Chapter Title: Alien Reading: Text Mining, Language Standardization, and the Humanities

Chapter Author(s): JEFFREY M. BINDER

Book Title: Debates in the Digital Humanities 2016

Book Editor(s): Matthew K. Gold and Lauren F. Klein

Published by: University of Minnesota Press

Stable URL: https://www.jstor.org/stable/10.5749/j.ctt1cn6thb.21

PART III

DIGITAL HUMANITIES AND ITS PRACTICES

This page intentionally left blank

Alien Reading: Text Mining, Language Standardization, and the Humanities

JEFFREY M. BINDER

For all the talk about how computers allow for new levels of scale in humanities research, new debates over institutional structures, and new claims to scientific rigor, it is easy to lose sight of the radical difference between the way human beings and computer programs "read" texts. Topic modeling, one of the most touted methods of finding patterns in large corpora, relies on a procedure that has little resemblance to anything a human being could do. Each text is converted into a matrix of word frequencies, transforming it into an entirely numerical dataset. The computer is directed to create a set of probability tables populated with random numbers, and then it gradually refines them by computing the same pair of mathematical functions hundreds or thousands of times in a row. After a few billion or perhaps even a few trillion multiplications, additions, and other algebraic operations, it sutures words back onto this numerical structure and presents them in a conveniently sorted form. This output, like the paper spit out by a fortune-telling machine, is supposed to tell us the "themes" of the texts being analyzed. While some of the earliest computational text-analysis projects, like Father Roberto Busa's famous collaboration with IBM on the *Index Thomisticus,* began by attempting to automate procedures that scholars had already been doing for centuries, topic modeling takes us well beyond the mechanical imitation of human action (Hockey). When we incorporate text-mining software into our scholarly work, machines are altering our interpretive acts in altogether unprecedented ways.

Yet, as Alan Liu has argued, there has been relatively little interchange between the scholars who are applying these computational methods to literary history and those in fields like media studies who critically examine the history and culture from which this computational technology emerged ("Where Is Cultural Criticism in the Digital Humanities?"). Many scholars of technology, including Lisa Gitelman, Wendy Hui Kyong Chun, Tara McPherson, and David Golumbia, have argued that the seemingly abstract structures of computation can serve ideological ends; but scholars who apply text mining to literary and cultural history have largely skirted

the question of how the technologies they use might be influenced by the military and commercial contexts from which they emerged (Gitelman, *Paper Knowledge*; Chun, *Control and Freedom*; McPherson, "Why Are the Digital Humanities So White?"; Golumbia, *Cultural Logic of Computation*). As a way of gesturing toward a fuller understanding of the cultural context surrounding text-mining methods, I will give a brief account of the origins of a popular technique for topic modeling, Latent Dirichlet Allocation (LDA), and attempt to situate text mining in a broader history of thinking about language. I identify a congruity between text mining and the language standardization efforts that began in the seventeenth and eighteenth centuries, when authors such as John Locke called for the stabilization of vocabularies and devalued "literary" dimensions of language such as metaphor, wordplay, and innuendo as impediments to communication. I argue that, when applied to the study of literary and cultural texts, statistical text-mining methods tend to reinforce conceptions of language and meaning that are, at best, overly dependent on the "literal" definitions of words and, at worst, complicit in the marginalization of nonstandard linguistic conventions and modes of expression.

While text-mining methods could potentially give us an ideologically skewed picture of literary and cultural history, a shift toward a media studies perspective could enable scholars to engage with these linguistic technologies in a way that keeps their alienness in sight, foregrounding their biases and blind spots and emphasizing the historical contingency of the ways in which computers "read" texts. What makes text mining interesting, in this view, is not its potential to "revolutionize" the methodology of the humanities, as Matthew Jockers claims, but the basic fact of its growing influence in the twenty-first century, given the widespread adoption of statistical methods in applications like search engines, spellcheckers, autocomplete features, and computer vision systems. Thinking of text-mining programs as objects of cultural criticism could open up an interchange between digital scholarship and the critical study of computers that is productive in both directions. The work of media theorists who study the ideological structures of technology could help us better understand the effects that computerization could have on our scholarly practice, both in explicitly digital work and in more traditional forms of scholarship that employ technologies like databases and search engines. On the other side, experimenting with techniques such as topic modeling in a critical frame could support a more robust analysis of the cultural authority that makes these technologies seem natural at the present moment, baring the ideological assumptions that underlie the quantification of language, and creating, perhaps, a renewed sense of the strangeness of the idea that words can be understood through the manipulation of numbers.

