The Coding of Literary Form

Data mining and the information structure of historical texts

Dallas Liddle
Department of English
Augsburg College
Minneapolis, Minnesota
liddle@augsburg.edu

Abstract—This working paper argues that many data-mining projects in the humanities limit themselves by choosing words as their default unit of analysis. Some authors, problems, and forms are better illuminated by analysis of individual textual symbols, others by examination of multiword constructions. Insights about the nature of code from mathematical information theory, long but perhaps prematurely rejected by humanists on theoretical grounds, may give researchers less subjective and more powerful tools by which to measure the information characteristics of texts and the innovations of specific historical writers.

I. Words, Words, Words

"Data-mining" in the humanities has traditionally meant some kind of *word*-mining. From the first digitally produced concordances of the 1940s and 1950s through the "little words" approach of John Burrows' *Computation into Criticism* (1987) and the topic modeling in Matthew Jockers' *Macroanalysis* (2013) and the work of Ted Underwood and Andrew Goldstone, to today's word searches on Google Books and the "wordles" we generate to supposedly represent key idea content of texts at a glance, the majority of data-mining in the Humanities and nearly all of that done by literary scholars is lexical at its core, using some form of the "bag of words" approach of treating words as the category for which we search, and whole texts--once digitized--as matrices in which these words occur at certain frequencies.

This assumption and its enabling practices have proven genuinely useful for researchers needing to find needles of content in haystacks of text, the sorts of project Michael Hancher has appropriately termed "*datum* mining." For many other DH projects, however, literary scholars and historians word-mine not because we are interested in instances or statistical frequencies of words for their own sake, but because we assume we must apply digital tools in this way or not at all. Scholars seeking insights into the projects of authors and patterns of texts at higher cognitive levels would prefer to use our tools to "distant read" historical texts or genres or oeuvres for ideas, information, and meaning, but we know no way to measure *information* more directly--or at least no way to train a computer rather than a graduate student to recognize it. We therefore work on words, which are easy for computers to manipulate. It was possible for George Kingsley Zipf to digitally tabulate every word in James Joyce's *Ulysses* by 1949, when computers could do very little else [1]. Many scholars in the tradition of Burrows, including Massimiliano Morini, have undertaken extensive analyses of textual forms using the bag-of-words approach in the belief that sufficiently detailed statistical descriptions of sufficiently large bodies of text will produce inferential glimmers about the authorial practices, relationships, and innovations we are really interested in [2]. The results, as we know, have been mixed. Literary data-mining as a methodology has produced puzzlingly few new insights about either literary history or forms, even when its sophistication is carried to impressive lengths.

For much of what we really want to know about literary forms, bag-of-words data mining may put us in the position of focusing powerful perceptual tools on the wrong object. In an old joke, a drunken man drops his keys in the middle of a dark street, but walks over to a lamppost before beginning to look for them. He says he does this because the light is better there. Is it possible that we spend so much time counting and clustering the words in literary data not because the answers we want are likely to be illuminated by such processes, but because we assume we cannot do better--that the light of such methods is the best we have?

This short working paper proposes that we have other lights. It describes a small group of conceptual experiments in humanities data-mining that attempt to engage the problem of understanding the *information structures* of texts and groups of text by other means than through the default level of individual words. The first cases considered are small information structures below or above the level of the single word; later ones consider the forms taken by larger textual structures. As the experiments and their results are described, their theoretical implications are discussed.

II. Structures of Textual Information

Case no. 1: News by the "en"

Historians of the Romantic-era newspaper in Britain know that words were never the preferred unit of measure for the complex and high-stress production system that put out a London daily paper. Hand-press printers measured their work by the *token* of 250 single-sided page impressions; writers and editors by the column or the line of typeset copy, while compositor performance was measured by the character--the "en" of type physically set for printing. Compositors in this period were paid by the en according to rules set out in major trade agreements such as the Book Scale of 1785, superseded for news compositors by a News Scale in 1820 [3]. In other words, for newspaper practitioners at this era text mattered and was measured not as a lexical but as a volumetric phenomenon.

