistic positivism and skilled quantitative analysis. What processes of parameterization, sampling, and assessment do the results reflect? Text-mining results are not statements about texts but expressions of what algorithms can extract from texts derived from a model of what is being searched. Quantitative approaches are always limited by the partial, skewed, and heterogeneous evidence in the cultural record. Close reading has been criticized for attending to a handful of supposedly exceptional canonical texts and for its theologically modeled hermeneutics against which corpus analysis proposes an inclusive view. But even if collection building, policies of exclusion and inclusion, and the history of cultural-repository formation could be mapped fully, access to replete material history is impossible. The result is not lies built on lies but claims made on partial evidence according to models that may not produce reliable results.

The need to qualify data production and text mining in relation to the terms on which parameterization, modeling, and sampling occur is evident, and the results—that is, data production—will vary considerably depending on the ways these basic starting conditions are set. Once the data are produced, other difficulties arise in presenting and reading results. Developed in fields far from the humanities, information visualizations—metrics expressed as graphics—are pressed into service. In the disciplines that use them, scales, order, sequence, arrangement, choice of color, texture, shape, and other variables have all been studied systematically.[9] But the ease of producing a graph from a spreadsheet obviates the need for a studied approach. At the click of a spreadsheet button, bar charts and continuous graphs, pie charts and scatter plots, whisker plots and tree maps, as well as other standard forms, all appear.

These visualizations are representations, elaborately constructed expressions following legible conventions through a series of interpretative decision points that are all concealed in a final statement that passes itself off as a presentation. The phrase *This is*, with all the conditions of equivalence that the copula contains, is rampant in captioning or glossing these images. Often the visualizations are faulty to the point of being misleading. One common error is to create a continuous graph from discrete data points so that rates of change, differences of value from one point to another, are read as lines whose angle is taken as semantically meaningful. Network graphs, whose screen presence is created through spring-loaded algorithms that optimize legibility in display, are often read for their graphic arrangement without understanding the ways centrality and frequency work in the processing of the data. And yet, time and again, the results of automated text analysis—that is, of distant reading—are presented in scatter plots, charts, and network graphs whose meaning is the result of display protocols, not a semantically significant expression of the data. Edward Tufte's oft-cited principle of graphical excellence "Above all else, show the data" implies that data have a direct correlation to graphical form (92). But in practice, we have good and bad presentations of the same data set, and a single data set can be used to generate any number of different graphical expressions (charts, plots, and so on). The specific semantics and rhetoric of visual epistemological systems are underanalyzed, and the elaborate processes of mediation and remediation are generally overlooked. We are reading (in the cognitive-hermeneutic sense) the artifacts of a process as if they are the actual phenomena. The visualizations are assertions read as declarations.

Text analysis has value. It exposes start points for study and permits the investigation of social and cultural issues in texts at a scale no representative single selective exegesis can produce. It shifts from the symptomatic to the systematic as a mode of inquiry. But often the patterns it shows are the patterns of the processing algorithm and its underlying model. Scientists assessing the vitality of fish populations realized they needed to shift from measuring the largest member of a species to measuring the size of populations in specific locations to map oceans' vulnerable ecologies.[10] The tools of text analysis are just that, tools, and their value is only as good as the models on which they are made, the protocols used to implement those models, and the qualifications that can be attached to the results. But these tools don't read, except in the most mechanical sense.

When Franco Moretti began his research into the use of computational techniques for text analysis, around the turn of the millennium, the tools and platforms for this work were relatively unknown in literary circles, though they had been in development for decades in linguistics. Custom programming was required even for simple tasks. Now basic text-analysis platforms, like *Voyant*, are online and ready to use, and higher-level modeling as well as compression, analysis, and parsing are well-documented and researched. Moretti did not invent the techniques, nor write the algorithms or code, for his projects. But he effectively brought the concept of distant reading to the attention of literary scholars.

The techniques of distant reading will augment the work of scholars, students, researchers, and the public. But will it relieve us of the task of reading? Rather than answer yes or no, we should ask, Why would we want it to? The symbolic provocatively plays itself in our heads, in the chain of exchanges among human beings, cultures, times, geographies, and attitudes.

Text analysis and distant reading embody the peculiar aspiration of the humanities to achieve the conditions of authority perceived to inhere in the natural sciences and their empirical methods. But human perception is characterized by infinite variation and a high degree of specificity, alongside a capacity for ambiguity, ambivalence, contradiction, and association. Reading is an act of self-production, subject enunciation, and, as such, is an emergent and shifting process in human cognition. Computationally sorting, counting, matching, and finding are mechanistic acts. Perhaps adjusting the claims that respect these qualities and qualifications would put the processes in their right relation to our work. We can engage in critical conversation about the differences between mechanistic and hermeneutic work as they inform visualizations and data production and influence cultural practices. Distant reading, when properly understood, is neither mechanistic nor hermeneutic. Its literalness makes it the closest form of reading imaginable. What distant reading lacks is distance. That distance is critical; it is the space between the literal text and the virtual text, between the inscriptional, notational surface and the rhetorical, cognitive effect that produces a text.