Models of Language

"Topic modeling" does not refer to any single method, but rather to a number of distinct technologies that attempt to determine the "topics" of texts automatically. The

implementation most commonly used in the humanities is a program called MALLET, developed by Andrew McCallum and others and based on an algorithm developed by David Blei (Blei, Ng, and Jordan, "Latent dirichlet allocation"; McCallum, *MALLET*). Provided with a collection of text files, MALLET can produce "topics" that look, in the output of the program, like this:

> passions passion pleasure person love pride object hatred humility
> men interest natural society property actions justice human moral
> reason nature give principles general observe relations common subject
> idea ideas objects existence mind perceptions impressions form time
> object imagination relation effect present mind idea experience force

This model was trained using the text of David Hume's *Treatise of Human Nature,* divided into Hume's relatively short sections. Each line represents a "topic"—a cluster of words that tend to appear together in the same section. MALLET presents these topics as lists of words (e.g., "passions passion pleasure . . ."), starting with the word most strongly affiliated with the topic and proceeding downward. The program associates each text with one or more of these topics, which constitute a guess as to what that text is "about." There is no certainty to this process; the topics are produced by an approximate method and so the results are slightly different every time the program is run. The meaning of the results is further complicated by the fact that the "topics" in the output do not necessarily correspond to anything for which a simple description might exist. In many cases, they seem to be based more on sets of discursive conventions than on what we normally think of as "topics," and the results often include one or more topics that are totally inscrutable. Interpretation emerges as a key issue, especially given that the method depends on a complex set of assumptions that are colored by the institutional situation from which topic modeling emerged.

The idea of using a computer to automatically identify "topics" is in large part a product of the desire to exploit the increasingly large amount of text that was being distributed electronically in the late twentieth century. While the earliest attempts at automated "topic detection" go back to the 1960s, the field expanded greatly starting around 1990. (For an example of a very early attempt, see Borko and Bernick.) Many of the efforts from the 1990s dealt primarily with text from newswires and were designed for applications in finance and national security. The major accomplishments of this period include a software package known as SCISOR (System for Conceptual Information Summarization, Organization, and Retrieval), developed in the early 1990s, and the DARPA-funded Topic Detection and Tracking initiative, which ran from 1996 to 1997 (Jacobs and Rau; Allan et al.). The primary goal of the DARPA initiative, which drew participants from Carnegie Mellon University and the University of Massachusetts, was to come up with a way of automatically detecting the occurrence of major world events, such as volcanic eruptions and political

elections, through the text analysis of news feeds. The methods developed for this project mostly worked in a different way from the topic-modeling software that is now most familiar in the humanities. Instead of dealing with static collections of texts, they were designed to process continuous text feeds that changed from one topic to another at irregular intervals. One of the primary functions of the software was to determine when these transitions took place.

The topic-modeling techniques most commonly used in DH emerged around the same time as these projects, but they came from an area of research that was more oriented toward static collections than continuous news feeds. One of the most influential methods to emerge from this area is Latent Semantic Indexing (LSI), which was introduced in 1990 by Deerwester et al. ("Indexing by Latent Semantic Analysis"). Unlike the methods designed for the "segmentation" of newswire text, LSI and the other methods it inspired are meant to work with collections of discrete documents. The most common LSI-derived methods of topic modeling also depend on the "bag of words" assumption. Under this assumption, the computer takes no account of sentence divisions, syntax, or even the relative positions of words, considering only how frequently each word type appears in each document. Using these frequencies, LSI identifies clusters of associated words ("topics") and links them to particular documents, something that can serve two major purposes. First, as the name Latent Semantic Indexing suggests, it can be used as a subject index, helping users find documents that are relevant to particular topics; and second, it produces a "reduced description" of the corpus that can be used to visualize patterns in a large amount of text.

In the original version of LSI, the topics are computed through a more-or-less arbitrary procedure that was empirically found to produce reasonable results for the test dataset, a collection of information science abstracts (Deerwester et al., 19; Blei, Ng, and Jordan, 994.). In 1999, Thomas Hofmann developed a new variant of LSI based on a probabilistic model, a mathematical construct that offers a sort of rationale for the method (Hofmann, "Probabilistic Latent Semantic Indexing"). Texts, the model asserts, are composed of mixtures of certain "topics," each of which has an associated vocabulary; a text about, for instance, fishing and economics is most likely to contain words that are strongly associated with these topics. Hofmann's procedure can be used to determine the "topic" definitions that best fit a given collection of text based on this model. In 2003, David Blei, Andrew Y. Ng, and Michael I. Jordan introduced a further modification of the method called Latent Dirichlet Allocation (LDA), which is the variant used by MALLET and remains the most popular form of topic modeling among humanists (Blei, Ng, and Jordan; McCallum). LDA uses a similar model to Hofmann's, but it adds a mathematical function called a *Dirichlet distribution* to determine the probabilities of certain topics occurring together. Adding this function makes the model fully *generative,* which means, in a computer-science context, that it offers a complete mathematical description of the process by which the input texts were (hypothetically)

generated, including a way of determining the probabilities of specific outcomes. (For a general introduction to generative modeling in relation to other forms of machine learning, see Jebara.) The generative model underlying LDA is something like this: first, the writer picks a "mixture" of topics to write about; then the writer constructs the text word-by-word by first randomly choosing a topic from the mixture and then picking a word based on the probability table for that topic. This generative model allows a computer to perform two complementary operations: a topic-modeling program can "learn" what words are associated with what topics based on a corpus of text; then, it can use this model to infer the likely topics of other texts.