Commercial newspaper databases such as the *British Newspaper Archive* and *Times Digital Archive* are, like almost all humanities databases, built around word-search assumptions, and do not enable researchers to measure textual volume. If we change the questions we ask, however, the tools built for us can change their answers. Gale Cengage, owners of the *Times Digital Archive*, generously agreed to a request to data-mine the archive's original database for the total number of characters recognized by the optical character recognition (OCR) software for each discrete issue of the newspaper. Charting change over time in characters per paper between 1785 and 1810 reveals a striking pattern in how the production systems of the newspaper adapted to the French Revolution and subsequent outbreak of war with France. A visualization of this data for the first full century of the newspaper's development, soon to be published, reveals more striking patterns of how printed newspaper artifacts adapted to the competitive information pressure on newspapers that William Wordsworth described in the revised "Preface" to *Lyrical Ballads* in 1802 as the "craving for extraordinary incident, which the rapid communication of intelligence hourly gratifies."

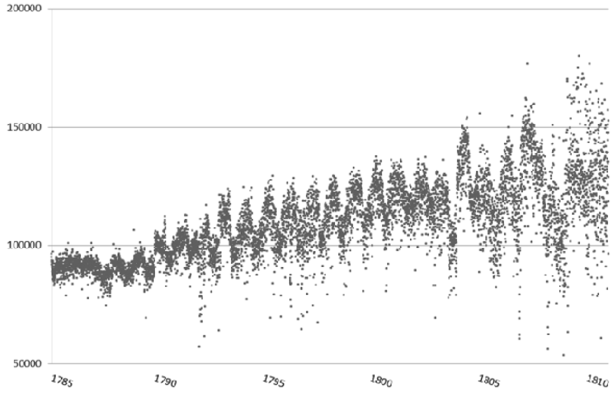

Fig. 1. Typographical characters detected per issue of the *The Times* from 1785 to 1810. Source of data: Gale Cengage.

The newspaper system appears to have responded to competitive pressure to provide readers with greater information content during wartime with a range of techniques including cyclic annual production patterns, radically increased day-to-day variability in the paper's textual volume, and a steady push to encode larger and larger absolute volumes of text onto a page of largely fixed dimensions. Henry Fielding had joked in Tom Jones (1749) that a "News-Paper" of his era "consists of just the same Number of Words, whether there be any News in it or not," but counting the actual textual volume of eighteenth- and nineteenth-century British newspapers shows them responding to demand for more information with much greater volumes of text published for substantial additional costs (this visualization also directly measures increased costs of composition, with newspaper managers paying for every additional en compositors set). Moreover, the development of the daily newspaper as a textual form in this era appears to have been intriguingly sophisticated, regular, and complex in ways that it is hard to imagine word-level searches of its texts would have revealed.

Case no. 2: Novels by the phrase

Jane Austen may be the most extensively word-investigated author in British literature after Shakespeare. Available in in numerous corpora of British texts, her *oeuvre* has been independently statistically analyzed many times--most recently by Matthew Jockers, but before him by Massimiliano Morini, and most famously by John Burrows, whose *Computation into Criticism* not only counted every word in her novels but sorted the results into parts of speech and further broke out the counts by character, by discourse (narration or dialogue), by sentence length, and even by free indirect discourse (FID), which Burrows called "character narration." Even this extraordinary project's results were not notably illuminating of Austen's abilities and projects, however. A comparison of the words Austen most commonly used with those most commonly used by her contemporary Sir Walter Scott, so stylistically different in readers' subjective experience, shows why great statistical subtlety seems requisite to draw any conclusions about Austen from her choices of individual words. In the list below, the ten most common words in the collected works of each author are listed in order.

