11

MARK ALGEE-HEWITT, ERIK FREDNER ,
AND HANNAH WALSER

The Novel as Data

A radically expanded notion of "reading" – which can now be close, distant, or take place at the surface – has emerged in recent scholarship, one that demands a reconsideration of the ways in which a reader gains information through an encounter with a novel.[1] In response to these evolving practices of textual interpretation, our conceptions of the novel itself have begun to change, shifting dialectically from the novel as a medium of narratives and ideas to a plurality of textual and contextual data that can be accessed, transformed, and even "read" through computational methods. Though we have always known that novels are more than the sum of such parts, even a single novel considered in this way can yield a wealth of information about word frequency, syntax, and structure – the building blocks of literary fiction.

Interpretation usually happens at the critic's scale of reading: maybe a few thousand novels in a lifetime, from which a handful of texts are considered at any given moment to develop an idea or advance an argument. But when we ask questions about genre – "What was *the* novel?" instead of "What were *these* novels?" – a turn toward data can offer historical, cultural, and generic insights without direct antecedents in past models of literary study precisely because the information offered by these insights does not operate on a human scale, though they are drawn from and informed by it. Thinking of the novel as data is by no means an attempt to circumvent reading. Rather, like other critical frameworks, it is better understood as a new way of seeing what has long been before us.

In this chapter, we consider these multifaceted transformations as we move from novel to data and, in the process, back to the novel. Although relatively new to the work of literary criticism, research in this vein has already yielded key insights in, for example, the interleaving of the conversional narrative with its formal structures or the loci of geographic attention in American fiction of the nineteenth century.[2] In these and other studies that understand the novel as data, what kinds of new knowledge become possible,

and what kinds of knowledge are lost? Can the phenomenological and interpretative aspects of a reader's encounter with a novel *become* data? And what, if anything, does this teach us about the novel itself? As we shall show, the intellectual and technological transformation of the novel into data is neither as radical as it might seem at first, nor is it a move unanticipated by the novel itself.

Data and Information

What does it mean to study the novel as data? Data, first of all, are closely connected to information. Whether read for pleasure or studied as objects of cultural interest, novels are a communicative medium and, as such, are carriers of information themselves. The information typically gained by a reader includes concepts familiar to literary criticism: plot, theme, characterization, and conflict. A close reading looks beneath these surface features to detect hidden elements, whether they are structural and aesthetic aspects that speak to the function of the text or information relating to the sociocultural conditions under which the novel was written or published. Both of these readings, however, fundamentally locate the information in the novel within higher-order concepts and ideas. In either critical epistemology, the novel ceases to be an ordered collection of lexemes or words and instead becomes an assemblage of themes, histories, and meanings, detached from the language through which they are communicated to the reader. It may have more or less in common with other novels of the same genre, time period, nationality, or author, but as a textual object, it remains irreducible in and of itself. As a part of a literary critical argument, then, a novel is anecdotally important. It can function as a piece of evidence if one of its higher-order concepts or facets confirms a pattern also found in related novels, but it needs to be explored as a unique object and, crucially, as a whole.

This strategy of reading is predicated on a set of mental operations that dictate how we extract information from a text in the first place: in seeking to understand a particular concept, plot, narrative, or theme, for instance, we often elide the specific words (and the specific order of the words) through which that concept is communicated. Words are, to put it another way, the means by which reader comprehension happens but not that which is comprehended. A reader studying *David Copperfield* can speak of the theme of class, or education, but he or she cannot precisely state (or be expected to know) how many times "the" is used to communicate these concepts. But in these words lie the patterns that construct the novel, that form the characters and communicate the ideas (and the ideology) that the author codes into the

text. For example, in *David Copperfield*, the frequency of a word such as "the" reveals that it, unlike most nineteenth-century novels, deemphasizes descriptive noun phrases at the beginning of the text in favor of the personal narrative (marked by "I").[3] Even such a "content-free" word as "the" reveals aesthetic facets of the novel, and its relationship to the other texts of its period, that are inaccessible to the traditional forms of literary interpretation. Reading the novel as data transforms both the text itself and our relationship to it.[4] It is reduced to a collection of words, in a particular order, each of which can be treated independently from the novel as a whole. Concepts, topics, and themes become assemblages of these words that emerge through the abstraction of our reading experience. These higher-order features no longer function themselves as concrete pieces of information and objects of interpretation; instead, they are revealed to be the end result of interpretative processes themselves.

Reading novels as data, then, involves not just a change in the novel as an object but also, and more importantly, a change in our strategies of reading. In 1961, in *The Future of Data Analysis*, statistician John W. Tukey described the differentiation between statistics, a pure mathematical science, and data analysis, an applied science that seeks to determine meaning from a given data set. The transition from reading to data analysis is less severe but still involves many of the same processes that Tukey describes: "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."[5] As it is applied to novels, then, data analysis becomes a new form of reading, a new way to interpret textual information.

Implied in this transformation from novel to data is quantification. Word counts of "tokens" (the individual words in the text) can be aggregated into counts of "types" (the unique words), turning raw words into a numerical representation of the novel. In quantification, what we lose in the immediate interpretability of the results of the analysis we gain in the ability to make use of statistical and computational operations to parse text at scales, both large and small, that are unavailable to readers. This is the metamorphosis that lies at the heart of Claude Shannon's information theory: the transformation of words into bits, into quanta of communication whose order, repetition, and frequencies not only can represent the text down to the smallest detail but also can be added, multiplied, or statistically modeled.[6] As early as the mid–twentieth century, the representation of text as quantity merged with literary criticism, inspiring Roman Jakobson's theory of communication functions. Jakobson's work sought to systematize text (including the novel)

in terms of its ability to transmit messages, adding the "poetic function" to explain the role of literature in textual data analysis.[7] If, in quantifying novels, we lose the ability to interpret the quantified text as readers, we gain the ability, as analysts of the resulting data, to study micro- and macroscopic linguistic structures at scales ranging from an individual sentence to a corpus of thousands, or millions, of texts.[8] The resulting evidence, though less easily subsumable into a narrative argument, has the potential to turn theoretical concepts into replicable, data-driven operations.

Why Think about the Novel through Data?

Reading the novel as data involves not just a transformation of substance (words into numbers) but one of scale as well. As the words of the text become numbers (raw and normalized frequencies, probabilities, and sequences), a comparison between novels becomes akin to a mathematical operation: a comparison of quantities, frequencies, percentages, and likelihoods. Whereas a close reading of an individual novel resolves the information of the text into a series of discrete features (themes, concepts, and narratives) that may be more or less qualitatively similar to other novels that the scholar has read, the quantitative comparisons enabled by the transformations of the novel into data are subject to a different set of operations. Similarities and differences between texts become a function of shared or differing word probabilities: whether these probabilities represent the presence/absence of words or the statistical likelihood of them appearing in given sequences, their transformative power remains the same.[9] Whereas meaning, in a close reading, lies in the reader's ability to apply larger cultural or aesthetic phenomena familiar to him or her to a text from the outside in a quantitative analysis, meaning becomes a set of relationships within and between texts, an assemblage of presences and absences that together offer a different way to reconstruct the sociocultural, aesthetic, or critical phenomena that shape the novel. It is this large, macroscopic scale that is most often cited as a primary advantage of such a quantitative analysis.[10] Rather than arguments based on, at most, tens of books, we can now craft arguments based on comparisons that are orders of magnitude larger. And yet, a quantitative analysis of the novel depends primarily on both of the scales at which data operate: the very small and the very large, the microscopic and the macroscopic. As these analyses count individual function words within thousands of texts, they are able to reveal new kinds of resemblances and differences that are only apparent within these magnitudes of scalar reading.

