CA Journal of Cultural Analytics

Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel

Dallas Liddle

02.21.19

Peer-Reviewed By: Katherine Bode, David Bamman

Clusters: Genre

Article DOI: 10.22148/16.033

Dataverse DOI: 10.7910/DVN/M7AGWJ

Journal ISSN: 2371-4549

Cite: Dallas Liddle, "Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel," Journal of Cultural Analytics. February 21, 2019.

"More matter with less art."- *Hamlet* 2.2

Literary scholars in the information age, whatever issues we may differ on, have long shared a belief that the amount of "information" in literary language is not a phenomenon that can be empirically investigated.[1] We tell our students, and remind each other, that the meaning in literary texts is inherently subjective, culture-specific, and contextual rather than fixed. Scholar Katherine Bode reaffirms this article of faith in a new book on digital literary methodology, writing that "literary data are inevitably constructed and transactional, whether they are

[1] The author is grateful to the editor and reviewers of CA for generous feedback on earlier drafts of this article, which led me to ask new questions, run new tests, and seek out resources I had entirely missed. Everything still missing in the current version is my sole responsibility. Some parts of the work described here were supported by a 2010 NEH Summer Stipend, others by a 2016 Advanced Collaborative Support Grant from the HathiTrust Research Center.

explicitly designated so or not."[2] Long prose forms such as the novel seem especially immune to empirical assessment. As Thomas Pavel notes, the wide-open form of the novel allows "for any imaginable kind of confabulation without constraint,"[3] while the fact that each reader can reinterpret and recontextualize a novel anew seems to guarantee that the meanings latent in such texts can never be fully described, much less quantified. Probably because both digital and traditional humanists hold these beliefs, most data mining, topic modeling, sentiment analysis, and other distant reading methods used on fiction search for subtle patterns or relative thematic trends across time, or for faint signals of style detectable in genres or the works of individual writers. We assume that the total quantity of information contained in novels could never be computationally estimated by any measure humanists would care about.

The reported outcome of a digital humanities experiment by Mark Algee-Hewitt of the Stanford Literary Lab, published online in the multi-authored Pamphlet 11, "Canon/Archive: Large-Scale Dynamics in the Literary Field" (2016), is therefore both anomalous and interesting.[4] Algee-Hewitt and colleagues compared two large sets of nineteenth-century novels: 250 works in a proprietary e-text collection,[5] representing the canon, and a larger number of texts from the general population of novels published in the same era, representing the archive. They divided all texts into collections of bigrams (word pairs) which they assessed for mathematical "redundancy" using equations based on the discipline of Information Theory founded by Bell Systems engineer Claude Shannon in 1948.[6] The canonical novels displayed different statistical characteristics than the archival ones, and the effect was large: "that three-fourths of the Chadwyck-Healey collection

[2] Katherine Bode, *A World of Fiction: Digital Collections and the Future of Literary History* (Ann Arbor: University of Michigan Press, 2018), 96.

[3] Thomas G. Pavel, *Fictional Worlds* (Cambridge: Harvard University Press, 1986), 2.

[4] Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, Hannah Walser, "Canon/Archive: Large-scale Dynamics in the Literary Field," *Stanford Literary Lab Pamphlet 11*. Jan. 2016.

[5] For a description of the collection see "About Nineteenth-Century Fiction," *Chadwyck-Healey Literature Collections*. c. 1996-2016.

[6] The mathematical Information Theory invoked and used by Algee-Hewitt et al. is not a standard area of undergraduate mathematics, but a number of excellent introductions and guides have been written. For a brief and readable overview I recommend the background section of M. B. Plenio and V. Vitelli, "The physics of forgetting: Landauer's erasure principle and information theory," *Contemporary Physics* 42.1 (2001): 25-60, available as a PDF from the Imperial College London website. A longer introduction aimed at the generalist is John R. Pierce, *An Introduction to Information Theory: Symbols, Signals, and Noise*, rev. ed. (New York: Dover, 1980), which was reviewed in manuscript by Claude Shannon. The standard textbook for Information Theory at the graduate level is Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory* (New York and Chichester: John Wiley, 1991). To go back to the source see Claude E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal* 27 (July and October 1948): 379-423, 623-656, a reprint of which is available on the website of the Mathematics department at Harvard.

would be less redundant than three-fourths of the archive was a *much* stronger separation than we had expected to find."[7] The Stanford authors seem to have found that for nineteenth-century fiction a purely statistical feature of texts—the redundancy of their words, indicating higher or lower information density in mathematical terms—has a strange power to predict the value that would be assigned to those texts by historical readers, markets, and scholars.

The Stanford authors' somewhat conservative interpretation of this result was to see it as confirming literary conventional wisdom: of course less-celebrated authors would use language more redundantly. "The clarity of the contrast had simply confirmed a received idea: forgotten authors used language in a redundant fashion; if they had remained unread, it was because they weren't really *worth* reading [….] Not exciting, corroborating a received idea."[8] When they attempted to determine more exactly what canonical novels did differently than archival ones, however, the Stanford authors found the unreadably large lists of bigrams generated by their method a methodological impasse: "Here, statistical significance seemed impervious to critical meaningfulness."[9] They proceeded to lexical (type-token) and grammatical (part of speech) comparisons of the canon and archive texts, but these generated less striking results, and they concluded that their bigram redundancy test had led to a dead end.

In presenting their experiment as interesting but uninterpretable Algee-Hewitt et al. demonstrate modesty and scholarly scrupulousness, but seem to miss the most significant implications of their result. It is true that we expect skilled writers not to write redundantly, but not true that any reputable literary humanist of the past half-century familiar with what Claude Shannon called the mathematically-defined "information" of a message would expect the *semantic* redundancy human readers recognize in literary texts to covary meaningfully with the *statistical* redundancy computers can calculate, or that any mathematical analysis could predict a judgment as complex, contingent, and human as the literary value of a book. Generations of scholars have believed the opposite. A young Noam Chomsky, formulating his own framework for linguistic investigation in the 1950s, coined the nonsense line "Colorless green ideas sleep furiously" specifically to demonstrate that mathematical information theory's probabilistic approach could never explain high-order human use of language.[10] After Chomsky, Fr. Walter Ong attacked information theory as simplistic and mechanistic in a succession of books and articles over more

[7] Algee-Hewitt et al., 6.
[8] Algee-Hewitt et al., 6.
[9] Algee-Hewitt et al., 7.
[10] Noam Chomsky, "Three Models for the Description of Language," *IRE Transactions on Information Theory* 2.3 (1956), 116.

