CHAPTER 1

# *Introduction*

This book is admittedly unusual in that its goals are twofold. First, it is an interdisciplinary study of literary dialects, their history, and their connections to British imperial ideologies. Second, it is an argument for and a demonstration of an approach to digital research: dynamic reading. Before getting into the details of either the study or the approach, I think it is important to explain why the book is set up in this way.

The first reason relates to the nature of the approach and its differences from other computationally oriented methods of textual analysis in the humanities. Typically, demonstrations of method (such as Jocker's (2013) excellent *Macroanalysis*) are executed through a series of small case studies, showing, for example, how we can uncover cycles of genres or stylistic differences based on nationality. This is done because scholars want to illustrate the broadest potential application of their analytic tools. Dynamic reading certainly shares an interest in further building the case for computational methods in the humanities. Its name is partly an indebted nod to Moretti (2005) and others who have argued that computationally assisted forms of textual analysis constitute a form of "reading." Moretti's specific formulation is "distant reading," a name that emphasizes an analytical perspective operating from a bird's-eye remove. However, as the name *dynamic reading* implies, my focus is not solely on the distant. It is on the articulation between the distant and the close, on the ability of analysis to operate at different levels of resolution.

That articulation is formulated through a mix of quantitative and qualitative methods and seeks to marshal the combined power of digital tools and digital archives. The argument for qualitative close reading as an important resource in our digital toolkit – one that widens the pool of available data and one that can complement computational analysis – is a defining characteristic of dynamic reading. It is also more effectively illustrated through a longer case study than a series of shorter ones. Unlike quantitative techniques whose efficacy can be quickly gleaned from well-

designed and well-presented visualizations, the persuasiveness of qualitative evidence is built up through repeated examples and sustained discussion.

The trade-off, of course, is that a longer case study showcases less breadth of application than do shorter ones. This is mitigated somewhat because this book analyzes how three different groups of speakers have been voiced in British fiction: African diasporic, Indian, and Chinese. Each of these literary dialects has a distinct history. Those histories are shaped by particular political events, ideological currents, social anxieties, and aesthetic fashions. Exploring those histories requires a continual reassessing and adjusting of methods. Thus, within the larger case study are embedded three smaller variations, which provide at least a somewhat broader look at the approach and its applications.

The second reason for the dual goals is that each is equally important, though important for different reasons and likely interesting to different audiences. The project began not as methodological experimentation but as a (not very successful) qualitative study of written representations of African diasporic vocal culture. That project eventually morphed into the more computationally intensive one that is presented here. Despite some radical changes in methodology, however, my original motivations remained. I wanted to explore how routines of linguistic mimicry propagate, as well as how they are implicated in the perpetuation of racist ideologies and asymmetries of power. Although these routines are just a small part of larger apparatuses that serve to uphold the social, political, and economic interests of the dominant culture, they provide insight into related processes: how mechanisms (like perceptions of language variation) often operate below the level of consciousness; how societal paranoia in response to shifting demographics can infect discourses surrounding community practices; and how these discourses can serve as proxies for other kinds of evaluations that are used to justify fear, oppression, and violence.

My analysis focuses specifically on the period beginning in the late eighteenth century and ending in the early twentieth and on the literary dialects that British authors used during that time to voice African diasporic, Indian, and Chinese characters. The results identify the vast array of features (related to grammar, spelling, and vocabulary) exploited in representing these vocal cultures, features that differentiate their fictional dialogue from the dialogue of other characters and the narration. Further, they show how conventions for representing speakers change over time, as well as how dialogue can not only align with those conventions but also deviate from them. Those various descriptions of similarity and of

difference, of structure and of change, are then examined in their relationship to the imaginings of empire – its subjects, conditions, and purposes.

While the contours of the project reflect my own interests – in language history and in social justice – the methodological principles that this book details have much broader application. The identification of patterns in language, after all, is relevant to a range of fields and disciplines. In fact, part of what makes this study a compelling example of dynamic reading is its interdisciplinarity, bridging as it does corpus linguistics, historical sociolinguistics, and colonial discourse studies. But before diving into the specifics of the project, it will be helpful to outline how distant and close reading might function within a comparative framework, particularly in light of the sometimes contentious debate within literary studies.

In his remarks on the relationship between close reading and distant reading, Moretti (2005: 74) is clear on where he stands:

> Between interpretation (that tends to make a close reading of a single text) and explanation (that works with abstract models on a large groups of texts) I see an antithesis. Not just difference, but an either/or choice.

Jockers (2013: 9) expresses a similar view in his discussion of macroanalysis:

> The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random "things" gathered from a few, even "representative," texts. We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly "uninteresting" texts.

Despite the binaries that such statements may appear to construct, I want to be clear that I do not see dynamic reading as a rebuke of Morettianism or some kind of accommodation between it and traditionalism. I see it, instead, as an approach with a somewhat different focus than distant reading but allied with it, nonetheless. Moreover, I do not believe that the way close reading is conceived of within the framework of dynamic reading necessarily contradicts what either Moretti or Jockers argues in the preceding quotations.