It should be apparent from my admittedly rough description that this form of statistical modeling carries a heavy weight of epistemological baggage. Prominent among the disciplinary norms that govern the legitimacy of evidence and methodology in machine learning is the idea that the performance of the tools should be judged against a "gold standard" that defines the correct output—a practice that assumes the desired result to be both fixed ahead of time and accessible through some means outside of the method itself (Juckett). The institutional formation from which text-mining software emerged has also influenced the sorts of language for which it is designed. As I noted previously, the original version of LSI was initially tested with information science abstracts. The paper that introduces LDA draws its examples from AP and Reuters news articles, while Blei's later work has included models based on articles from the journal *Science,* the *Yale Law Journal,* and the *New York Times* (Blei, Ng, and Jordan; Blei, "Probabilistic Topic Models"; Blei, "Topic Modeling and Digital Humanities"). The sorts of text on which these methods are generally tested have a number of commonalities. They are primarily written in a standard dialect and orthography; they tend to privilege the informational over the aesthetic dimensions of language; and they primarily consist of prose. Many of the examples used in testing these methods are also, it is worth noting, the sorts of text that the military-industrial apparatus would have a clear interest in mining. The language commonly used in articles about topic modeling—articles by Hofmann and Blei describe users going on "quests" for information in collections of texts that bear "hidden" or "latent" meanings—is suggestive of the ultimate purpose of the technology (Hofmann; Blei, "Probabilistic Topic Models"). By automatically determining what large numbers of documents are "about," the software can help operators find and "extract" what they need from texts that are assumed to be repositories of information.

While LDA has proved to work reasonably well when applied to texts that are outside of its original purview, including nineteenth-century novels, literary criticism, and early eighteenth-century essays, it is reasonable to ask whether the results it produces are affected by the assumptions that went into the development of the software.[1] One scholar who has considered the biases of topic modeling while employing it in humanistic research, Lisa Marie Rhody, argues that topic models of poetry must be read in a different way from those based on scientific journals

("Topic Modeling and Figurative Language"). A reason, she suggests, is that poetry characteristically uses a relatively large amount of figurative language and produces meaning in a much wider variety of ways than do "non-figurative" texts. An attempt to topic model poetry thus encounters particular interpretive difficulties, but it also "illustrates how figurative language resists thematic topic assignments and by doing so, effectively increases the attractiveness of topic modeling as a methodological tool for literary analysis of poetic texts." In the topic model that she produced based on the Revising Ekphrasis[2] collection of poetry, Rhody finds that some of the most interesting "topics" correspond less to what poems are "about" than to particular poetic traditions. She argues, in particular, that a topic with the top words "death life heart dead long world blood earth man soul men face day pain die" corresponds to the language of elegiac poetry, and she uses it to highlight elegiac qualities in poems by African American poets that are not explicitly about death.

While Rhody frames her argument in terms of a "caricatured" view that hyperemphasizes the figurative nature of poetry, the distinction between figurative and non-figurative language is slightly misleading as an explanation of why topic modeling works particularly well with scientific texts. Many phrases occur repeatedly in scientific abstracts that are arguably figurative: ideas being "underlined" and "highlighted," "first steps" toward solutions, "root causes" of problems. From the perspective of topic modeling, what is important is not that words be used literally, but that the vocabularies of texts correlate with their topics in a uniform fashion. Scientific language fits this requirement particularly well in part because it is produced within a system of overlapping subdisciplines that have distinctive lexicons. The slippery question of whether well-worn phrases like "root cause" are figurative is beside the point; what is important to a topic-modeling program is that this formula is repeated commonly in engineering abstracts but is relatively rare in physics, which makes it a potential distinguishing factor between the two disciplines.[3] In addition to having to work with highly specialized technical vocabularies, scientific writers are encouraged to stick with established usages rather than inventing novel expressions for things that have already been described, which creates highly repetitive patterns in word usage that facilitate the detection of topics.[4] Because of the tight control of the vocabulary to be used within each specialization, the process of scientific writing hews very close to the generative model by which LDA assumes texts were written—much closer, I venture, than the process of writing poetry, although Rhody's example shows that some poetic traditions do have distinctive vocabularies that topic modeling can detect.