than a decade, while Colin Cherry, who had experimented with it, came to describe the theory as a "blind alley" for human communication.[11] The belief that information theory was useless for humanist inquiry seemed to have the endorsement even of Claude Shannon, whose 1948 paper on the mathematics of communication stipulated that "semantic aspects of communication are irrelevant to the engineering problem" with which his theory was concerned.[12] The current 3rd edition of the *Oxford English Dictionary* defines "information" in the statistical sense Shannon established (2c.) as "a mathematically defined quantity divorced from any concept of news or meaning," and adduces a quotation from M.A.K. Halliday that "Information theory, which has a place in the quantitative description of a language, implies nothing about the relative efficiency of languages or the effectiveness of language activity."[13] The leading authority on cybernetics in modern fiction, N. Katherine Hayles, has long opposed use of information mathematics for literary study, writing as recently as 2014 that for humanists "a theory totally divorced from meaning has little to contribute."[14]

The consensus of so many major literature and communications scholars would carry weight even if the idea of a relationship between statistical information and literary value did not seem so unlikely on its face. It is counter-intuitive to say the least to imagine that a Victorian publisher circa 1850 who divided the texts of novel manuscripts into bigrams and laboriously calculated and summed the likelihood p_i that the bigrams would appear, using Shannon's equation for the information content of a message source,

$$H = -\sum p_i \log p_i$$

would have had in H a reasonably good indicator of which works might go on to become classics. The Algee-Hewitt et al. result suggests that, despite all intuition,

[11] Cherry told an interviewer in the 1970s that Information Theory "doesn't help us" much to understand human communication [….] It was a blind alley." See Carol Wilder, "A Conversation with Colin Cherry," *Human Communication Research* 3.4 (1977), 356.

[12] Shannon's position on this issue is usually overstated, however. While he did write that the semantic aspects of communication were irrelevant to the engineering aspects, he never claimed the engineering aspects were equally irrelevant to semantic communication. The metaphor that statistical information is "divorced" from meaning is therefore misapplied. Divorces make a mutual separation; Shannon's claim of separation went only one direction.

[13] "information, n." OED Online. June 2017. Oxford University Press. (accessed December 21, 2017).

[14] N. Katherine Hayles, "Cognition Everywhere: The Rise of the Cognitive Nonconscious and the Costs of Consciousness," *New Literary History* 45 (2014), 216-217.

this would have been true. Their finding of an apparent association between subjective literary value and objective information density has potential to call some of literary studies' longest-standing assumptions into question.

It should be emphasized that the experiments that follow exploring this potential association do not attempt to directly replicate the Stanford methods—fortunately so, since Pamphlet 11 is thin on the methodological details that would enable replication.[15] As a historian of genre and the nineteenth-century novel my interest is less with quantifying canonicity than with studying how or whether the information dynamics Algee-Hewitt et al. may have detected could have influenced historical authorship and the novel's development. Technology historians established decades ago that competitive pressure to employ more mathematically efficient codes shaped the development of communications technologies from signal beacons and semaphore flags to the telegraph, telephone, and fiber-optic cable.[16] In recent years literary scholars have sometimes characterized the novel as a technology also.[17] But no researcher seems to have carried the metaphor far enough to ask what it now seems possible to at least ask: whether the mathematics of information could have shaped the developing practices of novelists as well as those of engineers and coders.

Measuring the "information" in nineteenth-century novels: method and results

Luckily for literary historians looking for ways to test a potential relationship between Information Theory and literary history, theoretical and applied computational linguists have had fewer objections to statistical information than their humanist counterparts. The reader probably uses many practical achievements

[15] Neither Pamphlet 11 nor the Literary Lab website share the original results, list of archival texts, or the code of the algorithm used to generate their redundancy figures. Like the other Stanford pamphlets, the work was also not subjected to peer review. For the reasoning behind the latter decision see Franco Moretti, "Literature, Measured," Stanford Literary Lab Pamphlet 12, April 2016.

[16] An excellent study of how the development of digital and non-digital communication networks were influenced by the same patterns and forces is Gerard J. Holzmann and Björn Pehrson, *The Early History of Data Networks* (Washington and Brussels: IEEE Computer Society Press, 1995).

[17] See for example Richard Menke, *Telegraphic Realism: Victorian Fiction and Other Information Systems* (Stanford: Stanford UP, 2008); Cara Murray, *Victorian Narrative Technologies in the Middle East* (New York: Routledge, 2008); and Tony Jackson, *The Technology of the Novel: Writing and Narrative in British Fiction* (Baltimore: Johns Hopkins UP, 2009).

of computational linguistics daily: spelling correction in word processors, next-word predictors in text messaging, spam filters in email, and speech recognition tools in digital assistants are all built on information-theoretical algorithms. As Kenneth W. Church recently explained, Shannon's "noisy-channel model makes lots of sense for speech recognition, OCR, and spelling correction" while Naïve Bayesian statistical algorithms are valuable for spam filtering, author identification, sentiment analysis, and word sense disambiguation.[18]The assumption behind most of these systems is that they are only ways to enable computers to *seem* to process language as humans do, however, not that they reproduce or even illuminate real human processes. Church notes that speech recognition systems distinguish ambiguous words using "confusion matrices" created by supervised algorithms that have processed trillions of words of natural-language corpora—not quite the way people do it.

Researchers in comparative linguistics have established at least one method by which an information-theoretical tool may help illuminate issues of human language use, however. Patrick Juola, Max Bane, and others have used a simple information-theoretical algorithm to measure the information densities of the same document across different languages, such as translations of the Bible and European Union Constitution, as they explore the question of whether all human languages are equally complex.[19] The quantity these researchers seek to determine for each translation of the same text is its *Kolmogorov complexity*, an alternative way to determine Shannon information in which the information content of a text string is defined as the length of the shortest computer program that would output that string and then stop. To show what that means, these two text strings both have exactly 38 ASCII characters:

[1] CoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCoCo

[2] Colorless green ideas sleep furiously.

The strings may contain the same absolute number of text characters, but do

[18]Kenneth W. Church, "Statistical Models for Natural Language Processing," *Oxford Handbook of Computational Linguistics 2nd edition,* ed. Ruslan Mitkov. Online publication date: Aug. 2016. DOI: 10.1093/oxfordhb/9780199573691.013.54.