Walter Scott	Jane Austen
the	the
of	to
and	and
to	of
a	a
in	her
I	I
his	in
that	was
he	it

With the clear exception of *his* and *he* (Scott) for *her* (Austen), the lists make the two authors appear remarkably similar--their first five words are the same, and three of those even appear in the same order. We see no immediate confirmation in these word-level statistics for so many readers' impressions of Austen's prose as dense and tight with signification, and of Scott's as characterized by relative ease and diffuseness. Jockers, who has literally written the book on how to use the statistical package R to manipulate and compare such lists, has thoroughly analyzed the characteristic patterns of supposedly unconscious use of such "little words" by scores of authors including both Scott and Austen, and based on that analysis has reached the conclusion that Austen in particular was rather stylistically unsophisticated than otherwise. "No offense to Austen fans intended," Jockers writes, "but Jane is easy to detect--her style is, as it were, an open book" [4].

Not everyone agrees that the characteristics of discourse are best measured at the level of the word, however. Linguists

Susan Conrad and Douglas Biber have found that discourse in actual "natural language" use is also significantly structured at the level of the multiword sequence or "lexical bundle" of three to five words [5]. The fact that studies have shown spoken discourse corpora to often be characterized by greater use of these lexical bundles, and written discourse corpora by fewer, suggests that information density in discourse may sometimes be coded more by the multi-word information *structures* writers and speakers construct than by the individual words they choose. In the case of Austen and Scott, expanding the search filter from one word to three, and taking the ten most common three-word constructs or "trigrams" in the collected works of each author instead of single words, immediately begins to show clearer traces of the different prose styles we expect from these two authors.

Walter Scott	Jane Austen
as well as	I do not
which he had	I am sure
one of the	she could not
at the same	it would be
the same time	in the world
part of the	as soon as
out of the	a great deal
of his own	would have been
the name of	she had been
a sort of	it was a

Scott's list of most frequent three-word sequences is dominated by prepositional phrases--six instances of "of" alone appear in his top 10, with all the 10 most common phrases clearly functioning to relate, coordinate, and qualify other ideas rather than directly express information themselves. His characteristic trigrams are low-information connectors. In Austen, by contrast, the most common forms seem to be much higher-information constructs. Seven of her first ten trigrams contain verbs, for example, while Scott has only one. Two of her common phrases use the "I" that indicates dialogue; two are negations using "not" and two are conditionals using "would." The pattern is confirmed if the lists are expanded: Austen's first 30 trigrams contain nine negations, six "I" phrases, and 20 verbs, but only three "of" phrases. Scott's top 30 trigrams contain two negations, 14 verbs, three "I" phrases, one conditional "would," and 10 instances of "of." Austen's most common trigram of all, "I do not," is simultaneously a character's dialogic utterance, a characterization of that character, a negative, and a verb--a very highly concentrated string of signification.

Data-mining Romantic fiction and Romantic news at levels above and below that of the word, in other words--by en and by phrase, respectively--immediately suggests that the structural choices encoded in historic texts do indeed have the potential to reveal important characteristics of their authors and genres. These experiments also suggest a possible need to rethink a term and category that has long been either narrowly defined or banned outright from much humanistic scholarly discourse: the word *information* itself.

III. Defining information

The small case studies just described seem to show that non-word data-mining projects have at least enough validity to demonstrate phenomena confirmable in other ways--that greater absolute volumes of text were required and used to convey more information to readers in the British newspaper in wartime, for example, and that Austen's prose characteristically contains more information, phrase for phrase, than Scott's (an assessment Scott himself seems to have shared). Describing collections of text or small passages of as "high-information" or "low-information" in the way I have been doing above, however, is usually considered both technically and theoretically invalid in literary studies. Claims about the amount of information measurably contained in text are in fact usually treated among humanists as a category error confusing a term belonging to humanistic scholarship with a more technical usage proper only to the world of communications engineering.