The assumption that word choice follows uniformly from the "topic" of a text—whatever we take the "topics" to represent—presumes a sort of linguistic standardization that is historically bound up with structures of authority. As John Guillory has noted, while we now tend to see a plain and direct style as the default for most forms of writing, rhetoricians in the early-modern period placed a greater

value on *copia,* a style that involves a profusion of different ways of saying the same thing ("The Memo and Modernity"). The idea that words should have fixed meanings was largely a product of the latter half of the seventeenth century, when authors like John Locke, Thomas Sprat, and John Wilkins began to see the fluidity of language as an obstruction to clear thought.[5] The compilation of dictionaries for European languages, which began in earnest around that time, created a newly sharp division between standard and nonstandard uses of words, largely based on which usages were "authorized" by their inclusion in the works of eminent writers. It also sharpened the divisions between languages; as Benedict Anderson argues, the "lexicographic revolution" led Europeans to see languages as the property of particular groups, creating imagined communities of German, French, and English speakers where previously there had been a profusion of local dialects (84). This shift in the way Europeans thought about language enabled the creation of highly uniform styles like the ones now used in scientific writing and gave a greater stability to the word usage of many other forms of writing, reducing both variation for variation's sake and many regionalisms. The reason topic modeling does particularly well at identifying themes in technical and informational genres like news articles, abstracts, and encyclopedia entries is that they actively strive to follow this sort of standard, sticking for the most part to usages that have the force of authority behind them.

While topic modeling's affinity for uniform language can be accounted for if the collection of texts is fairly homogeneous, as in Blei's collection of newspaper articles and Rhody's collection of poetry, it becomes a more difficult problem when the method is applied to a corpus that includes texts of various types with varying relations to linguistic standards. When trained on collections that include mostly standardized text but some text that follows other conventions—as is the case with many collections of nineteenth-century fiction—topic models tend to relegate the nonstandard words to a small number of topics while excluding them from the rest of the model. For example, Matthew Wilkins's topic model of the Wright American Fiction corpus[6] includes these two topics: "uv wuz ez hev wich hed sed sez ther ef" and "dat master slave negro massa slaves white black dis dey" (Wilkins, *100-topic model of the Wright American Fiction corpus*). The language of the mock-rustic characters of humorists like David Ross Locke, Charles Farrar Browne, and George William Bagby gets its own topic, while the spellings that some novelists used to represent African American speech are mixed together with words having to do with slavery. Topic modeling is fairly good at distinguishing languages, something that could potentially be useful, but this tendency to separate linguistic conventions could easily become problematic if we are not extremely careful in how we interpret the results. To the extent that it is relegated to its own topic, orthographically distinctive text is prevented from influencing the other topics in the model. If we are to use the other topics as a way of tracking themes or "discourses" in the collection,

we are effectively excluding the words of characters who are presented in caricature from affecting our results, repeating the structure of authority that enables their speech to be coded as nonstandard.

LDA is not just tuned to work best with standardized (and, one might say, hegemonic) forms of language; it also structures its results in a way that encourages interpretation in terms of the standardized meanings of words. In chapter 45 in this volume, Tanya Clement discusses a property common to many text-mining techniques, a dependence on the assumption that "the Word" is a stable and inherently meaningful unit of language ("The Ground Truth of DH Text Mining"). The tendency of text-mining programs to accentuate the stability of words results, in part, from the way in which statistical methods tend to smooth out individual discrepancies so as to emphasize the overall patterns in a dataset. This smoothing is not an accident, but a necessary result of the need to avoid what statisticians call *overfitting* (Dietterich). A model that exactly accounts for every nuance of a dataset tends to be too complex to be useful—to take an image from Jorge Luis Borges, it is like a map that is as large as the territory it represents—and thus, some cases that deviate from general trends have to be ignored (Borges, "On Exactitude in Science"). Though the practicalities of modeling require the smoothing-out of differences, this process is an ideologically loaded way of dealing with language, and the much-vaunted comprehensibility of topic models depends on it. Each "word" in the output of a topic-modeling program stands for many instances of that word in the input, each one with a unique syntactic context that the model largely ignores. An interpretation of these aggregate-words can easily slide into the assumption that all of these instances can be encompassed by a single meaning.

An example of this smoothing-out of instabilities in word meaning occurs in Matthew Jockers's book *Macroanalysis*. After introducing the idea of topic modeling, Jockers presents two topics from a model of the Stanford Literary Lab's corpus of novels. Jockers proceeds to interpret the appearance of the word *stream* in the list of top words for one of these topics, alongside *indian, indians, chief, savages, warriors, men, party,* etc.:

> In conjunction with the much larger company of other words that the model returns, it is easy to see that this particular use of *stream* is not related to the "jet stream" or to the "stream of immigrants" entering the United States in the 1850s. Nor is it a word with any affiliations to contemporary "media streaming." This *stream* refers to a body of flowing water. (127)

Here Jockers seems to be doing something familiar to literary critics: determining the meaning of a word based on context. But the "particular use" of *stream* to which Jockers is referring is neither a word type (the word *stream* considered in the abstract) nor a word token (a particular instance of the word in a text)—it is an

entry in a probability table that was generated through an approximate optimization method. This unit does not correspond to a single "use" of a word in any usual sense, but rather derives from patterns among many different instances of the word in the corpus. Although some of these instances might refer to a body of flowing water, there is no guarantee that they all use the word in the same sense—there are, for instance, at least a few dozen references to a "stream of settlers" in nineteenth-century texts that discuss conflicts between Europeans and Native Americans, and if these are present in Jockers's corpus they would likely be included in the topic he discusses.[7] In attempting to determine what the *stream* in the topic model "refers to," Jockers interprets this abstract composite as if it were the same type of thing as a word token in a literary text, a move that presupposes the stability of the word's signification in the parts of the corpus covered by the topic. This sort of interpretation-in-aggregate is not necessarily illegitimate if we recognize it for what it is, but Jockers's application of simple and familiar terms of interpretation to a topic model belies the very complex and potentially problematic set of assumptions that underlie what he is doing in this passage.