[19]See for example Max Bane, "Quantifying and Measuring Morphological Complexity," in *Proceedings of the 26th West Coast Conference on Formal Linguistics*, ed. Charles B. Chang and Hannah J. Haynie (Somerville, MA: Cascadilla Proceedings Project, 2008), 69-76. Available: www.lingref.com, document #1657. Patrick Juola, "Assessing linguistic complexity," is in Matti Miestamo, Kaius Sinnemaki and Fred Karlsson, eds., *Language Complexity: Typology, Contact, Change* (Amsterdam/Philadelphia: Benjamins, 2008), 89-108. The compressibility of 21 different translations of the EU Constitution was tested by Markus Sadeniemi, Kimmo Kettunen, Tiina Lindh-Knuutila, and Timo Honkela in "Complexity of European Union Languages: A Comparative Approach," *Journal of Quantitative Linguistics* 15.2 (2008): 185-211.

not have the same quantity of mathematical information. In Shannon's original terms, that is because the first string is so predictable—after reading the first few characters the reader could easily pick the next in line until the end. In Kolmogorov's terms, the same attribute can be measured by finding the length of the shortest computer command needed to recreate each string. String [1] is so highly patterned and predictable that it could be reproduced by a simple instruction to print "Co" 19 times. String [2] is so unpatterned and unpredictable that it would have to be reproduced at close to full length as part of any machine instruction. String [2] contains greater mathematical information than string [1] by both Shannon's and Kolmogorov's criteria of measurement.

An exact final value for the Kolmogorov complexity of a text string is formally incomputable,[20] but linguists consider it well established that Kolmogorov complexity and compressibility are so nearly the same phenomenon[21] that the Kolmogorov complexity of a text can be estimated using off-the-shelf file compression programs such as WinRAR, WinZip, gzip, and Zlib.[22] As Juola writes, "If [...] Kolmogorov complexity represents the ultimate possible file compression, a good file compressor can be seen as an attempt to approximate this kind of complexity within a tractable formal framework."[23] By compressing multiple translations of *Alice in Wonderland* as tightly as possible and comparing the ratios of the full-sized and compressed versions, Katharina Ehret and Benedikt Szmrecsanyi produced a table of comparisons of ten languages' relative complexity in line with what more orthodox complexity notions would lead one to expect" (e.g. Hungarian led the overall list).[24] The method is not limited to use on parallel texts, however. In a different project they themselves call a "proof-of-concept study," Ehret and Szmrecsany assessed essays by second-language learners of English at various degrees of mastery, finding that "essays by more advanced learners tend to be more Kolmogorov complex than essays by less advanced learners."[25]

[20] See Ming Li and Paul Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, 2nd ed. (New York and Berlin, Springer, 1997).

[21] "Fortunately, it can be shown that all reasonable choices of programming languages lead to quantification of the amount of 'absolute' information in individual objects that is invariant up to an additive constant. We call this quantity the 'Kolmogorov complexity' of the object. If an object contains regularities, then it has a shorter description than itself. We call such an object 'compressible.' " Li and Vitányi, v.

[22] Ehret and Szmrecsanyi use the utility gzip, while Sadeniemi et al. use Bzip2. Juola, who has used both, reports they produce similar results; see Juola, 102n3.

[23] Juola, "Assessing Linguistic Complexity," 92.

[24] Katharina Ehret and Benedikt Szmrecsany, "An Information-theoretic approach to assess linguistic complexity," in R. Baechler and G. Seiler, eds, *Complexity, isolation, and variation* (Berlin/Boston, MA: De Gruyter, 2016), 74.

[25] Katharina Ehret and Benedikt Szmrecsanyi, "Compressing learner language: An information-theoretic measure of complexity in SLA production data," in *Second Language Research* (2016): 10.

In the experiments below, instead of using a word-level assessment such as the bigram tool the Stanford researchers employed, full-text corpora of large numbers of novels were assessed for the compressibility of their texts at the more fundamental level of the byte and ASCII code character by the Kolmogorov complexity method used by the linguists. The compression utility chosen was WinRAR, set to best possible compression using the "Zip" method, so that the algorithm applied would be DEFLATE. This industry-standard compressor uses a combination of Huffman coding, which finds the relative rates at which alphanumeric characters, read as codes in ASCII, are used overall in a file, and substitutes the most space-efficient codes possible (the commonest characters get the shortest codes, the rarest characters the longest), and Lempel-Ziv coding, which searches for repeated strings of code at the byte level and replaces these with index markers, creating a much smaller file capable of re-generating the original text. Text files packed by such combinations of Huffman and Lempel-Ziv coding often require as little as a third of the storage space needed by the original file.

To make the experimental praxis as transparent as possible, only freely available textual archives have been used, so that any reader with an Internet connection should be able to replicate the experiments. Large sets of texts were also sought; Ehret and Szmrecsanyi note that the compressibility ratios of individual files are hard to interpret in the abstract, and best evaluated in comparative contexts.[26] All other tools and methods used are identified as they appear, and all spreadsheets used to create the visualizations shared. For terminology, since the numbers calculated are actually file compression ratios, that is the language I will use for them, rather than claiming that they represent (rather than approximate) Kolmogorov complexity.

A first attempt to assess the evidence for information-theoretical pressures on the historical development of the novel using this method was performed on Andrew Piper's freely available .txtLAB collection of 451 novels in English, French, and German published between the 1770s and 1930s.[27] Each novel in the sets of 150 or 151 novels from each national tradition was compressed with DEFLATE with WinRAR and the ratio of full-size to compressed size for each work in the English, French, and German collections calculated and displayed over time. The strong expectation literary historians should surely have for such an experiment—and also the null hypothesis—would be that the machine compressibility of a given literary text will turn out to be meaningless metadata, which is how literary scholars have always treated it. Compressibility from text to text could be expected to vary slightly based on random factors and stylistic peculiarities of individual

[26] Ehret and Szmrecsanyi, "Compressing Learner Language," 19.
[27] Andrew Piper, txtlab Multilingual Novels, 2016. Figshare.

works and authors, but seems unlikely to trend over time, and the corpus as a whole should show noise rather than signal. A result that found a real trend in such data, however, would suggest that the average information density of fiction in one or more of these national traditions really did change over time, though it would not in itself identify the reason.

Scatterplots of the relative compressibility over time for the English, French, and German novels in the .txtLAB collection are visualized side by side in Figure 1 below, made using R with ggplot2.[28] The reader is encouraged to download the original archive from Figshare, and/or the spreadsheet associated with Figure 1, to reproduce and explore the patterns for themselves.

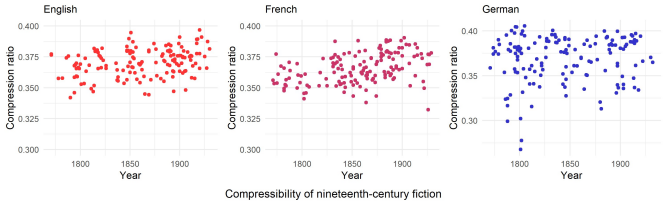

Figure 1. Compression ratios ("packed" size divided by "unpacked" size) of 150 German, 150 English, and 151 French novels published between the 1770s and 1930s. Higher numbers indicate lower compressibility. Source of data: Andrew Piper, .txtLAB.