Yet reading at these scales depends on a process of abstraction. A quantitative analysis, by its nature, requires a loss of information to attain

interpretability. Many analyses that compare individual word frequencies across thousands of texts lose the ability to keep track of where these words occur and in what order in any individual text. They retain minute knowledge of frequencies while sacrificing any knowledge of sequence. Studies that elect to focus on relationships of sequence find that this information is incompatible with frequencies. And in both cases, to collapse comparisons involving millions of words into an interpretable measure, analyses must collapse the many dimensions of similarity and difference into single measurements of comparison.[11] To operationalize a novel, then, is to measure it, and to measure it is to reduce it to discrete points of meaning.[12] In any process of abstraction, information is lost. But even close reading studies sacrifice information for clarity. In a close reading, a reader reduces the manifold complexity of the text down to a single point of conceptual understanding, often abstracted to single-word concepts: "time" in *To the Lighthouse* and "identity" in *David Copperfield*. Such an analysis excludes the information in the novel that could be used to compare *David Copperfield* along other axes (social critiques of industrialization or Victorian gendered social codes). The difference is that this information loss occurs at the beginning of the analysis, at the point of reading, rather than at the end of the analysis, at the point of interpretation. A reader focusing on identity will abstract the language of the book to this one concept and then compare the concept across texts (or within a single novel). A quantitative analysis retains much more of the complexity of the text in the comparison but abstracts the dimensions of comparison in order to create an interpretable measurement. Both deal in abstraction; the divergence between the two analyses lies in *when* it occurs.

In the process of abstraction, the quantitative textual analysis gains the ability to operate at extremely large scales, across corpora of texts rather than individual texts. The advantages of this are manifold, particularly in the possibilities it offers for changing our understanding of the literary history of the novel. Scholarship on the novel has always been constrained by the pace of reading. Take, for example, Ian Watt's and Michael McKeon's divergent accounts of the development of the novel through the eighteenth century in *The Rise of the Novel* and *The Origins of the English Novel*, respectively. While the conclusions both draw are vastly different, the texts that they chose to analyze, out of the hundreds of novels published during the timeframe of their arguments, are remarkably similar: *Don Quixote* (1605 and 1615), *Pilgrim's Progress* (1678), *Robinson Crusoe* (1719), *Gulliver's Travels* (1726), *Pamela* (1740), *Clarissa* (1748), *Tom Jones* (1749), and *Tristram Shandy* (1767), among others.[13] The scholarly canon of novels, whether that of the eighteenth, nineteenth, or twentieth century, is, in part, a response to

the inability of the critic (let alone the reader) to account for the vast number of novels published. Faced with the entirety of literary history, the institutional response has been to cull the numbers of novels into what Matthew Arnold in 1869 called the "best which has been thought and said in the world," the literary canon.[14] In this reduction, though, what historical transformations and local effects are lost in the vast field of what Margaret Cohen has called "the great unread"?[15]

In a provocative response, Franco Moretti begins "The Slaughterhouse of Literature" with a list of titles from Columbell's circulating library catalog: "*Arabian Tales, Aylmers, Annaline, Alicia de Lacey, Albigenses, Augustus and Adelina.*"[16] The form of the list itself causes us to wonder what these books were. Many sound like they might well have been novels, but few have been studied. Framing the logic of the list in Darwinian terms, if we imagine the novels that have "survived" (i.e., that we still read) as the fittest iterations of a particular subgenre of the novel, then, through the parallel processes of education and canonization, we come to take these works' distinctive features as salient expressions of the ideas, affects, and other cultural-historical forces that a subgenre of the novel responded to or expressed. But when we look to the vast number of texts that are no longer read or taught or written on, we begin to see contingencies: other histories, other ideas of the novel than the ones we have inherited. Do these texts Moretti lists, known as novels in their time but unmentioned in accounts of the novel such as Watt's or McKeon's, have less purchase on the category because they did not survive the "slaughterhouse"?

Expertise, *literary* expertise, consists, then, of knowing a fraction of a fraction of the field. The aperture through which critics hope to shed light is quite small. Scholars of literature interested in quantitative textual analysis need not (and, for that matter, do not) claim that the recent empirical turn somehow invalidates previous literary scholarship on the basis of mere sample size.[17] The study of literature has always focused on outliers, those privileged positions in Pierre Bourdieu's field, at the expense of almost everything else in it precisely because the works that people read and rediscover contribute more to literary culture than unread and forgotten works do.[18] But one of the great opportunities offered by a quantitative analysis of text lies in putting the small fraction described by Bourdieu in the context of the vast unread discussed by Cohen and Moretti. As early as 1500, when the total number of books and manuscripts in Western Europe hovered around 11 million,[19] the prospect of "reading it all" had already become quixotic. Today, with nearly 3 million new books published annually, merely *finding*

all the novels published in a given year would be hard enough, never mind the matter of actually reading them.[20] When we consider how little of the field we can actually read – and how deliberately unrepresentative what we actually read is – we can get a sense for the scale at which claims made about "the novel" necessarily operate and how transformative a quantitative analysis that could take a much larger account of the field as whole could be.

If these vast corpora offer a new set of possibilities for novelistic information, then they are only made feasible by the minute features at the other end of the scale through which quantitative analysis operates. From a detailed attention to the individual frequencies of individual words, we can compare thousands of texts across hundreds of years. Seen in this way, the differences between quantitative analysis and close reading become transformations not only of scale but also of structure. If we imagine the text to be a fabric, then close reading cuts a swatch and examines its colors, patterns, and texture; a quantitative analysis, by contrast, pulls apart the weave to examine individual threads whose use can be compared across vast amounts of cloth. Close reading focuses on the synthetic effect, whereas reading novels as data allows us to explore the texts' constitutive elements. The former ties the range of potential readings to the theories and concepts that the closer reader brings to the text; the latter offers the potential of new readings based on the unanticipated patterns that emerge from the data. Yet, in the analysis of these patterns, we can assemble new synthetic models of the text whose conclusions work hand-in-hand with the literary practices of close reading.

Indeed, one research outcome for a project that reads the novel as data is a new way of characterizing the exceptional and conventional aspects of a literary novel when put in dialogue with a larger field. Our ideas about what makes a novel such as *David Copperfield* exceptional might be developed on aesthetic, historical, and cultural grounds and then bolstered by readings of Dickens' novel, along with comparative texts, informed by previous theoretical and critical work in the field. An extension of this traditional modality of criticism then asks whether such qualitative claims can be verified, refined, modulated, or upended by means of a quantitative scaling up.