Notes

1. For one succinct, critical statement about *distant reading*, see Underwood. For a more technical discussion, see Stubbs.

2. For discussion of Paul Otlet, a pioneer in information science, see Wright; Day. For outdated machines, see Shirriff; Hess.

3. Turney and Pantel provide a technical discussion of natural-language-processing systems.

4. Among the many studies, see Acerbi et al.; Pennebaker; Liu.

5. For more on the Old Bailey project, see *The Proceedings of the Old Bailey* (www.oldbaileyonline.org/). For more on the index, see *The Getty Research Institute* (www.getty.edu/research/tools/provenance/search.html).

6. For a vintage article on the development of automated systems, see Luhn; for recent work, see Renear and Palmer; Grimes.

7. Thanks to my colleague Francis Steen for comments to the effect that natural language processing is seriously stupid and also for the reference to an article by Petrov on parsing that shows the mechanical challenges of the field.

8. For discussion of gender modeling, see Rybicki; Mandell.

9. For a classic work on techniques of statistical graphics, see Schmid; for the classic of semiology, see Bertin. For reliable work on visualization, see Ware; Yau.

10. One study of population modeling is Pope et al.

Works Cited

Acerbi, Alberto, et al. "The Expression of Emotions in Twentieth-Century Books." *PLOS One*, 20 Mar. 2013, dx.doi.org/10.1371/journal.pone.0059030.

Bertin, Jacques. *The Semiology of Graphics*. Translated by William Berg, U of Wisconsin P, 1983.

Bruns, Gerald. *Hermeneutics: Ancient to Modern*. Yale UP, 1992.

Day, Ronald E. *The Modern Invention of Information*. Southern Illinois UP, 2008.

Grimes, Seth. "Brief History of Text Analysis." *BeyeNetwork*, 30 Oct. 2007, www.b-eye-network.com/view/6311.

Hess, Whitney. "Historical Technology: Machines of Times Gone By." *Pleasure and Pain*, 11 Apr. 2011, whitneyhess.com/blog/2011/04/11/historical-technology-machines-of-times-gone-by/ .

Liu, Bing. *Sentiment Analysis: Mining Opinions, Sentiment, and Emotions*. Cambridge UP, 2015.

Luhn, H. P. "A Business Intelligence." *IBM Journal of Research and Development*, vol. 2, no. 4, Oct. 1958, pp. 314–19.

Mandell, Laura. "Visualizing Gender Complexity." 9 June 2016, Universität Hamburg. Available at lecture2go.uni-hamburg.de/l2go/-/get/v/19498.

Pennebaker, James. *The Secret Life of Pronouns: What Our Words Say about Us*. Bloomsbury Press, 2011.

Petrov, Slav. "Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source." *Google Research Blog*, 12 May 2016, research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html.

Pope, Kevin L., et al. "Methods for Assessing Fish Populations." *Inland Fisheries Management in North America*, edited by M. C. Quist and W. A. Hubert, 3rd ed., American Fisheries Society, 1 Jan. 2010. *Digital Commons at University of Nebraska, Lincoln*, digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1072&context=ncfwrustaff.

Renear, Allen H., and Carole L. Palmer. "Strategic Reading, Ontologies, and the Future of Scientific Publishing." *Science*, vol. 325, no. 5942, 14 Aug. 2009, pp. 828–32.

Rybicki, Jan. "Vive la Différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies." *Digital Scholarship in the Humanities*, vol. 31, no. 4, 8 July 2015, dsh.oxfordjournals.org/content/early/2015/07/07/llc.fqv023.

Schmid, Calvin. *Statistical Graphics: Design Principles and Practices.* John Wiley and Sons, 1983.

Shirriff, Ken. "Inside Card Sorters: 1920s Data Processing with Punched Cards and Relays." *Ken Shirriff's Blog*, www.righto.com/2016/05/inside-card-sorters-1920s-data.html. Accessed 27 Nov. 2016.

Stubbs, Michael. *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*. Wiley-Blackwell, 1996.

Tufte, Edward. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

Turney, Peter, and Patrick Pantel. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research*, vol. 37, Feb. 2010, pp. 141–88.

Underwood, Ted. "The Real Problem with Distant Reading." *The Stone and the Shell*, 29 May 2016, tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/.

Ware, Colin. *Information Visualization: Perception for Design*. Morgan Kaufman, 2012.

Wright, Alex. *Cataloguing the World: Paul Otlet and the Birth of the Information Age.* Oxford UP, 2014.

Yau, Nathan. *Visualize This.* Oxford UP, 2011.