All three visualizations are noisy, but all also suggest an upward trend over time toward less compressibility and greater density of information in each nation's novels. To assess whether the trends are real or only statistical noise, Mann-Kendall tests were performed using A.I. McLeod's R package Kendall.[29] Mann-Kendall is a test for change in independently collected observations, often used to look for trends in time series of environmental samples. The result for the German novels was to show only a slight upward trend (tau = 0.0633) in information density, and not to a statistically significant degree (2-sided pvalue = 0.25128). The English novels, however, showed a stronger upward trend (tau = 0.183) at high statistical significance (2-sided pvalue = 0.0009166). The French novels showed an even more marked upward trend (tau = 0.289) and higher statistical significance (2-sided pvalue = 0.00000011921). The apparent increase over time in the information density of English and French fiction, at least, does not initially seem to be a function of random variation alone.

[28] R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org; and H. Wickham, "ggplot2: Elegant Graphics for Data Analysis" (New York: Springer-Verlag, 2016).

[29] A.I. McLeod, "Package 'Kendall'" 16 May 2011.

Statistically significant or not, the English and French results are so odd that it seemed important to try to rule out alternative explanations for what seems to be a shared movement by literally hundreds of French and British novelists to write more mathematically information-dense fiction as the nineteenth century went on. Other explanations for these numbers than a genre-wide evolution toward more dense information coding practices for fiction are possible, and more intuitively likely. Perhaps over this period the average novel in both countries got longer or shorter in absolute terms, and the apparent increase in information density only shows the compression utility's efficiency in packing larger or smaller files. Perhaps discourse in this period underwent a more general cultural change, and novelists' changing language reproduced a larger cultural trend toward more varied, concrete, or otherwise information-dense expression. When Stanford's Ryan Heuser and Long Le-Khac discovered a trend away from abstract words and toward more concrete ones in British fiction over this same period, a cultural trend was their preferred explanation, and other scholars have found this idea convincing.[30]

To test the first alternative explanation for the increasing information density of these novels, that the effect is due only to novels changing in absolute size, the simple uncompressed file size of all three sets of novels was visualized and more Mann-Kendall trend tests run.

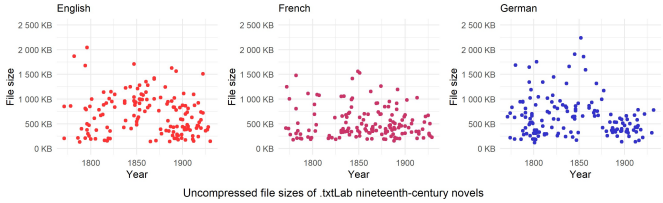

Figure 2. Absolute size of the original text files in Andrew Piper's collection of 451 English, French, and German novels.

In all three cases the average size of each nation's novels decreased very slightly over time (English tau -0.0939, French tau -0.008539, German tau -0.0962), but none of the decreases was statistically significant at $p < 0.05$ (2-sided pvalues 0.088572, 0.87979, 0.080945). When the relationship between compressibility and original file size was checked with the full suite of correlation tests available

[30] See Ryan Heuser and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," Literary Lab Pamphlet 4, May 2012. Alan Liu's endorsement of their approach appears in "The Meaning of the Digital Humanities," *PMLA* 128 (2013): 409-423. More recently Matt Erlin, in "Topic Modeling, Epistemology, and the English and German Novel," *Journal of Cultural Analytics* 1 May 2017, Article DOI: 10.22148/16.014, Dataverse DOI: 10.7910/DVN/3J38FX, interprets new research findings of his own as corroborating theirs.

in R (Pearson, Kendall, and Spearman), however, the compression ratio of the French novels, though not the others, did covary significantly with the original size of the file. This should not have been unexpected, since many more French novels in the Piper collection are short, and the comparative linguists who have used this method observe that file compression can behave erratically on smaller files.[31] To control for this effect the original texts of all three sets of national novels were concatenated into single very large text files which were then divided into exactly equal 1-megabyte slices (99 English, 77 French, 94 German). The compression algorithm was re-run on the slices, with the result shown in Figure 3. When more Mann-Kendall tests were performed, English fiction continued to show a statistically significant rise over the period (tau = .199, 2-sided pvalue = 0.0034964), French fiction maintained its still more significant increase (tau = .403, 2-sided pvalue = 2.3842e-07), and German fiction continued to show no significant effect over time (tau = -0.03, 2-sided pvalue = 0.67112). The apparent historical increase in the information density of French and English novels does not appear to have been generated by the compression utility.

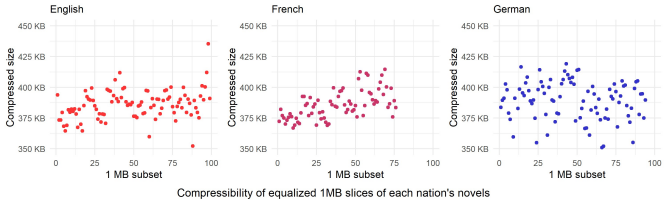

Figure 3. Post-compression sizes of 1MB text file slices of the .txtLAB collections of English, French, and German novels.

The second alternative explanation, that the result reflects some change in discourse within the wider culture, is more difficult to rule out. Corpus linguist Michaela Mahlberg, who studies represented speech in Dickens, has expressed frustration that corpus-to-corpus comparisons of dialogue in fiction with actual Victorian conversations are literally impossible to make, and also that in her experience peer reviewers expect them anyway.[32] Fortunately, nineteenth-century Britain does have one body of continuously "captured" public discourse, which even if far from ideal at least provides a body of single-genre comparison text cre-

[31] Ehret and Szmrecsanyi, "Compressing Learner Language," 19.

[32] In a 2016 lecture Mahlberg observes, "I get this a lot. When you send an article to a journal, then these reviewers give you these comments, 'This is quite a nice idea, but actually you really need to compare this to a real spoken corpus,' and like, that's a very good idea, can someone show me a real spoken corpus of nineteenth-century language and I'll happily do this." Michaela Mahlberg, "Corpus linguistics and the challenges of close and distant reading," 11th Sinclair Open Lecture, University of Birmingham, 20 Jun. 2016. *Youtube,* time index 20:40-21:07.

ated over the same period as nineteenth-century fiction. From the 1780s onward London morning newspapers posted teams of reporters in Parliament, Charles Dickens briefly among them, to work in shifts to take notes and write out textual replicas of the debates (in compressed third-person for minor Members, "full" first-person for party leaders) to pass to newspaper compositors for immediate publication in the next morning's newspaper.[33] These on-deadline transcripts were collected, edited, and republished as a quasi-official public record in *Hansard's Debates*, the nineteenth-century series of which are now available on the web.[34] If British novels in this era became denser only because of a culture-wide movement toward more information-dense discourses, the compressibility ratios of parliamentary transcripts over the same period might reflect the same trend. To keep the comparison fair, the calculation was limited to the period from 1803 until 1878, the year *Hansard* began to use a reportorial staff of its own in addition to collecting and collating the newspaper reports.[35] The trace this experiment produced is visualized in Figure 4 below.