Rendering the novel as data does not merely provide a new approach to old questions, although it can. Rather, transforming the novel into data radically expands the set of texts against which literary critics and historians must test their assumptions and theories about both literary and nonliterary works. Andrew Piper has argued that the case for a quantitative intervention rests, in part, on the inability of the humanities to generalize in a systematic

and intellectually rigorous way an area where quantitative research has excelled.[21] The generalizations produced by data-analytic approaches are unusual for literary studies because they are capable of being wrong: not in the sense that their claims can be contested, since this applies to any mode of literary criticism, but rather in the sense that the findings can be reproduced or, if not, falsified.[22] Here lies a critical difference between the analysis of the novel as data and the practices of close reading. In the quantitative analysis, there are aspects to the researcher's hypothesis *that are unknown* at the outset of the study and that may prove the initial hypothesis false. In the close reading, the categories of true and false, in the scientific sense, are meaningless because all information present in the analysis is *already known* to the researcher at the outset. Interpretation from such an analysis can be "reproduced" by persuasion, but what persuades one reader may not be enough for another. Theories of the novel pieced together from close readings are abandoned as they fall out of fashion or are superseded by new cultural norms, but rarely are they proven "false." And yet both skeptics and practitioners of data-oriented literary criticism constantly ask the same questions of these methods. Can quantitative aspects of a text really have purchase on the qualitative experience of reading or broader notions of literary history? If they can, which do? And if we are looking to the right metrics, how do we determine their significance?

Working with the Novel as Data

"Data" does not only connote computation. The spreadsheet program and the database, for instance, have become the object metaphors that we turn to when we think about data today, but these have analogue origins in such things as bookkeeper ledgers and shipping manifests. Indeed, in thinking about the novel as data, we could, in fact, look back at least as far as 1907, when Mary Williams published her *Dickens Concordance*, whose title page describes the book as "a compendium of names and characters and principal places mentioned in all the works of Charles Dickens."[23]

Williams restructured Dickens' corpus around three simple questions. Which character was introduced or place visited? In which book? And when? The *Dickens Concordance* is thus partly indexical, partly cartographic, and partly temporal, privileging each of these categories over the many other aspects encoded into Dickens' works (see Table 11.1). Because of this structure, Williams' *Dickens Concordance* is not only an early example of the novel reimagined as data but also a way of reencoding prose as a means of data storage and retrieval.

TABLE 11.1: Partial Reproduction of Table from *Dickens Concordance*

Character or place	Book	Chapter
Cook's Court, Cursitor Street	*Bleak House*	10
Coper Augustus – "Boz"	*Dancing Academy*	–
Copperfield, Mr. and Mrs.	*David Copperfield*	1
Copperfield, David	*David Copperfield*	1
Coppernoze, Mr.	*Mudfog Papers, 2nd Meeting*	23
Corney, Mrs.	*Oliver Twist*	23

Source: Mary Williams, *The Dickens Concordance, Being a Compendium of Names and Characters and Principal Places Mentioned in All the Works of Charles Dickens*. Francis Griffiths, 1907, p. 87, available at https://archive.org/details/dickensconcordan00willrich.

So can Williams' way of restructuring Dickens' work teach us anything about the novels themselves? We are surely unsurprised to find that we meet David Copperfield in the first chapter of *David Copperfield*. But for Dickens' many bit players and functionaries, we might readily assemble a character network out of such introductions, a timeline of entrances tied to the work or even Dickens' larger oeuvre, and start to more fully describe what Alex Woloch has termed the "character space and character systems" of the individual novels in Dickens' fictional universes.[24] For instance, we might learn from patterns readily visible in Williams, but more obfuscated in Dickens' narrative, whether the timing of a character's introduction betrays their (in-)significance to the novel's plot. We can address questions such as these more readily with Williams' data than we could with Dickens' novels despite the fact that both contain sufficient information. It is Williams' transformation of this information, the way that she makes it available as data, that makes it available for such a reading. As we see through Williams, the novel considered as data can help to organize observations that elevate the primacy of a set of questions.

Yet one could not read Williams and feel that one had read Dickens. This is true in spite of the fact that from the perspective of Williams' questions, both contain the same information. The incommensurability of Williams' work with Dickens' – and, by extension, the transformation from novel into data – also encodes loss. There is no single data representation adequate to address a broad range of literary critical inquiry; instead, the way that we represent data is contingent on the questions that we wish to ask of them. The very act of representing data (through a graph, a spreadsheet, or a database) already encodes a research hypothesis into the representation.

Computation and Criticism

The *Dickens Concordance* is certainly an early example of a critical work that imagines the novel as data, but it is limited in its ability to speak about the genre of the novel more broadly. At approximately 3.8 million words, Dickens' body of work is massive by any standard, and Williams had the drive to ask, and answer, quantitative questions about character and place across Dickens' entire corpus. But as corpora begin to exceed 100 million words, such a strategy becomes impractical or even impossible.[25] We need computational processes that can transform elements of those texts into observations of literary phenomena.

A first-order problem for studying novels as data is the form of the novel itself. Paper, glue, and ink do not readily compute. If we want to work faster than Williams did, then, we need digital copies of the texts. This is the first transformation the novel undergoes to become data, and like other means of producing data we have looked to so far, the process necessarily entails a loss of some aspects of the work. For the novel, the loss is primarily paratextual: printing and formatting choices, material histories, and other means of mediation change as books are converted into digital files.[26] After this process, then, what remains? The novel's words, punctuation, paragraph breaks, and their sequencing.

Once we have these files, though, how can the computer begin to "read" them? Any reference here or elsewhere to computational reading is a metaphor of convenience. To the extent that computers can "read," they frequently do so by moving through the text one word at a time, often tracking one or more elements, patterns, and so on, but it remains different

TABLE 11.2: A Portion of the Document-Term Matrix for *David Copperfield*.

Word	Frequency
aback	1
abandon	6
abandoned	13
Abase	1
...	
daughterlike	1
daughters	9
David	65
Davy	165

in that the computer does not parse for semantic meaning. And yet the digital representation of the quantities of words in the text offers opportunities for analysis that are otherwise unavailable to a reading that focuses on meaning. One such representation is the "document-term matrix," a table that contains each of the types extant in a given document or documents along with the number of tokens (see Table 11.2).

Here we find ourselves with quantities rather than qualities and a "text" reorganized in such a way that it is scarcely recognizable. But this defamiliarization provides new opportunities for analysis: from the distribution of words and characters in a single text to a new analysis of the relationship between lexical patterns in novels across the wider nineteenth century.

Word Distributions

The frequency of words in a novel such as *David Copperfield* belies our expectations as readers. In order to be noticed by a reader, a word must be frequent enough that it is repeated within the text but unique enough to be recognized as such. Words such as "the," "and," or "of," for instance, are used too often to accrue any special meaning within the novel, but infrequent words such as "nestle," "plotted," and "outstanding" – all used only once in the novel – are so rare that they are lost among the 330,000 other words in the text. Unless it is used in a particularly striking way, however, a low-frequency word vanishes into the linguistic background. But if reading comprehension takes place at a middle scale, between the ultrahigh- and ultralow-frequency words that occupy the ends of the spectrum, what does the distribution of frequencies in a text, or in a corpus, actually look like?