Professional engineers and other practitioners who study, design, and innovate digital information storage and communications systems have since the 1950s used the term "information" to refer to a statistical quantity based in the mathematics of probability. This kind of information can be defined by equations, and its objectively measurable properties have proven accurate enough to become part of the theoretical foundation of digital data storage, data compression, machine code languages, and error correction. Humanistic scholarship since the 1960s, however, has explicitly rejected the engineers' "information theory" as a methodological tool. Humanists frequently argue that information should be considered a wholly subjective human phenomenon, roughly translatable to "meaning," and impossible to measure by objective or quantitative methods. The brief mid-century controversy that resulted in this consensus, although it involved many of the most famous names of later twentieth-century literary theory, was over so quickly and decisively that most modern literary scholars probably do not know this brushfire skirmish before the Theory Wars happened at all. The settled consensus among those who do know it is that the mathematical definition of information has been determined to be a technical sub-usage or misusage of the word, with no important application to humanistic inquiry. N. Katherine Hayles rejects "hard" information as a tool for humanistic scholarship in her well-received *How We Became Posthuman* (1999) with an ideological critique of the very idea of treating information as an abstractable statistical entity, while Lisa Gitelman et al. have gone so far as to evangelize for the conversion of engineers to the humanists' definition of information in *"Raw Data" Is an Oxymoron* (2013).

If one theoretical reason more than any other causes humanistic data-mining to almost always be *word* mining, this consensus may be that reason. Humanists long ago decided--and at intervals apparently re-decide--not to recognize as valid any measurable intermediary or connecting phenomenon between the level of the words on a page and the level of meaning in a human mind. If there was such a level, and we admitted it, it might make sense to data-mine texts at that level. Since we do not, we count words instead.

Information theory may never have been as alien to conceptual and linguistic phenomena as we have persuaded ourselves, however. Claude Shannon, the Bell Systems engineer who founded a new sub-discipline of the mathematics of communication in a landmark paper of 1948, began his work from essentially the same three conceptual premises that his near contemporary Ferdinand de Saussure had been using to theorize the nature of human language. Both Shannon and Saussure believed (a) that the signs used for communication needed to be studied independently of their referential contexts, (b) that any given sign functioned only because it could be distinguished from other signs, and (c) that the relationship of difference that gives signs meaning could be understood at least initially in binary terms. Saussure famously used these insights to theorize language as an unimaginably large and complex web of signification governed by the mutual relationships among all interacting signs at a given historical moment. Shannon took the same premises in another direction entirely: into the realm of statistical mathematics. Building on the work of R.V.L. Hartley, who had mathematically modeled binary communications systems in an attempt to find the theoretical limits on information sent by electric telegraph, Shannon added the key idea that the amount of information communicated by any given amount of code had to be a function not just of the absolute number of symbols sent or the range of different symbols available, but of the amount of *uncertainty,* a probabilistic quantity, that a given quantity of code symbols could be used to resolve for the system that exchanged them. Shannon's equation for the information H in a given amount of transmitted code separated and defined all the probabilities (p_i) that each possible sign might appear, multiplied each probability by its own logarithm, then summed the negatives of the products to give the total amount of information the variable could convey:

$$H = \sum -p_i \log p_i \qquad (1)$$

For engineers, this equation and others by Shannon simultaneously revealed both the core dynamic of communicative systems and the best ways to make such systems more powerful and efficient. The key was to closely match the amount of actual *information* conveyed by a given message--defined probabilistically as the amount of uncertainty it resolved--to the specific characteristics and capabilities of the coding symbols and channel used to send it. Not just the hardware of communications channels but their coding systems themselves, once the mathematical basis of communication had been described, were technological forms that could be systematically improved. The improvements could be made in just four ways: (1) by sending or storing larger total amounts of information-bearing code (as the Romantic newspaper did), (2) by reducing the amount of "redundant" or low-information code used to the bare minimum required to confirm the message, (3) by increasing the amount of information per unit of message sent, and most effectively of all (4) by "multiplexing" the system by using its channels to send or store more than one message simultaneously (the latter three of which Austen did). All these operations, though they require very different authorial and stylistic techniques, have the same mathematical result of increasing total information H in the equation above either by expanding the size or number of sets to be summed, or by increasing the average or absolute information of elements in those sets.