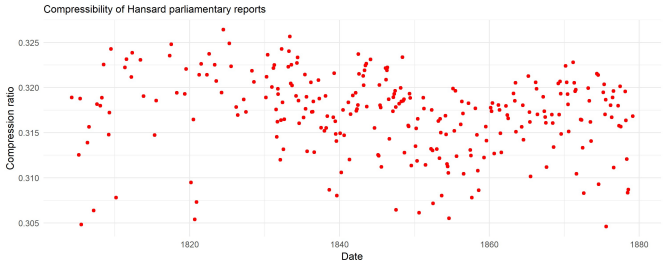

Figure 4. Compressibility of *Hansard* parliamentary debate transcript text files from 1803-1878. Source of data: Hansard.

A visible trend—unexpectedly more down than up this time—in the information

[33] An authoritative and readable overview of this system and its history is John Vice and Stephen Farrell, *The History of Hansard* (London: House of Lords Hansard and the House of Lords Library, 2017).

[34] See "Hansard Archive (debates from 1803)."

[35] Some other manipulations of this data were necessary and should be noted. The text files are kept at http://www.hansard-archive.parliament.uk/, but the index to sittings per volume is at https://api.parliament.uk/historic-hansard/volumes/index.html. The two resources are not perfectly aligned; 40 volumes of Hansard are listed in the index but not present in the zipped files, while 24 volumes had a zipped text file but did not appear in the index. In the data set these volumes are simply omitted. Another five volumes had both index entries and a file but the dates of the sittings were listed as "unknown." For these I inserted dates between those of the volumes that came before and after, as follows: s1v9 (15-8-1807 to 20-1-1808), s1v13 (8-3-1809 to 10-4-1809), s1v21 (25-7-1811 to 16-3-1812), s1v24 (5-5-1812 to 3-11-1813), s3v108 (2-8-1849 to 25-2-1850). Finally, the Hansard files download heavily laden with xml codes; to make sure compression was being done mostly on the original report texts these codes were removed using the text editor Notepad++.

density of Hansard-collected newspaper reports over the first three quarters of the nineteenth century seems to appear, but statistical significance was assessed anyway using the R Kendall package. The result was a negative trend (tau = -0.166) in the information density of debate transcript texts, to a strong degree of statistical significance (2-sided pvalue = 0.000040773). Over the period English fiction was apparently getting more information-dense, Parliamentary discourse, at least as newspapers reproduced it, was getting a bit less so.

How can this result be explained? One potentially important factor is suggested by the changing volume of parliamentary discourse transcript the teams of newspaper reporters were expected to produce for each parliamentary sitting, shown in Figure 5 below, which increased steadily, substantially, and significantly over the period (tau = 0.543, 2-sided pvalue = 2.22e-16). A weak but significant inverse correlation (cor = -0.1626734, p-value = 0.006967) appears to hold between the amount of transcript reporters were required to submit for each sitting and the information density of the resulting transcript. In other words, as the total volume of language the reporters had to produce to the same daily deadline for a given evening of speeches increased, the information density of the text those reporters produced decreased.[36]

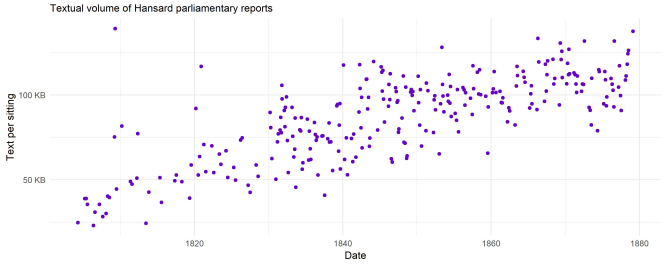

Figure 5. Amount of transcript text published per Parliamentary sitting in Hansard, 1803-1888, in average amount (in KB) of uncompressed text per event. Source of data: Hansard.

The *Hansard* results do not appear to support the alternative hypothesis that novels in Britain might have become more information-dense over the nineteenth

[36] Interestingly, at least one veteran newspaper reporter of this era appears to have hypothesized a necessary tradeoff between time pressure and textual density. Charles Ross, who reported Parliamentary debates for *The Times* from 1820 until 1883 and was its Chief of Reporters from the mid-1850s, told a Parliamentary subcommittee in July 1878 that deadline pressures had made reporting "too much of a mechanical art at present; giving the mere words." Ross preferred that "there should be compression instead of expansion; you want a man's opinions, not his words," but told the subcommittee that "under the existing system it cannot well be done." Great Britain, Parliament, *Reports from Committees*, 1878. Vol. 8. "Report from the Select Committee on Parliamentary Reporting," 39.

century simply because discourse or language use in the wider culture was doing so. If anything, novels became more information-dense in this period despite a slightly contrary trend in public discourse, although one experiment on a corpus created for other purposes is not strong evidence that a wider trend existed at all. The results do suggest, however, that pressure to produce more text in the same amount of time acts to lower the information density of the text thus produced.

Could the same cause have produced the same effect on literary composition, and could the relative compressibility of novels by individual authors over time perhaps even reveal information histories within individual artistic careers? With a trend toward greater information density in nineteenth-century English and French novels not disconfirmed by the Piper and Parliamentary experiments, it seemed important to ask whether this effect might also be visible at the level of the author. A productive author with a high total count or *n* of text files seemed the best way to produce meaningful results. Anthony Trollope published nearly 50 novels, in addition to short story collections and several nonfiction works, over a long career that also happens to have been divided into two quite different approaches to composition. Until 1860 Trollope's productivity as a novelist, though impressive by modern and most Victorian standards, was slow by comparison with his own later pace. The son of popular novelist Frances Trollope, Anthony Trollope wrote and published novels in this early era at intervals of a year to a year and a half or more, winning little critical or financial success and keeping his primary professional focus on his work as a Post Office civil servant. Once he was chosen by William Makepeace Thackeray to become a regular contributor of serial fiction to the *Cornhill Magazine*, however, Trollope became one of the most productive major writers of the Victorian era, rapidly completing multiple large novels per year for most of the rest of his life using the disciplined daily methods and word count targets described in his *Autobiography* (written by 1878, published 1883).[37]

To see if Trollope's shift to a faster textual production pace had any effect on the statistical information density of the resulting work, all 47 of his novels were downloaded from Project Gutenberg and the texts of each file stripped—for this purpose only—of editorial apparatus and Project Gutenberg legal language.[38] Originally a Python routine was written to cycle through the text files, compress each, calculate the degree of compression, and plot the result with Matplotlib.[39]To keep methods and comparisons consistent in this paper, the experi-

[37] Anthony Trollope, *An Autobiography* (Oxford and New York: Oxford University Press, 1980).