Figure 11.1 shows the full distribution of the frequencies of every word in *David Copperfield*, from most to least frequent. Each one is given a bar that rises to the height of its raw frequency, from 13,436 at the left of the graph ("the") to 1 at the far right of the graph ("zigzag"). The vast majority of the plot appears to be empty, simply because the frequency falls off so quickly that only the words on the very left side of the graph appear to have bars that are visible on the axes. This is because the distribution of frequencies within any text document (or collection of text documents) follows "Zipf's law," an exponentially decreasing distribution in which each rank, moving from highest to lowest, decreases in frequency by half.[27] The first few frequencies demonstrate the pattern: "the," 13,436; "I," 11,961; "and," 11,642;

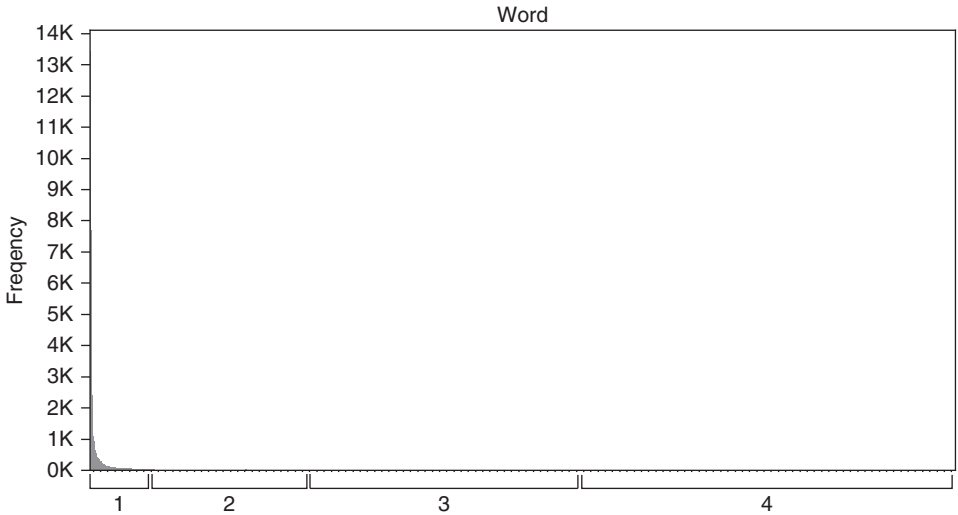

Figure 11.1: Distribution of word frequencies in *David Copperfield* ordered from high to low. The four zones below the *x*-axis represent specific frequency ranges.

"to," 10,242; "of," 8,579. The four numbered zones below the *x*-axis divide the chart into meaningful segments based on the frequencies of words they demarcate. Zone 1 captures the 1,000 most frequent words in the corpus, which have frequencies from 13,436 to 28. These are primarily composed of function words that are predominantly invisible to the reader even though they structure the language and grammar of the text. Zone 2 shows the next group of words: with frequencies from 27 to 6; they are repeated enough to be noticed but much less so than the function words that populate zone 1. Zone 3 contains words with frequencies from 5 to 2; these are typically words that have some value in repetition but do not play a significant part in the text. It is worth noting that many minor character names appear here. Finally, zone 4 contains words that are only used once in the text. It is the largest single group of frequencies, as it typically is for any text or corpus.

In their article "Mind-Modelling with Corpus Stylistics in *David Copperfield*," Peter Stockwell and Michaela Mahlberg draw on cognitive poetic models, coupled with a close reading of the text, to explore the mental structures of the characters in the novel. In the course of their investigation, they isolate lexical choices made by Dickens to articulate the cognitive space of particular characters, focusing, for example, on words "from the domains of perception, thought and belief: *think,*

looked, doubtfully . . . hope, anxiously, concocted."[28] The authors argue that these terms define a particular set of mental processes in their repetition by and around characters. If we place them on the graph of word frequencies in Figure 11.1, they would fall squarely into zones 2 and 3: high frequencies relative to the words on the right, which only occur once, but still orders of magnitude less frequent than the terms of the left. Yet it is here that the authors, like many before them, place their attention. This is the space of close reading on the word distribution: words that may contain a weak signal in and of themselves but that strike the perfect balance between remarkableness and infrequent repetition that signals attention to the reader. An argument such as Stockwell and Mahlberg's explores the use of these words as data in and of themselves but is only able to provide these terms through a close reading of the cognitive models of the text.[29] But what of the work of the remaining four-fifths of the graph? By isolating these words in service of their argument, they not only ignore the thousands of other words from the same frequency rank but also the tens of thousands of other words at different frequency ranks. By quantifying the words and making them available to computation, we can bring these ranks into visibility.

The words in Figure 11.2, which shows only words with frequencies over 500 (the top ninety-five words in the corpus), contain a number of

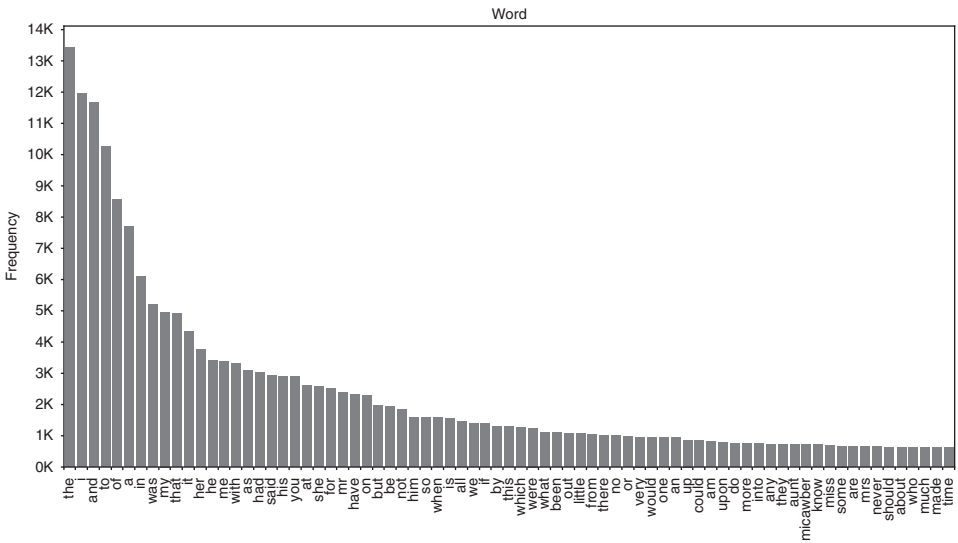

Figure 11.2: The top 95 most frequent words in *David Copperfield*, ordered by frequency.

powerful signals. Strongest of all is authorship: the frequencies of these terms, aggregated over the text (and, ideally, many texts by Dickens), reveal a pattern that is unique to Dickens' writing style.[30] They also serve to differentiate the language of a text written in Victorian England from texts written in different times and places (a late-twentieth-century American novel, for example). Between authorship and period signals, most of the raw information in a text is contained within this first frequency rank. Yet it receives little to no attention in close reading studies: the patterns of high-frequency words that identify authorship, or period, are simply too subtle, or take place over too much text, to be recognized by a reader (at least consciously). By quantifying the novel's lexicon, by turning the text into data, we can move from close readings into larger pattern analyses that reveal relationships between this text and others. It is through the silent but extremely active register of high-frequency words that texts reveal their *non*uniqueness – their participation in systems of genre, authorship, period, and nationality that surpass both authorial intention and readerly awareness.[31]

One way that quantitative literary critics have adjusted to the fact that high-frequency words tend to mark style more closely than semantics involves removing what are known as "stopwords." The stopword category includes many of those function words that are more indicative of authorial style than bearers of meaning ("the," "and," "of") as well as prepositions ("at," "into") and contractions ("I'm," "you'd," "won't"). To readers, these words, the connective tissues of prose, are almost invisible.[32] More important, many words from zones 1 and 2 are of critical interest, but their relationships are harder to see with the intervening stopwords. By counting only nonstopword frequencies, however, we can get a much better sense of a text's recurring themes, interests, and investments.