Shannon's work has helped enable the extraordinary capabilities of modern communications systems we all know: full-length feature films stored on 16-gram plastic disks; hand-held e-readers with the text and images of over 6,000 standard books. It also helps explain the findings of historians such as Gerard Holzmann and Björn Pehrson that the same stages have arisen independently and repeatedly to structure the developmental histories of successful telegraphic communication systems since the torches and beacons of ancient Greece, all of which in practice worked out essentially the same information efficiencies, often by the same stages and in closely parallel ways [6]. In Romantic Britain, as Jane Austen was writing her novels and John Walter was putting out his newspaper, the admirals of the Royal Navy were inventing and implementing a coding system based on signal flags that was so information-efficient that its active service use persisted for over a hundred years. The so-called "Trafalgar code" was in fact still being preferentially used early in World War I, well after the advent of wireless. If so many real-world systems for communicating meaning show the same developmental patterns, it begins to seem much less likely that what humanists mean by information and what coding engineers mean by it are concepts as incommensurable as we have been led to believe.

IV. INFORMATION STRUCTURES IN TEXTUAL FORMS

For the engineers who create code compression algorithms, information theory provides the conceptual tools needed to find the smallest possible code size into which any message can be reduced for storage or transmission without losing any information ("lossless compression"). The key is to discover the characteristic regularities in the message, both because any regularities represent redundancy or lack of uncertainty to be re-coded (perfectly compressed code appears completely random), but also because the regularities help reveal the structural tools by which the message or other information artifact was constructed. In the last part of this paper I want to try to show that data-mining for information structures rather than words can be used not only to confirm what we mostly already knew--that newspapers grew, and that Scott can be wordy--but to reveal important patterns and practices of information compression previously unsuspected.

Franco Moretti's 2009 article "Style, Inc. Reflections on Seven Thousand Titles" is well known among digital humanists as an example of the "distant reading" methodology at the center of Moretti's recent career [7]. Most of my readers probably know how Moretti used existing catalogues to create a title list of 7,000 novels published in Britain over 110 years, counted the words in each title, and charted their lengths. Mid-eighteenth century novel titles could be any length from one word to hundreds, but frequent long titles of 15 to 40 words or more functioned as summaries or abstracts of the novel's contents. Over about two generations, however, novel titles become much shorter, their average length dropping to just a few words, and doing so along an interestingly geometrical curve.

Still more interestingly, however, the information density of titles continued to change even after they reached their smallest average size, aided by the progressive invention of structures used to pack as much information as possible into their words. Early novel titles were information-light, often using just the names of protagonists to characterize their contents--a man's name as a title signaled an episodic adventure or *bildungsroman*, a woman's a marriage plot. Moretti finds that the first alternative constructions to this were "article-noun" and "article-adjective-noun" titling patterns. An effective article-noun title--*The Monk* or *The Italian*--worked by invoking a single already striking or disruptive concept. After all, Moretti writes, "[i]f all that is in the title is a noun, then that noun must guarantee an interesting story all by itself, and vampires and parricides are a very good choice." Addition of an adjective to the title, however, meant the noun chosen could be more familiar, since the adjective could then function to introduce an unexpected modification to the noun: ″infidel fathers and posthumous daughters." Writes Moretti, "Without adjectives we are in a world of adventures; with adjectives, in a destabilized domesticity."

Somewhat later to develop, Moretti finds, were the "the x of y" structure and the abstraction. "The x of y," frequently used for Gothic fiction, works by invoking space as a threat at two levels. "The Castle of Otranto: there is a building; there is a town; they are both gothic. Escape from the castle, you're still in southern Italy. There is no way out" (157). Abstractions, introduced around 1790, took the form of a single word such as *Persuasion* or a conceptual pair such as *Sense and Sensibility*, and "made titles meaning-ful: nothing but meaning, as if the essence of the novel had been distilled and purified of all narrative contingency."