The bias toward standardized forms of language is present not only in topic modeling, but in many other text-mining methods that depend on statistical analysis of words. The affinity of these methods for particular forms of language becomes readily apparent in an exchange between Jockers and Annie Swafford about Jockers's *Syuzhet* program (Jockers, "Revealing Sentiment and Plot Arcs with the Syuzhet Package"). This package uses sentiment analysis software to guess the emotional valence of each sentence of a novel and plots an "arc" that is derived from these results. In a blog post, Swafford points out a number of problems with this method, among them the inability of sentiment analysis to account for the nuances of literary language ("Problems with the Syuzhet Package"). Responding to the latter problem, Jockers admits his frustration: "Things like irony, metaphor, and dark humor are the monsters under the bed that keep me up at night" ("Some thoughts . . ."). The difficulty of accounting for these aspects of language in projects like *Syuzhet* seems to stem from the fact that all of the sentiment analysis methods presently available are designed to suit the language of, in Swafford's words, "a tweet or product review" ("Continuing the Syuzhet Discussion"). In other words, Jockers's analysis depends on a tool designed to suit the contemporary descendants of the Enlightenment project of rationalizing language and standardizing the meanings of words. Although some of the problems that Swafford points out are specific to the software that *Syuzhet* uses, many other text-mining techniques share the tendency to work best with texts that straightforwardly follow standard usages while treating the existence of "non-literal" language, when they deal with it at all, as a problem to be solved. If we are to adopt text-mining tools in humanistic research, we will need to take account of the assumptions they make about language and how those assumptions could serve ideological interests.

Alien Reading

Although these observations suggest that there are good reasons for scholars to be wary of the adoption of text-mining software in the humanities, it would be a mistake simply to dismiss it as irrelevant to our concerns. In his 2014 article, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," Ted Underwood forcefully points out that literary scholars outside of the digital humanities have already been using text-mining software on a regular basis for decades in the form of databases and search engines, but we have done little to theorize the role that these technologies play in scholarly practice (64). Many of the databases scholars commonly use already depend on generative models and other text-mining techniques for correcting scanning errors and accounting for spelling variants. In the present day, it is virtually impossible for scholars to avoid text-mining software altogether, even if many of us only encounter it indirectly through platforms like Google or JSTOR. If, as scholars, we are to engage with these technologies on our own terms, then we will have to find a way of making their roles in humanistic research a matter of active concern. Experimenting with text-mining programs in English departments could serve as a safeguard against the possibility that we unknowingly absorb these tools into our practice without reflecting on the assumptions about language and knowledge that underlie them and considering the effects they could have on our work.

The unreflective computerization that Underwood points out presents a particular problem in the present moment because of a current trend toward user interfaces that cover up the complexity of what goes on inside the machine. As Lori Emerson argues in *Reading Writing Interfaces,* the naturalistic interfaces of modern computers make their operations seem much simpler and more familiar than they really are, encouraging a passive, consumer-like orientation towards the computer rather than a deep understanding of it (1–19). The interfaces of tablet computers especially make heavy use of elements that mimic the behavior of physical objects, appealing to very familiar intuitions about how objects behave. This makes the devices easy to use up to a point, but it gives the typical user little insight into how they work. Search engines can similarly be much more complex than their user interfaces suggest, employing sophisticated algorithms for cleaning up and indexing texts, identifying synonyms, and determining the "relevance" of results that depend on strong assumptions about language and that could potentially introduce biases into research. While many text-mining programs present their results using familiar terms like *word, topic,* and *similarity,* the mathematical structures underneath are often fundamentally different from the ways in which human beings ordinarily understand these concepts. The apparent simplicity of interfaces like the search box allows us to use these technologies in our scholarship without confronting the complexity of what they do and the ways in which their designs might conflict with our precepts as scholars.

While Underwood responds to this problem with an embrace of statistical modeling, it is also possible to employ text-mining programs without accepting the thinking behind them, pushing back against naturalistic user interface design by drawing attention to aspects of the software that conflict with a humanistic view of interpretation. This approach would involve encountering text mining as an alien form of reading—alien both in the fact that it emerged from a discipline with very different concerns from our own and the fact that it is performed by a machine, the sort of nonhuman agent that Ian Bogost has sought to understand with his idea of *alien phenomenology* (Bogost). Rhody's work with topic modeling is one example of a project that employs text-mining technologies while keeping in mind the ways in which their assumptions might clash with the concerns of humanists. I would like to suggest an approach that goes further into a critique of the technology itself, engaging with text-mining tools as embodied, historically situated cultural productions that are potentially problematic. Understanding the extent to which our use of digital tools can reinforce hegemonic views of language requires a sort of scholarship that takes up a critical, perhaps even antagonistic attitude toward computerized modes of processing language. One thing that we, as humanists, can do to further this goal is to experiment with text-mining programs in a context that enables us to brush them against the grain, analyzing their assumptions and showing how they are positioned in the wider intellectual and cultural scene of the twenty-first century—writing, as it were, the *Tristram Shandy* to information science's *Essay concerning Human Understanding*.