In *David Copperfield*, the most frequent nonstopwords are "said," "mr.," "little," "one," "upon," "know," "micawber," "aunt," "now," "miss," and "peggotty." Like the prevalence of "I" in Figure 11.2, we can understand the dominance of "said" as it relates to Dickens' novel writing. Character names ("Micawber" and "Peggotty") and forms of address ("Mr.," "Aunt," and "Miss") unsurprisingly also appear quite frequently. But even this short list points to some words bearing more conceptual and stylistic weight: Dickens' interest in representing the young and the downtrodden (both rendered sympathetically with the occasional diminutive "little," surprisingly high in third place), David Copperfield's development of social and philosophical epistemologies ("know" and its pair "not

know"), and the roles of both coincidence and memory in Dickens (what happened in the deictic "now," and what from then has come to be known "now").

This movement between scales, from close to distant, happens constantly in the preceding microreadings of that word list. We necessarily refer to both Dickens' broader intellectual project and practice as an author and the specific contexts in which the words that appear most frequently in *David Copperfield* are used. To make an implicit point explicit, we enter into a familiar interpretative mode of literary criticism but with a different object than the paragraph, chapter, or work in looking at these frequencies.[33] Interpretation can be guided – and occasionally overruled – by information derived from the quantitative analysis of a text. And, more important, not every computational result is accessible to *literary* interpretation. That is, even though the results of algorithmic inquiry sometimes look like text, they are not always readable as texts. Just as literary critics can mistakenly amplify or dampen the significance of a passage or motif, so too can incautious quantitative analysis attempt to read signal into noise. And, by explicitly resisting the urge to *read* the results, to link sequentially the results of a topic model or narrate the historical axis of a graph, the analysis of the novel as data can offer even more radical reconfigurations of our understanding of the text.

Character Networks

Word order is also open to computation. Most models of textual frequency operate as a so-called bags-of-words. That is, they are quantified according to frequency, sacrificing any information on their order or position in the text. Yet this positional data are as informative as the frequencies, and through them, even the most interpretable lexemes in the text can carry meaningful quantitative information. Character names, for example, can reveal relationships in the novel simply by proximity. Two characters mentioned in the same paragraph (a function of proximity and word order) suggest a link between those two characters, whether they are acquaintances, enemies, or simply linked by the narrator of the paragraph. By quantifying this proximity, we can map the social space of the novel through character interactions.

Figure 11.3 shows the social space of *David Copperfield* as a network of interactions. Here each node (circle) is a character in the text, and an "interaction" occurs whenever two characters are named within the

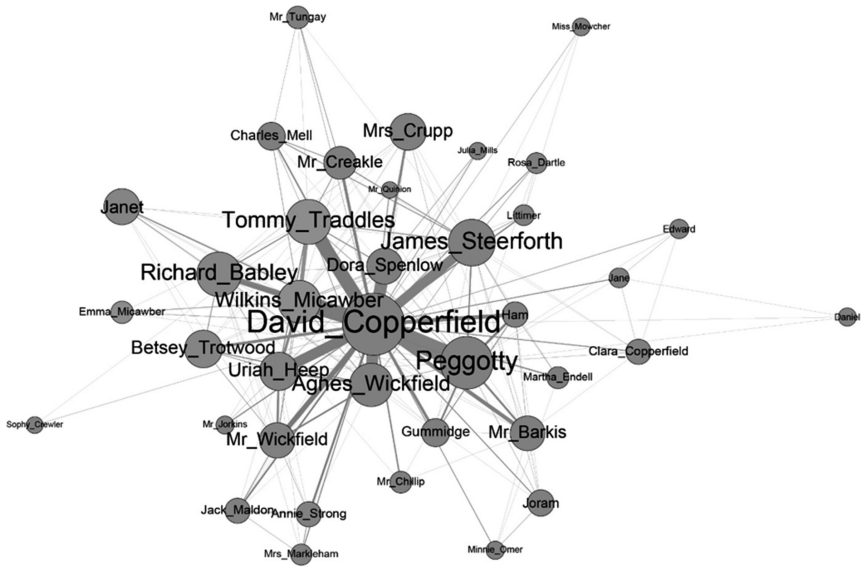

Figure 11.3: Social network of *David Copperfield*. Nodes are characters, edges are determined by the appearance of characters in the same paragraph, and the nodes are sized by eigenvector centrality. Eigenvector centrality measures the influence of a given node in a network. Its basic assumption is that a given node's influence should increase in proportion to (1) the number of other nodes to which it is connected and (2) the number of connections each of its connected nodes has. In a social network, if one person has a large number of connections and his or her connections also have large numbers of connections, that person would have a high eigenvector centrality. Cf. Ernesto Estrada, *A First Course in Network Theory*, Oxford University Press, 2015.

same paragraph. The more interactions that occur between two characters, the thicker is the edge between them. Although a reading of the text can offer an equally comprehensive (and more precise) understanding of the relationships between the various characters (including how they change over time), through the network, not only can we calculate the specific metrics that define these interactions, but we can also study the social universe as a total synchronic set of relations rather than a simulated dynamic system that emerges through the narrative of the text. As such, patterns of relationships that were not apparent from a reading of the text begin to emerge.

For instance, through this visualization of the novel's sociality, the most famous characters (besides David himself), Uriah Heep and Wilkins Micawber, appear as less important than Tommy Traddles, James Steerforth, and Richard Babley, all of whom co-occur with David more often than either Heep or Micawber. This, in turn, points to the novel's

asymmetrical representation of the social: by working from David's perspective, we learn more about his friends (Traddles and Steerforth) and their connections than we do about characters who arguably have a greater impact on David's life (Heep's machinations, Micawber's help to reveal Heep's fraud), even though he spends less time and mental effort on them.

Virtual Spaces

Moving beyond the text itself, quantifying the language of *David Copperfield* allows researchers to draw comparisons between it and other texts in Dickens' corpus. Using the frequencies of words, we can assess how similar *David Copperfield* is to other novels by Dickens.[34] And because these similarities (or dissimilarities) are functions of patterns of words, we can also see the ways in which two texts are like (or unlike) each other. In this model, two texts are similar if they share the same words in the same frequencies. Texts that contain substantially different words, or different frequencies of the same words, are less like *David Copperfield* than texts that share its vocabulary and proportions. Each word in this model of similarity acts as a dimension of comparison. So, across the full table of words, we can compare Dickens' novels in 13,441-dimensional space.[35]

Yet here we run into two problems. First, by including all the possible dimensions, we weight our model heavily toward the far end of the frequency distribution, toward words that appear once or twice in *David Copperfield* and not at all in any other text (e.g., the names of minor characters such as "Julia," who is only mentioned twenty-nine times by that name in the text). At the same time, the very high-frequency words receive an outsized weight, strengthening the authorial style signal rather than the semantic content of a given text. If the goal is simply to compare Dickens' corpus intratextually, then this is less crucial; however, if we want to assess the similarity between Dickens' texts and those of his contemporaries, then the author signal would reduce any other patterns to noise. In this analysis, we reduced the dimensionality of our comparison in two ways: first, by limiting the total number of words to 500 relatively high-frequency words and, second, by reducing the text of the novel to nouns only. A comparison between two texts based on this set of metrics, therefore, reveals the similarity of their most frequent nouns (rather than infrequent words or function words). Novels, through

this method, are similar when they share similar objects, people, or things in the same proportions. This necessarily limits our analysis to only one of many possible dimensions of comparison between texts, but one whose logic allows for meaningful conclusions to be drawn from the process of quantification.