[38] I assure Project Gutenberg's legal representatives that the truncated text files were created for research purposes only and have not been shared.

[39] This part of the project would not have originally been possible without technical help and advice provided through a 2016 Advanced Collaborative Support Grant from the HathiTrust Research

ment was rerun with files compressed with WinRAR and visualized in R to produce Figure 6. Once more, the null hypothesis was that the compressibility of novels by a single author should tend to be stable over time, consistent with the finding of numerous digital humanists that works by individual authors display a distinctive career-long "signal" of largely unconscious patterns of usage, especially in rates of use of "little words."[40]

Figure 6 seems to show that Trollope's change to more rapid textual production in 1860 did coincide with a significant drop in the information density of his fiction (overall tau -0.338, 2-sided pvalue = 0.0008437). The early novels written over longer periods are so much *less* compressible than the later ones, in fact, that they occupy an entirely different space in the visualization, hardly overlapping the later fiction at all. Quantitative stylistics practitioners might be surprised by the ease with which a test for file compressibility could apparently enable a scholar to identify a text as early or late Trollope, to a high degree of accuracy, simply by zipping it. A compressibility ratio of .361 seems to be the cutoff: before 1860 Trollope was never below this, while for the remaining 35 novels of his career he rose above it only for *The Way We Live Now* and *Harry Heathcote of Gangoil* (both 1873), the science fiction experiment *The Fixed Period* (1881), and *The Landleaguers* (1882). As with the newspaper reporters who created the Hansard transcripts, stronger pressures (in this case self-imposed) to produce more text in the same amount of time seems to have covaried with a significant decrease in the information density of the text produced.

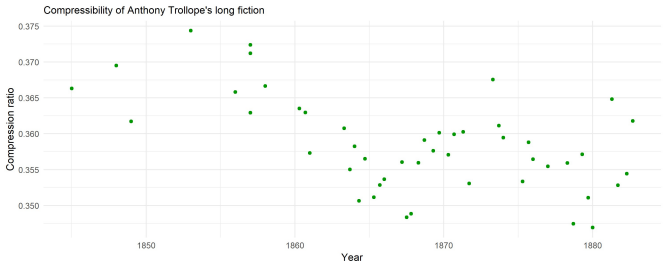

Figure 6. Compressibility of plain text files of Anthony Trollope's 47 novels by date. Source of files: Project Gutenberg.

Center. Grateful thanks are due to Peter Organisiak, Ryan Dubnicek, Eleanor Dickson Koehl, and Nandana Nallapu.

[40]This "little words" analysis was pioneered by J. F. Burrows, *Computation into Criticism: A Study in Jane Austen's Novels and an Experiment in Method* (Oxford: Oxford University Press, 1987). For modern discussion and application of stylistic analysis using modern statistical software see Matthew Jockers, *Macroanalysis: Digital Methods and Literary History* (Chicago: University of Illinois Press, 2013).

At the same time, however, the Stanford hypothesis that the *canonicity* of a work of fiction is the variable most directly related to its redundancy is not well supported by the Trollope results. Popular early novels such as *The Warden* (1855) do score high for information density, but so do the almost unread early works that preceded it, such as *The Kellys and the O'Kellys* (1848). The only factor with a consistent relationship to information density is the period in Trollope's career in which a work was written, and by implication the speed at which it was written.

To check this result against a second nineteenth century novelist with a large *oeuvre*, the same procedure was used on Sir Walter Scott, whose 26 full-length novels are all available from Project Gutenberg and were treated the same way. The short stories and nonfiction were excluded, as in Trollope's case, but Scott's long fictions also included the narrative poems *Lay of the Last Minstrel* (1805), *Marmion* (1808), *Lady of the Lake* (1810), and *Rokeby* (1813), which Scott's biographers agree were written with more care and over much longer periods than his novels. In fact, as numerous primary and secondary sources attest, Scott turned to novel writing mid-career with the deliberate intention of working in a genre he could produce more quickly than long poetry. The poems were unavailable from Project Gutenberg, so—at some risk of introducing a confounding variable—they were downloaded as plain e-texts from the University of Adelaide's eBooks@Adelaide, cleaned of extraneous code, and added to the set of Scott files as plain text UTF-8 files to match the file format of the fiction.[41]

Running the same compression routine on the Scott files produced the data shown in Figure 7, which seems to show that Scott's career as well as Trollope's had a mathematically detectable information history, and of just the sort we might expect. The long poems issued at three- to four-year intervals between 1805 and 1813 are compressible only to levels of .40 or more. The main series of Waverley novels, more quickly written and published beginning in 1814, are much more compressible, inhabiting the .38 to .39 range of the visualization. The final novels written after Scott had suffered several strokes, *Count Robert of Paris* and *Castle Dangerous*—generally considered his least valuable and successful—drop below .38, indicating (though with only two data points) that his condition empirically altered Scott's writing in ways that track with reader experience. Any other conclusions about the career as a whole are harder to draw. A small statistically significant decline in information density over Scott's career appears even if only novels are measured and the poetry excluded (tau = -0.286, 2-sided pvalue = 0.042578), but if the data are truncated to exclude the post-stroke novels the statistical significance of any gradual density decline over the Waverley set disappears (tau = -.159, 2-sided pvalue = 0.28616).

[41] eBooks@Adelaide, University of Adelaide Library, https://ebooks.adelaide.edu.au

These results make it more difficult to rule out a relationship between canonicity and information density in Scott's case than in Trollope's. Early well-received fictions such as *Waverley*, *Guy Mannering,* and *The Antiquary* do seem relatively less compressible than the other prose fictions, but some of Scott's most highly-regarded mid-career fictions, including *Heart of Midlothian*, also score lower in compressibility. Scott may simply have been a fairly consistent fictional coder during the main part of his career, or this experiment may point to limitations to the sensitivity of the compressibility ratio test. It certainly underscores the importance of using it on larger textual oeuvres.