This statement could perhaps be accused of encouraging the sort of navel-gazing focus on methodology that, as Cameron Blevins argues in chapter 26 of this volume, has characterized recent work in digital history; but we need not consider text-mining practices in isolation from the rest of the world ("Digital History's Perpetual Future Tense"). Sterne's *Tristram Shandy* is much more than just a satire of Locke's call for the stabilization of words; it situates this impulse among many other aspects of the life of the eighteenth-century English middle class and its relationship to the intellectual culture of its past. In a long view, text-mining software is a part of the same history that literary critics study, a twenty-first-century expression of a standardizing impulse that has had a productive (if sometimes hostile) interchange with imaginative literature for centuries and that bears complex relationships to older practices of industrial management, library organization, and philology.[8] These histories are relevant to many of the questions that more traditional forms of literary scholarship ask, bound up as they are with age-old practices of reading and writing like excerpting, cataloging, and the creation of grammars. A critical engagement with text-mining software can also help us understand those aspects of computational methods that are genuinely new, especially the use of statistical methods. Experimenting with text-mining software can highlight the strangeness of computational technology in comparison to what has come before—a strangeness that, to use a commonplace from the field of media studies,

those of us who live on the cusp of its emergence may be much better poised to see than future generations.

Engaging with text mining as an alien form of reading requires that we resist attempts to present computational results in forms that readily appeal to our assumptions and intuitions about language. The ease with which we can identify the "words" in the output of MALLET with our usual notion of the word makes it too easy to overlook the radical difference between how these units function in the program and the ways in which words can work in a human mind. While we cannot expect everyone who uses text-mining software to attempt a complete understanding of what is going on inside the computer, we should at least make an effort to appreciate the extent to which the tools we use are unknown to us, especially given the possibility that what happens inside could serve ideological ends. One can get a vivid sense of the gap between machine reading and our intuitive conceptions of language by examining the entries to Darius Kazemi's National Novel Generating Month,[9] an annual contest that challenges people to write a computer program that generates a 50,000 word novel (*NaNoGenMo 2014*). Most of the results are essentially unreadable, serving more as comments on process and algorithm than as ways of producing something that really resembles a novel, and they often ultimately direct our attention back on the role of computation itself in the generative process. For instance, Sean Connor's entry produces a randomized novel[10] by piecing together sequences of words and punctuation marks from L. Frank Baum's fourteen Oz novels. This is one paragraph of the output:

> " Same with me , please , " interrupted the girl Ruler for judgment. Again the passage turned abruptly , this time the huge scaly jaw of Choggenmugger was severed in twain and the beast advanced along the road . ("The Quantum Supposition of Oz")

The program that generated this text is based on a Markov chain model, the same sort of generative model that is commonly used in regularizing texts for the purposes of search engines, among many other applications. Although we cannot draw any major conclusions about how the technology works by reading specific examples of output, the practice of generating text using statistical models that were primarily designed for the processing of existing texts can be useful simply as a reminder of the fact that these models only loosely correspond to the way human languages work. The statistical methods that exist at present diverge in many ways from our ordinary expectations about what a text should look like, something that becomes much easier to see when one employs them for writing rather than for reading.

Text-generation programs have been employed for a number of purposes in the humanities, not all of which are specifically critical of the technology underlying them. Stephen Ramsay's *Reading Machines* proposes an Oulipan approach to literary criticism in which computerized transformations serve to enable richer

and more complex interpretations of texts (15–17). Others, such as Mark Sample, have connected text generation to Lisa Samuels and Jerome J. McGann's idea of "deformance," a practice that creates modified versions of texts as a way of exploring their autopoietic capabilities (Sample, "Notes towards a Deformed Humanities;" Samuels and McGann, "Deformance and Interpretation"). For instance, Sample's Twitter account "This is Just to Say" (@JustToSayBot[11]) produces randomly generated parodies of William Carlos Williams's poem of the same name, replacing the sweet and cold plums with something different every time. But it is also possible to think of text generation more as an interrogation of the technology itself than as a way of encountering literary texts. One way of understanding how a text generator could serve as a critique of technology is through Sean Sturm and Stephen Turner's idea of *digital caricature*. Drawing on the work of the philosopher Vilém Flusser, Sturm and Turner suggest that we think of computation as "a caricature of thinking," a diminished imitation of mental operations that can potentially be viewed as a joke (para. 30–31). Finding humor in the "drop-down menu-isation" that computers impose on design, they argue, involves understanding it not just in terms of symbolic logic, but also from the perspective of "a region of primitively evolved drives" that computers lack (para. 33). The failures of methods like the Markov chain model to produce convincing imitations of novels can serve as caricatures in just the sense that Sturm and Turner discuss, eliciting laughter because they reveal the machine's incongruity with the social world in which we expect writing to take place. While computer scientists will undoubtedly develop better software that can create more convincing imitations of human writing, employing these programs as jokes allows us to revel in their present limitations, taking the opportunity they provide to show how the mechanisms underneath the software differ from human intelligence. Given the increasing inescapability of digitally inflected modes of thought, Sturm and Turner suggest, the best way to understand what it means to be human today is to laugh at computers.