The use of frequent nouns solves the problem of the high- and low-frequency words biasing the results toward authorial or uniquely textual signals. Yet the question of how to visualize even a 500-dimensional comparison remains. Even if we can assess the similarities between novels based on these features, the purely mathematical or statistical results resist meaningful interpretation within the context of literary studies. Figure 11.4 therefore reduces the dimensions of analysis from 500 to 2 through a "principal component analysis" (PCA). A PCA combines the various points of comparison (the individual words) into discrete principal components, combinations of words that most strongly separate the texts in the corpus. In Figure 11.4, the first two principal components are plotted against each other. In this graph, each dot is a text by Dickens, and distance in the virtual space of the graph is a marker of lexical similarity. Two texts that use the same nouns, in the same proportions, are closer to each other on the graph. Two texts that use different words, or different frequencies of the same words, are farther apart. The direction of their separation (horizontal or vertical)

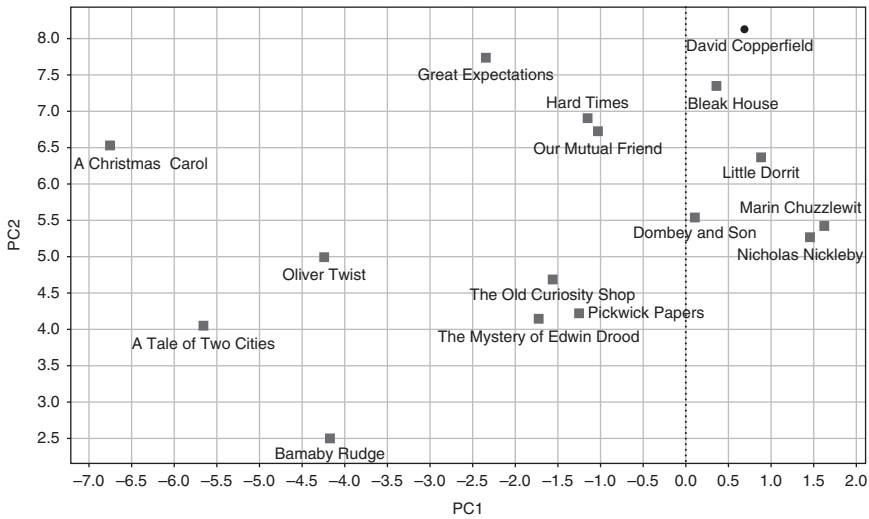

Figure 11.4: PCA of Dickens' corpus based on the 500 most frequent nouns in his texts.

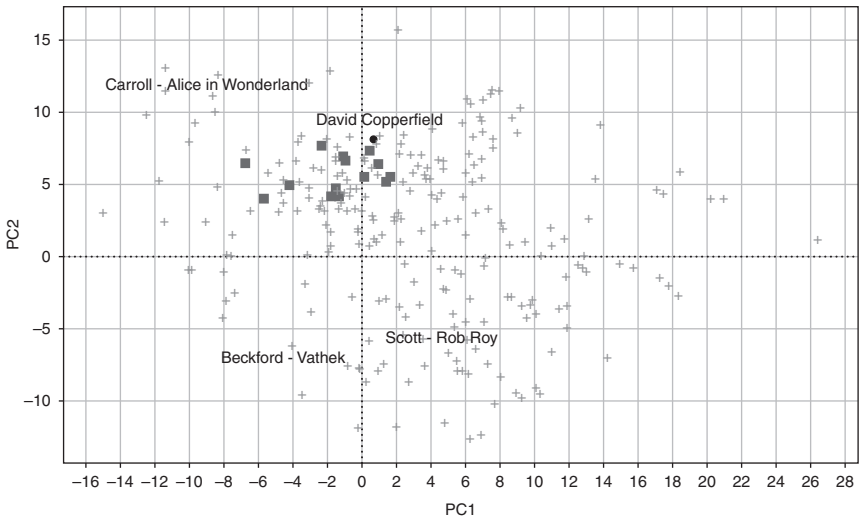

Figure 11.5: PCA of the Chadwyck-Healey corpus of nineteenth-century British and American novels based on the 500 most frequent nouns.

reveals some information on exactly which words are either pushing the texts apart or pulling them together. The virtual space of the PCA then translates complex lexical similarities into measurable distances at the cost of some precision.

From the graph we can observe the unique position of *David Copperfield* among the virtual space of Dickens' works. By itself, in the upper right corner, the text seems an outlier among the corpus, although not as much as *A Christmas Carol* on the opposite side of the chart, whose complete isolation speaks to its differences of method, genre, and intended audience within Dickens' oeuvre. By the measurement of its most frequent nouns, *David Copperfield* is most like *Bleak House*. This is a logical result: both texts deal with themes of class and deception, are set in cities, and are among Dickens' longest works (even after scaling to adjust for length as part of the comparison process, length still carries some weight in the analysis). Interestingly, *Little Dorrit* and *Hard Times* are the next closest novels, suggesting that these four retain some similarity in their most mentioned nouns, a function of their similar settings and conceptual concerns.

Again, among the advantages of a data-driven approach in literary studies is the ability to move between scales. Figure 11.5 shows the same arrangement of Dickens' texts, only now in relation to a larger corpus of British and American Victorian texts from the Chadwyck-Healey collection of novels.

This corpus, while not truly representative of literary production in the nineteenth century, nevertheless contains an expanded canonical sample of roughly 700 texts from the period.[36] From this graph, then, we can gain a better understanding of where Dickens, as an author, fits within the literary field of his period, as well as the location of *David Copperfield* in relation to the new texts. Again, the distances on the graphs are virtual representations of lexical similarity between the texts, specifically comparing their 500 most frequent nouns.

The positions of the other texts in the corpus reveal many of the forces dictating the specific arrangement of Dickens' texts. The two novels that are among the farthest from *David Copperfield* – *Oliver Twist* and *A Christmas Carol* – occupy the same quadrant that, at its extremity, contains Lewis Carroll's *Alice in Wonderland* and Anna Sewell's *Black Beauty*. This suggests that the objects in these texts, among the more popular in Dickens' corpus, are more in keeping with a certain kind of nineteenth-century children's literature than his more complex or satirical books. *A Tale of Two Cities*, however, and *Barnaby Rudge* and the unfinished *Mystery of Edwin Drood* all lie in the same direction as William Beckford's Gothic novel *Vathek*. Because *Barnaby Rudge* begins with a story of murder, which echoes the mysterious disappearance of Edwin Drood, it is logical that both should share, at least in part, a lexicon with historically set Gothic novels. Finally, although none of Dickens' corpus overtly resembles Walter Scott's historical novel *Rob Roy* (situated in the lower-right quadrant among Scott's other novels), the descriptive work of *Nicholas Nickleby* most closely echoes Scott's depiction of the sociocultural and economic conditions faced by both the highland and lowland Scots in Scott's *Waverley* novels.