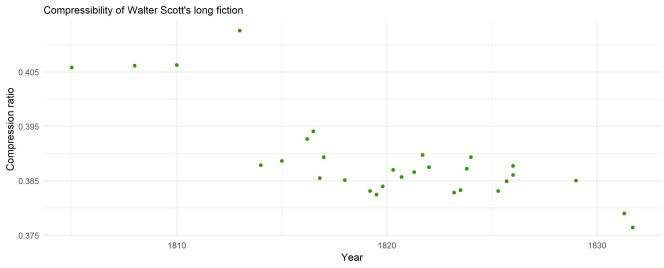

Figure 7. Compressibility of plain text of four long narrative poems and 26 novels by Sir Walter Scott, by original publication date.

Discussion: An information mathematics of fiction?

If multiple experiments using different corpora suggest that national histories of the novel, and even the careers of individual novelists, may reflect information-theoretical pressures in a statistically significant way, literary historians may have to take seriously the hypothesis that mathematical information density has some relationship to literary forms and literary history. It may be necessary to assess the implications of this idea, and to plan further investigations to confirm or disconfirm these apparent effects.

As has been noted, literary scholars since the 1960s have been disinclined to believe–for what appeared excellent reasons—that the information readers recognize in fiction might be the same kind of phenomenon, or have any of the same properties, as the electronic or digital "information" streamed over telegraph, telephone, or fiber optic cables, or stored in computer memory.[42] To most

[42] The fascinating story of how and why philosophers and humanists in the 1940s and 1950s briefly

scholars such a parallel has seemed too unlikely even to need refuting, although in the early excitement generated in the mathematical and scientific communities of the 1950s and 1960s by the engineering success of Information Theory many took time to refute it anyway.[43] Some have continued to do so, as when John Gunders wrote in 2002 that "the redundancy that Shannon discusses has very little to do with the surprise with which a text may confront readers or their ability to anticipate and select the appropriate response."[44] Over the same period a small group of humanists and critics, more strangely, have invoked Shannon in service of critical theory, attracted by the quasi-paradox that for engineers "noise" is the opposite of signal, but also the most mathematically information-dense kind of transmission, since white noise is completely random and unpatterned. This group includes Michel Serres, Jacques Attali, and Henri Atlan in France, Freidrich Kittler in Germany, and William Paulson and Philipp Schweigerson in the U.S. None of these scholars actually adopt Shannon's mathematical quantification of noise along with his definition of it, however, and some go out of their way to reject it.[45] Paulson writes that, "[W]e are not dealing with the direct application of a working scientific theory to a new empirical domain: the literary text cannot simply be placed in the position of electronic signals and analyzed quantitatively."[46] David Letzler goes further, classing the whole idea of applying information mathematics to literature with the poststructuralist misuses of science that inspired Alan Sokal's *Social Text* hoax.[47]

As recently as the early 2000s, this opinion that information theory was for computer engineers alone was widely shared in other disciplines as well. To that date even most linguistics scholars—in many cases approaching the idea of linguistic information through Chomskyan assumptions about generative grammar—left the mathematical model of information untouched, even though actual engineers, including computer scientists such as Peter Norvig, who was to become

engaged and then disengaged Shannon's Information Theory is well told in Bernard Dionysius Geoghegan, "From Information Theory to French Theory: Jakobson, Levi-Strauss, and the Cybernetic Apparatus," *Critical Inquiry* 38.1 (2011): 96-126.

[43] See, for one particularly convinced anti-Information Theory example, Carl H. Weaver and Garry L. Weaver, "Information Theory and the Measurement of Meaning," *Speech Monographs* 32 (1965): 435-447.

[44] John Gunders, "Signal or Noise? Information Theory and the Novel," *Double Dialogues* 3 (Summer 2002).

[45] For good explications of this position see William R. Paulson, *The Noise of Culture: Literary Texts in a World of Information* (Ithaca and London: Cornell University Press, 1988), and Philipp Schweighauser's entry on "Information Theory" in Bruce Clark and Manuela Rossini, eds, *The Routledge Companion to Literature and Science* (New York: Routledge, 2012), 145-156.

[46] Paulson, 65.

[47] David Letzler, "Crossed-Up Disciplinarity: What Norbert Wiener, Thomas Pynchon, and William Gaddis Got Wrong About Entropy and Literature," *Contemporary Literature* 56.1 (2015), 52.

a director of research at Google, were already successfully applying statistical modeling based on information theory to linguistic processing in information retrieval, and beginning to achieve the real-world results already discussed.[48]

In the past ten years, however, researchers' assumptions across a range of related disciplines about the value of information theory for investigating human language use has begun to alter. In experimental studies involving both corpus analysis and laboratory work with human subjects, researchers at MIT's Department of Brain and Cognitive Sciences, the University of Barcelona's Complexity and Cognitive Linguistics Lab, and other institutions have independently produced evidence that information theory may usefully model some aspects of how minds produce and exchange language. In 2011 psychologist Steven T. Piantadosi was able to show with corpus research that statistical information content accurately predicts word lengths in natural language, and in a separate study the next year that information-theoretical ambiguity in natural language improves its information efficiency.[49] In 2013 Kyle Mahowald et al. showed that speakers demonstrate information efficiency in making word choices.[50] Yang Xu and David Reitter of Pennsylvania State University recently proposed a formal "information-theoretic view of dialogue" after demonstrating that dialogue partners mutually adapt the information density of their speech as their discourse develops.[51] Laboratory studies of reading have also offered evidence that information mechanics may partly structure the experience of literary narrative. Rolf Zwann found that subjects given a passage of narrative and told that it is fiction read significantly more slowly than subjects told the same passage is a piece of journalism, suggesting that readers expect—have learned by experience to expect?—greater information density in fiction than in nonfiction, and slow down to better process it.[52]

Even other recent work in the literary digital humanities, revisited from this perspective, may support the idea that humanist information has more in common with statistical information than we have believed. Franco Moretti's 2007 *Critical Inquiry* article "Style, Inc." showed quantitatively that titles of novels published

[48] Norvig's readable essay evaluating Chomsky's continuing claims (as recently as 2011) that human language cannot usefully be statistically modeled is, "On Chomsky and the Two Cultures of Statistical Learning," reproduced at Norvig's website, http://norvig.com/chomsky.html

[49] Steven T. Piantadosi, Harry Tily, and Edward Gibson, "Word lengths are optimized for efficient communication," *PNAS* 108.9 (2011); and Steven T. Piantadosi, Harry Tily, and Edward Gibson, "The Communicative Function of Ambiguity in Language," *Cognition* 122 (2012): 280-291.

[50] Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson, "Info/information theory: Speakers choose shorter words in predictive contexts," *Cognition* 126 (2013): 313-318.

[51] Yang Xu and David Reiter, "Information Density Converges in Dialogue: Towards an Information-Theoretic Model," *Cognition* 170 (2018), 147-163.