Metaphorical titles, developing just as Moretti's data runs out in 1850, raised the information stakes highest of all. Titles such as *Loss and Gain*, *Flies in Amber*, and *The Swan's Egg* are not transparent, as the abstractions are, but deliberately opaque. Moretti writes that "by puzzling and challenging readers, metaphors induced them to take an active interest in the novel from the very first word." His project shows compellingly that those who titled novels changed their behavior successfully and systemically over 100 years toward developing techniques for conveying more information in fewer words. He interprets his own finding by imagining those people acted on by a market figured as evolutionary ecology, but change by the evolutionary mechanism of random variation and selective reproduction of successful variants is not actually what Moretti's data shows. If we hypothesize that the pattern of change Moretti observes may be more technological than biological, we immediately note that the better information structures are introduced in stages, and become more sophisticated over time, as in other kinds of technological practice, and also that--as information theory predicts--within novel titles, providing the reader with actual content knowledge and deliberately creating *uncertainty* about content appear to do exactly the same work. Three of the kinds of title Moretti identifies function by specifying: proper names, article-nouns, and abstractions all provide correct information about the story. The other three structures, article-adjective-noun titles, "the x of y," and metaphors, do their work by creating contradictions, puzzles, uncertainty. For this practitioner community, puzzling and specifying apparently perform precisely the same function. It is possible, then, that Moretti's findings do not show evolutionary change so much as they show an historical adoption curve of higher-information-density structural coding solutions to the problem of novel titles in particular.

Many historians of the novel celebrate multivalent *free indirect discourse* (FID) as the key stylistic innovation to the novel introduced early in the nineteenth century by individual writers including Austen and Balzac. If I am reading the implications of Moretti's research correctly, six more discrete structural means of coding more information within fewer words, as FID does, were being developed *just for titles alone* at the same historical moment that novelists including Austen were introducing this seemingly standalone stylistic innovation. If all these techniques are recognized as structural innovations to the information density of novelistic code, then free indirect discourse, titles, and even the kind of "multiplexing" of discourse represented by Mikhail Bakhtin's "laughter and heteroglossia"--may all actually have been parts of a much larger range and spectrum of information coding innovations being introduced for fiction over this era [8].

If this does prove to be the case, we should surely be emboldened to try new ways to use computers to quantify the unusual levels of information density and efficiency authors such as Jane Austen achieved, but that we have long believed could not be assessed in any quantitative way. For example, within the last few years some corpus linguists have shown renewed interest in the possible mathematical relationships between the information efficiency of a given discourse and the rank-frequency distribution of words it displays. Rank-frequency distributions, also known as Zipf or Zipf-Mandelbrot distributions, are one of linguistics' great unsolved mysteries: the relationship between the rank of a word by its usage in a large amount of text and the absolute frequency with which words at each rank are used generally plots as a nearly straight line with slope of -1 on a log-log scale, though the reason this would be, or why certain kinds of discourse display somewhat different slopes and shapes than others when thus graphed, has never been well understood. The mathematician Benoit Mandelbrot believed it had to do with the way language was structured in response to the pressures of information, specifically as defined by information theory. Could rank-frequency analysis ultimately help complete a task usually deemed impossible, and *quantify* exactly how efficient an information coder Jane Austen was? In the visualization below, rank-frequency distributions of the words in Austen's *Pride and Prejudice* and Sir Walter Scott's *The Antiquary* are plotted on a log-log scale together with a corpus of a kind of discourse much more recently invented specifically to optimize for clarity and information density: the discourse of commercial aviation communication as recorded in the Air Traffic Control Simulation Speech Corpus (ATCOSIM) of the Graz University of Technology. The resulting visualization is at least intriguing. Although Austen's fiction is more clearly similar overall in slope to Scott's than to the deliberately information-efficient discourse of air traffic controllers, she does fall between the two, and her pattern at both extremes of the scale bends toward theirs.