But while digital caricature can serve a useful purpose by provoking an awareness of the difference between human and machine reading, it cannot substitute for a historical perspective on these technologies. The absurd text created by novel generators can give us a visceral sense of how computational models differ from our intuitive understandings of language, but it can only get us so far in understanding how those models relate to ideology. For this we need to supplement our experimentation with text-mining methods with research that situates them historically—both in the short term, looking at the institutional contexts from which they emerged, and in the long term, looking at how they relate to the histories of linguistic thought, philosophy, communication, and labor organization. This is an area where scholars of literature and intellectual history could have a particularly productive interchange with media theorists who critically study contemporary technology. Text-mining systems are playing increasingly large roles in our lives, our teaching, and our scholarship, and digital humanists, especially those

who are versed in both statistical modeling and literary theory, are uniquely positioned to examine the linguistic ideologies that underlie them. Placing text mining in dialogue with the past could be useful not just for theorizing the implications of new scholarly tools like search engines, but also for interpreting historical texts in ways that are of particular relevance to the present shift from print to digital reading. To do this, we need a different form of scholarship from the one that applies a computer science methodology to the study of literary history. A media-studies approach would engage with programs like MALLET as cultural artifacts from the twenty-first century, products of a mechanization of language that is in some ways similar to views that have been put forth in the past, and that is in some ways new.

NOTES

1. Jockers models nineteenth-century novels in *Macroanalysis,* 118–153; Goldstone and Underwood apply topic modeling to literary criticism in "The Quiet Transformations of Literary Studies." Collin Jennings and I created a topic model for Joseph Addison and Richard Steele's *The Spectator,* available to view online at http://networkedcorpus.com/spectator/topic-index.html.

2. http://www.lisarhody.com/revising-ekphrasis.

3. I draw this conclusion from a search of the Thomson Reuters *Web of Science* database.

4. For an anecdote about a PhD student in biology being excoriated for writing "like a poet," see Ruben.

5. See Locke, *An Essay concerning Human Understanding,* especially 437–65; Sprat, *History of the Royal Society of London*; and John Wilkins, *An Essay towards a Real Character*.

6. http://wilkens.github.io/wright-topics.

7. The HathiTrust database returns 329 results for "stream of settlers" together with "Indians," constituting at least thirty distinct books.

8. On one connection between computers and industrialism, see McPherson. On accounting as a precedent for hypertext, see Duguid. On the relationship between computational linguistic techniques and philology, see Lennon.

9. https://github.com/dariusk/NaNoGenMo-2014.

10. https://github.com/spc476/NaNoGenMo-2014/blob/master/TheQuantumSuppositionOfOz.txt.

11. https://twitter.com/JustToSayBot.

BIBLIOGRAPHY

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. "Topic Detection and Tracking Pilot Study Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.* Lansdowne, Va.: February 1998.

Anderson, Benedict. *Imagined Communities: Reflections on the Origin and Spread of Nationalism.* London: Verso, 2006.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (2012): 77–84.

———. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2, no. 1 (Winter 2012). http://journalofdigitalhumanities.org/2–1/topic-modeling-and-digital-humanities-by-david-m-blei.

Blei, David, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

Bogost, Ian. *Alien Phenomenology, or What It's Like to Be a Thing.* Minneapolis: University of Minnesota Press, 2012.

Borges, Jorge Luis. "On Exactitude in Science." In *Collected Fictions,* trans. Andrew Burley, 325. New York: Penguin, 1999.

Borko, H., and M. D. Bernick. "Automatic Document Classification." *Journal of the ACM* 10, no. 3 (April 1963): 151–62.

Chun, Wendy Hui Kyong. *Control and Freedom: Power and Paranoia in the Age of Fiber Optics.* Cambridge, Mass.: MIT Press, 2008.

Connor, Sean. "The Quantum Supposition of Oz." *NaNoGenMo 2014.* https://github.com/spc476/NaNoGenMo-2014/blob/master/TheQuantumSuppositionOfOz.txt.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41, no. 6 (1990): 391–407.

Dietterich, Tom. "Overfitting and Undercomputing in Machine Learning." *ACM Computing Surveys* 27, no. 3 (September 1995): 326–27.