[52] Rolf A. Zwaan, "Effect of Genre Expectations on Text Comprehension," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (1994): 920-933.

in Britain in the eighteenth and early nineteenth centuries became shorter over time, but also with close reading that the new short title structures invented during this process conveyed greater significance in fewer words.[53] Andrew Piper, in "Fictionality," makes sophisticated use of the LWIC tool to discover a statistically significant rise over the nineteenth century of language associated with *perception* and with *doubt* in English-language fiction.[54] Ryan Heuser and Long Le-Khac used topic modeling to discover a century-long trend in long fiction away from generalization and abstraction toward specific and concrete language. In all these cases the authors ascribe causation to non-information-theoretical factors: Moretti attributed the change he observed to the workings of natural-selection-style evolution in a market, while Piper, Heuser, and Le-Khac proposed historical and cultural trends. Their findings are consistent with gradually improving information performance as well, however, because the language forms they find increasing over time are those that better concentrate the qualities information theory measures.

These research results suggest that the belief of literary scholars that information in texts belongs in a different epistemological category than information in machines—that it obeys different laws and is unreachable by empirical inquiry—may soon be unsustainable. Even our professional experiences as scholarly writers and teachers of writing may have been showing us a different set of forces at work. Like Victorian parliamentary reporters, many of us have had experience that when we write long texts to short deadlines the results are likely to be wordy and redundant. Some have probably observed that in enterprises where language must be used under constraints, or where accuracy is especially important, practitioners expand or contract their discourse to match their system's needs, developing techniques to balance production time, language volume, and information content. The ten-codes used by police, fire, and rescue services, the abbreviations and initialisms in military orders and hospital emergency rooms, the clarity of flight traffic instructions, the clipped codes exchanged by restaurant wait staff and line cooks, are all linguistic responses to systemically constrained information exchange. Long-form as well as short-form professional genres are altered by such pressures. Generations of journalism students have been taught a (somewhat exaggerated) story of how news organizations invented the "inverted pyramid" to concentrate important information at the beginning of a document when telegraph service became unreliable during the American Civil War. They are less often given the example of Victorian freelance "penny-a-line" reporters,

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

[54] Andrew Piper, "Fictionality," *CA: Journal of Cultural Analytics* 20 Dec. 2016. Article DOI: 10.22148/16.011, Dataverse DOI: 10.7910/DVN/5WKTZV

whose reimbursement was determined by the number of lines of their submissions the newspaper printed, and who responded to those incentives by coding ever-smaller quantities of newsworthy information into ever-larger quantities of verbiage.

If an information-theoretical mechanism for both Algee-Hewitt et al.'s result and the compression ratio results described here are confirmed, literary historians might need to find new kinds of stories to tell about the historical forces that helped shape the development of narrative fiction. To explain the increasing information density of English and French novels over the nineteenth century, for example, we might consider a hypothesis that from the mid-eighteenth century onward novelists faced a technical challenge inherent in the ideal of a prose fiction at once realistic and artistic. Poetry concentrates signification using specialized and aesthetically optimized language codes, but realist fiction has to use—or seem to use—the same narrative and speech genres as contemporary social interaction. The challenge this poses for artists is that social speech genres are mostly low-density, marked by high redundancy and the use of premade linguistic formulae (Douglas Biber and Susan Conrad's "lexical bundles") that render spoken language easy to produce in real time.[55] Novelists trying to construct high-information-density artistic texts out of low-information-density discourse genres may never have been in the enviable position imagined by critics such as Pavel, free to confabulate without constraint, and more in the one imagined by Vladimir Nabokov, required to create art by assembling "dazzling combinations of drab parts."[56] Raymond Chandler described the difficulty of writing dialogue in Hollywood at the mid-20th century: "the challenge of screenwriting is to say much in little and then take half of that little out and still preserve an effect of leisure and natural movement." The only solution to such problems, Chandler wrote, was "experiment and elimination."[57] The historical rise in the average information density of French and English fiction seen in Figures 1 and 3 may be a trace left by novelists addressing the complex information problem of artistic realist fiction by experiment and elimination also, gradually accumulating an expanded repertoire of more effective coding tools by successive variation in their discourse practices, and by copying the successful variations of peers.

An emerging information history of fiction might also help us make connections among a range of literary phenomena and critical theories we currently treat as separate. In communications engineering, reducing statistical redundancy is

[55] See for example D. Biber, S. Conrad, and V. Cortes, "*If you look at … :* Lexical bundles in university teaching and textbooks," *Applied Linguistics* 25 (2004): 371-405.

[56] Vladimir Nabokov, *Nicolai Gogol* (New York: New Directions, 1961), 56.

[57] Raymond Chandler, *Raymond Chandler Speaking,* ed. Dorothy Gardiner and Kathrine Sorley Walker (Berkeley and Los Angeles: University of California Press, 1997), 119.

only one way to increase the information performance of a channel. Coders can also increase the local uncertainty of the individual components of the message that remain; in the case of fiction, this might mean making uncertainty itself a central topical interest in fiction, as Piper seems to have found, or reducing the average generality of language used in favor of more concrete detail, as Heuser and Le-Khac showed. A still more effective solution to information bottlenecks in many technical communication systems is to "multiplex" the channel by finding ways to send more than one message simultaneously. Edison's "Quadruplex" telegraph used phase modulation to send one message and amplitude modulation to send another across the same wire at the same time; novelistic equivalents may include the stylistic inventions literary scholars know as Free Indirect Discourse, Mikhail Bakhtin's heteroglossia and polyvocality, and Lisa Zunshine's application to fiction of Theory of Mind,[58] which may not be mutually unrelated phenomena. Novelists under information coding pressure may value and seek ways to make single lines of text do double duty, and "multiplexing" novelistic discourse by simultaneously conveying a narrator's perspective and a character's, one character's thoughts and another's, or a character's own perspective together with its ironization, do this double duty effectively.

Humanists are within their disciplinary rights to require rigorous evidence before accepting claims that empirical measures can apply to human expression, or that mathematical constraints might have helped shape imaginative language. Even for the most cherished and traditional purposes of literary history and criticism, however, techniques that allow us to trace an information history for fiction may be valuable. If literary scholars are to appreciate as fully as possible the craft, cognitive powers, and artistry required by great fiction, we cannot afford to ignore any of the constraints under which—and against which—novelists have had to innovate and succeed.

[58] M.M. Bakhtin, *The Dialogic Imagination: Four Essays* (Austin: University of Texas Press, 1981); Lisa Zunshine, *Why We Read Fiction: Theory of Mind and the Novel* (Columbus: Ohio State University Press, 2006).