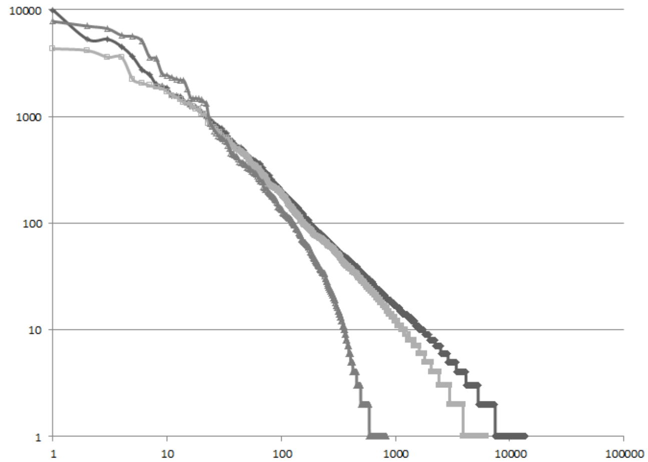

Fig. 2. Rank-frequency "Zipf" distributions for Scott, *The Antiquary* (dark shading), Austen, *Pride and Prejudice* (light shading), and the ATCOSIM flight traffic control simulation corpus (medium shading)

V. Conclusion

Modern Digital Humanists generally use the term *coding* to mean something we do to texts, not something their authors did to create them in the first place, and the nature and complexity of which can potentially be understood and measured. We set our methodological sights rather lower. For all the hopes and high expectations surrounding digital claims and projects, most humanists who use electronic tools still accept conceptual limits on the questions we can ask that have largely been inherited from poststructuralist theory. Our outward optimism, then, may well mask what Elijah Meeks and Scott B. Weingart call the "jaundiced realization that computational techniques like topic modeling ... are not an upgrade from simplistic human-driven research, but merely more tools in the ever-growing shed." [9] Many of us still believe, with the Terry Eagleton of *Literary Theory*, that it must be impossible to find, and therefore naïve to believe in, any pattern of textual encoding that would enable us on objective grounds to determine what information structures distinguish literature from non-literature, fiction from nonfiction, eighteenth-century fiction from nineteenth, realist fiction from romance, or, most of all, information-efficient fiction from inefficient fiction. Perhaps it was the generation of Theory Wars, fought to their enervating standstill, that taught so many of us to overestimate the complexity of our subject and the intractability of its major research problems, and to underestimate the capabilities of our conceptual tools.

I believe that the information-bearing codes of literary forms, while extraordinarily dense with signification, are probably not infinitely or unmeasurably so, even potentially. They are fundamentally just codes, succeeding by the same standards and governed by the same laws as all other communication codes. Once the use of digital tools to study texts in the humanities has become a mature theoretical enterprise, the most truly extraordinary discovery we could make would not be that historical discourses and literary texts have in fact been always been shaped by the mathematical relationships predicted by information theory. It would be far more remarkable to discover that literary communication alone has somehow been exempt from laws that govern, limit, and enable the way every other system exchanges information, from the level of the Internet to the level of the cell.

References.

[1] G. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Mass.: Addison-Wesley, 1949 [2] M. Morini, Jane Austen's Narrative Techniques: A Stylistic and Pragmatic Analysis, Ashgate, 2009.

[3] E. Howe, The London Compositor, 1785-1900. London: The Bibliographical Society, 1947.

[4] Jockers, Matthew L. Macroanalysis: Digital Methods & Literary History. Urbana, Chicago, and Springfield: University of Illinois Press, 2013, p. 93.

[5] S. Conrad, D. Biber, "The Frequency and Use of Lexical Bundles in Conversation and Academic Prose," Lexicographica 20 (2004), 56-71.

[6] G. Holzmann, B. Pehrson. The Early History of Data Networks. Washington and Brussels: IEEE Computer Society Press, 1995.

[7] F. Moretti. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850)." Critical Inquiry 36 (Autumn 2009): 134-158.

[8] M. M. Bakhtin. "From the Prehistory of Novelistic Discourse." The Dialogic Imagination: Four Essays. Ed. M. Holquist. Austin: University of Texas Press, 1981, 41-83.

[9] E. Meeks, S. Weingart, "The Digital Humanities Contribution to Topic Modeling," Journal of Digital Humanities 2.1, Winter 2012, http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modelin