By treating Dickens' novel as data, by quantifying the data and comparing them statistically with both the rest of Dickens' own corpus as well as those of his contemporary canonical authors, new patterns of resemblance begin to emerge. Although they serve only as a prompt for further interpretive work as we question the results and seek to explain the axes of differentiation, the novels of the nineteenth century, when transformed into data, quantified, and analyzed as numerical objects, reveal more and deeper patterns than human readers could possibly perceive. Again, the strength of the quantitative analysis lies in *relative* measures, not absolute statements of identity. Things are more or less like each other, allowing us to speak of the nineteenth century as a spectrum: in this case, a spectrum aligned along "object-ness" in Dickens.

Reading across Scales

But the patterns that we have seen in *David Copperfield*, and in Dickens' texts more generally, offer little meaning when viewed in the vacuum of a purely quantitative analysis. Instead, these objects give us new entry points into an interpretation of the text. But the return from a distant reading to the text itself is disorienting: informed by the computational study, what we notice about its ideas and emphases has shifted. The graph allows us to view a novel, or a corpus, at a glance. Seeing a novel not as pages turned over hours but as a simultaneous array of words in multiple relations with others and themselves is a position more attuned to the novel as an assemblage: words in an order, a node in a network, a point in vector space. These transformations, however, point us toward unexpected local manifestations of larger phenomena.

For example, the noun most distinctive of Dickens across his writing when compared with the larger nineteenth-century corpus is "head." In other words, compared with the other texts in the Chadwyck-Healey novel corpus, the word "head" appears significantly more often in Dickens' novels than we would expect if it were randomly distributed across the entire corpus.[37] It is therefore a lexeme that marks both a distinctive authorial habit and an important site of attention for a critic. Unlike the stopwords that we discussed earlier, "head" is both used significantly more frequently and is a particularly significant word, the double valence of *significance* signifying both statistical and interpretive importance. "Head," then, not only provides a lexical axis with which to differentiate Dickens' writing from that of his contemporaries, but it also bears a significant thematic and conceptual burden, perhaps best expressed in *David Copperfield*. Its very prevalence, which suggests its unimportance to the reader of *David Copperfield*, turns out to be a measure of its significance within the larger Victorian context. "Head" is a peculiar word to read closely considering that it is far more common and, as a result, more polysemic than abstractions such as "bildung" or "selfhood": "head" is unlikely to be a word on which a critic would build a close reading. On rereading Dickens' works with the knowledge of its surprising statistical significance, however, we can recognize an emphasis on the head as characteristically Dickensian in style and theme. In *Great Expectations*, Dickens is careful to describe Mr. Jaggers' "exceedingly large head and a correspondingly large hand," amplified by equally large fingers that Jaggers constantly bites in a gesture meant to menace but that

comes off as almost childish.[38] Like a caricaturist, Dickens inflates Jaggers' head to make it easy for the reader to see what one needs to know about him.

For Dickens, the head is as much the physical repository of brain and mind as it is a bodily vehicle of expressing the unsaid and the unsayable. In conversation, his characters constantly "shake," "nod," "motion with," "twist," "throw back," and "sit with their hand upon their head," among other gestures of affirmation or disavowal, as the situation demands. But heads also "confuse," "[get] heavy," "ache," and "grow light," precipitating narrative developments rather than merely reflecting on them. Most of all, thoughts, ideas, and impressions "come into" his characters' heads, which is to say that they rise to the foreground, not of perception, but rather of attention, thereby becoming part of the narrator's, and, by extension, the reader's interpretative framework.

David Copperfield's focus on the protagonist's developing mentality, as he forms a sense of self in a deceitful world while cultivating writerly sensibility, suggests the particular importance of the head as a thematic locus for the novel. But the connection between the word "head" and these concepts, which are crucial for a close reading, could only have come from a recognition of the significance of "head" in a quantitative analysis of Dickens' works. For example, consider these passages, separated by a chapter break, near the end of the novel:

> And Agnes laid her *head* upon my breast, and wept; and I wept with her, though we were so happy.
>
> [It] was like the beginning of a favorite story Agnes used to tell them, introductory to the arrival of a wicked old Fairy in a cloak who hated every body, it produced some commotion. One of our boys laid his *head* in his mother's lap to be out of harm's way, and little Agnes (our eldest child) left her doll in a chair to represent her, and thrust out her little heap of golden curls from between the window-curtains, to see what happened next.[39]

Across these two passages, we see that the head makes contact with the body ("upon my breast" and "in his mother's lap") in a gesture of physical and narrative intimacy that both concludes the marriage plot between David and Agnes and points toward the lives of their children. The mirroring of the head movement not only establishes a physical continuity between marriage and childhood but also suggests an epistemological continuity whereby one's mental life is physically

connected to and even "leans on" the family. Moreover, the children listening to the story, who want to be "out of harm's way" and yet "see what happened next," almost come to personify novel reading itself, heads tilted down and simultaneously looking out "between the window-curtains."

Conclusion

We have suggested that the difference between a close reading and a data-driven, quantitative analysis of the novel lies in a transformation from a synthetic approach to the novel as a singular object and a constitutive approach to the novel as an aggregation of features that is itself only one data point among a corpus of thousands. At stake in the transformation, then, is our understanding of the novel itself: what it is, what information it contains, and what it can do as both a literary and critical object.

Quantitative analysis offers new scales through which to read the novel. At the macroscopic level, it can include a wealth of new texts, forgotten or dismissed from literary history, against which to measure intuitions and conclusions that have been gleaned from the canon so far. We can compare texts mathematically or statistically across centuries, nationalities, and genres and create new constellations of resemblance and difference. At the microscopic level, we can trace the patterns of individual words through an author's corpus, or text, revealing the textual threads that hold together the synthetic conceptual features that lie at the level of reading. The information that we gain from either analysis is different in kind from that revealed by reading, even close reading, a single novel. If the former provides document-term matrices, eigenvector centralities, and principal components, then the latter offers ideas, concepts, and affect. The former offers ways of measuring the novel, of operationalizing the text along the axis of literary criticism or history. The latter offers an interpretative understanding of the text that reconciles the experience of reading and the reader's knowledge and intuition with the formal features of the novel. They are not, however, incompatible. Both rest on abstractions of the novel and, more importantly, both require interpretative acts on the part of the researcher to create meaning from the analysis.

And it is in interpretation that the power of recasting the novel as data reveals its true transformative potential. If quantitative analysis requires strategies of close reading in order to make meaning out of its

mathematical operations, so too do the practices of close reading require the new scales of analysis available only through quantitative study. The new methods that we have outlined work hand-in-hand with the methods of literary criticism and history as they have been traditionally practiced by scholars. Reading the novel as data while forgetting the disciplinary lessons of literary study foreshortens our ability to interpret the results of our analysis. But, by the same token, as quantitative textual analysis reveals the omissions, biases, and limitations of close reading, so too has it become equally necessary to attend to these new critical practices. Already they are on the brink of transforming our understanding of individual novels or authors, and yet the potential of this new method of literary criticism is greater still. As we mobilize these methods to access new ways of understanding novelistic prose, as we compare the canon and archive of the novel not only among themselves but against all of the writing, fictional and nonfictional, of a given historical period, we gain the potential to not just analyze a novel, or even many novels, but to finally assess the true role of *the* novel in shaping the ideas, values, and even the very culture of every period that the novel has touched.