Duguid, Paul. "Material Matters: Aspects of the Past and the Futurology of the Book." In *The Future of the Book,* ed. Geoffrey Nunberg, 63–102. Berkeley: University of California Press, 1996.

Emerson, Lori. *Reading Writing Interfaces: From the Digital to the Bookbound*. Minneapolis: University of Minnesota Press, 2014.

Gitelman, Lisa. *Paper Knowledge: Toward a Media History of Documents.* Durham, N.C.: Duke University Press, 2014.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45, no. 3 (Summer 2014): 359–84.

Golumbia, David. *The Cultural Logic of Computation.* Cambridge, Mass.: Harvard University Press, 2009.

Guillory, John. "The Memo and Modernity." *Critical Inquiry* 31, no. 1 (Autumn 2004): 123–29.

Hockey, Susan. "The History of Humanities Computing." In *A Companion to Digital Humanities,* ed. Susan Schreibman, Ray Siemens, and John Unsworth. Malden, Mass.: Blackwell, 2004. http://www.digitalhumanities.org/companion.

Hofmann, Thomas. "Probabilistic Latent Semantic Indexing." *Proceedings of the Twenty-Second Annual International SIGIR Conference.* New York: ACM, 1999.

Jacobs, P. S., and Lisa F. Rau. "SCISOR: Extracting Information from Online News." *Communications of the ACM* 33, no. 11 (November 1990): 88–97.

Jebara, Tony. *Machine Learning: Discriminative and Generative*. New York: Springer: 2004.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History.* Urbana: University of Illinois Press, 2013.

———. "Revealing Sentiment and Plot Arcs with the Syuzhet Package." *Matthew L. Jockers* (blog), February 2, 2015. http://www.matthewjockers.net/2015/02/02/syuzhet.

———. "Some thoughts on Annie's thoughts . . . about Syuzhet." *Matthew L. Jockers* (blog), March 4, 2015. http://www.matthewjockers.net/2015/03/04/some-thoughts-on-annies-thoughts-about-syuzhet.

Juckett, David. "A Method for Determining the Number of Documents Needed for a Gold Standard Corpus." *Journal of Biomedical Informatics* 45, no. 3 (June 2012): 460–70.

Kazemi, Darius. *NaNoGenMo 2014.* https://github.com/dariusk/NaNoGenMo-2014.

Lennon, Brian. "Machine Translation: A Tale of Two Cultures." In *A Companion to Translation Studies,* ed. by Sandra Bermann and Catherine Porter, 135–46. New York: John Wiley & Sons, 2014.

Liu, Alan. "Where Is Cultural Criticism in the Digital Humanities?" In *Debates in the Digital Humanities,* ed. Matthew K. Gold, 490–509. Minneapolis: University of Minnesota Press, 2012.

Locke, John. *An Essay concerning Human Understanding* (1690). London: Penguin Classics, 1998.

McCallum, Andrew Kachites. *MALLET: A Machine Learning for Language Toolkit*. 2002. http://mallet.cs.umass.edu.

McPherson, Tara. "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." In *Debates in the Digital Humanities,* ed. Matthew K. Gold, 139–160. Minneapolis: University of Minnesota Press, 2012.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism.* Urbana: University of Illinois Press, 2011.

Rhody, Lisa M. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2, no. 1 (Winter 2012). http://journalofdigitalhumanities.org/2–1/topic-modeling-and-figurative-language-by-lisa-m-rhody.

Ruben, Adam. "How to Write Like a Scientist." *Science Careers,* March 23, 2012. http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2012_03_23/caredit.a1200033.

Sample, Mark. "Notes towards a Deformed Humanities." *Sample Reality,* May 2, 2012. http://www.samplereality.com/2012/05/02/notes-towards-a-deformed-humanities.

Samuels, Lisa, and Jerome McGann. "Deformance and Interpretation." *New Literary History* 30, no. 1 (1999): 25–56.

Sprat, Thomas. *History of the Royal Society of London, for the Improving of Natural Knowledge.* Royal Society, 1667.

Sturm, Sean, and Stephen Francis Turner. "Digital Caricature." *Digital Humanities Quarterly* 8, no. 3 (2014). http://www.digitalhumanities.org/dhq/vol/8/3/000182/000182.html.

Swafford, Annie. "Continuing the Syuzhet Discussion." *Anglophile in Academia: Annie Swafford's Blog,* March 7, 2015. https://annieswafford.wordpress.com/2015/03/07/continuingsyuzhet.

———. "Problems with the Syuzhet Package." *Anglophile in Academia: Annie Swafford's Blog,* March 2, 2015. https://annieswafford.wordpress.com/2015/03/02/syuzhet.

Underwood, Ted. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127, no. 1 (Summer 2014): 64–72.

Wilkins, John. *An Essay towards a Real Character, and a Philosophical Language.* London, 1668.

Wilkins, Matthew. *100-topic model of the Wright American Fiction corpus.* http://wilkens.github.io/wright-topics/#.