NOTES

1. Franco Moretti, "Conjectures on World Literature," in *Distant Reading*, Verso, 2013, pp. 43–62; Stephen Best and Sharon Marcus, "Surface Reading: An Introduction," *Representations*, vol. 108, no. 1, 2009, pp. 1–21.
2. Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History*, vol. 46, no. 1, 2015, pp. 63–98; Matthew Wilkens, "The Geographic Imagination of Civil War-Era American Fiction," *American Literary History*, Vol. 25, no. 4, 2013, pp. 803–40.
3. Recent work at the Literary Lab has shown that the "average" nineteenth-century novel has a high frequency of "the" at both the beginning and the end of the text. In *David Copperfield*, the frequency of "the" is depressed at the beginning relative to the end, and the word "I" is actually more frequent than "the" at key moments of the narrative.
4. Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, University of Illinois Press, 2011.
5. John W. Tukey, "The Future of Data Analysis," *Annals of Mathematical Statistics*, vol. 33, no. 1, 1962, p. 2.
6. Claude Shannon and Warren Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, pp. 8–9.
7. Roman Jakobson, "Closing Remarks: Linguistics and Poetics," in *Style in Language*, ed. Thomas Sebeok, Wiley, 1960, pp. 350–77.

8. A "corpus" is any set of texts considered together. For instance, imagine a corpus of all novels published in 1899 or all the poems by the Lake Poets.
9. See examples of this in Andrew Piper and Mark Algee-Hewitt, "The Werther Effect: I. Goethe, Objecthood and the Handling of Knowledge," in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed. Matt Erlin and Lynne Tatlock, Camden House, 2014, pp. 155–84.
10. See, for example, Matthew Jockers, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, 2013, p. 28.
11. Which remain highly contingent on the methods of collapse. See, e.g., D. Sculley and B. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities," *Literary and Linguistic Computing*, vol. 23, no. 4, 2008, pp. 409–24.
12. Franco Moretti. "'Operationalizing': or, the function of measurement in modern literary theory," pamphlets of the Stanford Literary Lab, June 2013.
13. Michael McKeon, *The Origins of the English Novel, 1600–1740*, Johns Hopkins University Press, 2002; Ian Watt, *The Rise of the Novel*, University of California Press, 2001.
14. Matthew Arnold, *Culture and Anarchy*, ed. Jane Garnett, Oxford University Press, 2009.
15. Margaret Cohen, *The Sentimental Education of the Novel*, Princeton University Press, 1999, p. 23.
16. Franco Moretti, "The Slaughterhouse of Literature," in *Distant Reading*, 65.
17. Lee Konstantinou, "Am I Turning Empirical?," Arcade, Stanford University, available at http://arcade.stanford.edu/blogs/am-i-turning-empirical.
18. Pierre Bourdieu, *The Rules of Art: Genesis and Structure of the Literary Field*, Stanford University Press, 1996.
19. Eltjo Buringh and Jan Luiten van Zanden. "Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries," *Journal of Economic History*, Vol. 69, No. 2, 2009, pp. 409–45. Based on the accumulated data in Tables 11.1 and 11.2, across all the languages of Western Europe, the total number of manuscripts and printed books available in 1500 would have been roughly 10,906,675.
20. "Books in Print, Global Edition," Bowkers, November 16, 2016, available at www.booksinprint.com/Search/Results?q=%7b%7b(publisher-date%3d%5b2015-01-01~To~2015-12-31%5d)%7d%7d&ast=se.
21. Cf. Andrew Piper, "There Will Be Numbers," *Journal of Cultural Analytics*, May 23, 2016, available at http://culturalanalytics.org/2016/05/there-will-be-numbers/.
22. Here we use "falsifiable" in the sense of Karl Popper, who gives it a privileged position in his proposed revisions to the scientific method. To be falsifiable (and therefore empirical), a theory must be able to be wrong if one or more occurrences that it has ruled out occur. Karl Popper, *The Logic of Scientific Discovery*, Routledge, 2002, p. 68.
23. Mary Williams, *The Dickens Concordance, being a compendium of names and characters and principal places mentioned in all the works of Charles Dickens*,

Francis Griffiths, 1907, available at https://archive.org/details/dickensconcordanoowillrich.

24. Alex Woloch, *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*, Princeton University Press, 2009.
25. This would hardly constitute a research project on an unprecedented scale. For reference, Proust's *In Search of Lost Time* is 1,267,069 tokens, often printed in six volumes. A corpus with 474 such volumes would exceed 100 million words. By contrast, the complete HathiTrust corpus has over 814 billion tokens (cf. David McClure, "Counting Words in HathiTrust with Python and MPI," Techne, Stanford Literary Lab Blog, August 26, 2016, available at https://litlab.stanford.edu/counting-words-in-hathitrust-with-python-and-mpi/).
26. Sample files of this type can be easily found on sites such as Project Gutenberg or Archive.org, both of which maintain a mixture of manually entered and human-corrected texts that were typed by hand or scanned from the original and converted to plain text through a process known as "optical-character recognition" (OCR). These files are "plain" in the sense that they only allow minimal characters (letters, numbers, spaces, symbols, carriage returns, etc.) and eliminate other types of formatting or media (bold, italics, images, etc.).
27. George Zipf, *The Psychobiology of Language: An Introduction to Dynamic Philology*, MIT Press, 1935.
28. Peter Stockwell and Michaela Mahlberg, "Mind-Modelling with Corpus Stylistics in *David Copperfield*," *Language and Literature*, vol. 24, no. 2, 2015, pp. 129–47.
29. Ibid., p. 136.
30. David Hoover, "Statistical Stylistics and Authorship Attribution: An Empirical Investigation," *Literary and Linguistic Computing*, vol. 16, no. 4, 2001.
31. J. Rybicki, M. Eder, and D. Hoover, "Computational Stylistics and Text Analysis," In *Doing Digital Humanities: Practice, Training, Research*, ed. C. Crompton, R. J. Lane, and R. Siemens, Routledge, 2016, pp. 123–44.
32. Indeed, at times these words almost literally become invisible. Psychological studies have shown that readers frequently fail to notice unnecessarily duplicated and misspelled words. Cf. K. Rayner, S. J. White, R. L. Johnson, and S. P. Liversedge, "Raeding Wrods with Jubmled Lettres: There Is a Cost," *Psychological Science*, vol. 17, no. 3, 2006, pp. 192–93.
33. Dawn Archer, *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*, Routledge, 2012.
34. In making cross-novel comparisons, we scale word frequencies to account for differences in text length, so instead of simply transliterating the fact that "David" appears sixty-five times in *David Copperfield*, we convert it to a frequency, recording that it represents 0.0002 percent of words in the text.
35. This is possible because the novels use 13,441 types (unique words), with many millions of individual tokens.
36. "Nineteenth-Century Fiction." Chadwyck-Healey (ProQuest), November 15, 2016, available at http://collections.chadwyck.com.
37. The statistical significance of the use of "head" here was determined by a Fisher's exact test, which gives each word a *p*-value indicating whether it passes the threshold for significance within the text of a single novel or author. With

a p-value of 1.44×10^{-25}, it is clearly highly significantly less than the 0.05 threshold for significance.

38. Charles Dickens, *Great Expectations*, ed. Margaret Cardwell, Oxford University Press, 2008, p. 75.
39. Charles Dickens, *David Copperfield*, ed. Nina Burgis, Oxford University Press, 2008, p. 844 [our emphases].