I think proponents of computational methods (myself included) sometimes cast a skeptical eye on close reading because it can be used to undergird claims of generality that do not hold up to scrutiny. There is certainly nothing controversial about qualitatively analyzing a novel to make claims about that novel. However, we often want to do other things. We want to use that novel to instantiate claims about culture or politics or ideology more broadly. And this is where, I suspect, Jockers's incredulity

comes in. How can we possibly say that a novel is representative of anything larger given that our reading of it and its very selection as a text worth reading are subject to and the product of our own biases?

Computational analysis is one way of mitigating researcher bias. Rather than considering just one novel, we can consider many. And rather than presumptively assuming that a phenomenon exists and is represented by a work, we can see what patterns emerge from a collection of works, describe those patterns, and explore the contributions of individual examples. That research process does not preclude close reading. In fact, in fields like corpus linguistics (which is where my training comes from), it is standard practice to combine qualitative and quantitative data.

To sketch out how we might do this, let us consider a hypothetical study in which we are interested in how Native Americans are represented in fiction. We could build our own dataset, but for the purposes of this thought experiment we'll rely on data from the Corpus of Historical American English (Davies 2010–). One thing that we could do would be to look at the adjectives that occur before nouns that identify native characters such as REDSKIN. The results show that while a few positive adjectives modify REDSKIN (e.g., *noble* and *lithe*), most are negative (e.g., *vagrant, thieving, skulking, marauding, low-down, ignorant, hostile, ferocious, dirty, cruel, cowardly, bloody, bloodthirsty*). If we wanted to be even more rigorous in our computational methodology, we could use a measure such as Mutual Information to calculate the strength of association. In other words, does a token such as *ignorant* significantly associate with REDSKIN, or is their appearance together just an artifact of *ignorant* being a common word? We find that many of these pejorative adjectives have a high Mutual Information score: for *wily*, MI = 10.35; for *skulking*, MI = 9.48; for *bloodthirsty*, MI = 9.35; and for *ignorant*, MI = 5.05. (A significant association occurs when MI is greater than approximately 3.)

As a next step, we might want to examine their use in context. In corpus linguistics, this is usually done with what are called concordance lines or Key Words in Context (KWIC). This is where a kind of close reading comes into play. We want to get a clearer sense of how meaning is being made – in this case around our two-word phrase or what is called a "bigram." A small sample of concordance lines is presented in Table 1.1, and we could make a number of observations. For one, it is clear that these figure fictional Native Americans pejoratively on the whole, as we would expect. The fourth line, however, is ambiguous as the bigram *wily redskin* is in scare quotes. The following one, too, is preceded by the negator *not*, perhaps suggesting an equivocal or even romanticized portrayal. If we were

Table 1.1 *Concordance lines from COHA showing adjectives collocating with* REDSKIN *one token to the left*

| | |
|---|---|
| in too much hurry," rejoined the | **wily redskin**. "I was told the camp was but |
| dian's side, to make certain the | **wily redskin** was not shamming, he found Ye |
| frustrate all his plans. But the | **wily redskin** was not to be so easily caugh |
| bor. "We have had plenty of the | '**wily redskin**' kind of thing," I said to St |
| for the saints were not | **bloodthirsty redskins** but the descendants of Laman |
| s of dwarfed timber in which | **skulking redskins** could watch them come. That |

actually conducting this study, we would want to unpack our interpretations of these lines in detail. Doing so is intended to create a fuller and more defensible accounting of the pattern: quantitative analysis can demonstrate that adjectives with negative semantic resonances tend to co-occur with REDSKIN, and the qualitative analysis can explicate how those co-occurrences function in the surrounding discourse.

In broad strokes, that is one way that corpus linguistics integrates quantitative and qualitative analysis. About that integration, I would point out a few things. First, because the qualitative analysis follows from the quantitative, it is not "random" in the way that Jockers suggests other kinds of close reading of isolated texts can be. Second, it is not based on the interpretation of individual texts in the way that Moretti frames close reading. Even in the qualitative reading of concordance lines, corpus analysis of this sort is based on the accumulation of demonstrably related evidence. That said, I doubt that practitioners of distant reading would categorize a study like the one I have outlined as such. Distant reading is generally more interested in uncovering global, systemic patterns like the changing influence of gender on novelistic themes than on individual word collocations. Nonetheless, it illustrates the potential for computational explanation to drive qualitative interpretation.

This, then, brings me to dynamic reading and its differences from something like the hypothetical study I outlined. In corpus linguistics, all of the analyses – quantitative and qualitative – are generated from data internal to the corpus. What dynamic reading proposes is an articulation between quantitative analysis of data internal to a corpus and the qualitative analysis of evidences external to that corpus. Consider an artifact such as the mid-twentieth-century comic book panel in Figure 1.1. Alone, it is perhaps a suggestive text, but in the context of quantitative data that we generated from the Corpus of Historical American English it takes on new importance. It does not repeat any of the REDSKIN bigrams. In the panel,


Figure 1.1 A panel from the comic book *Western Thrillers* (Fox 1949)

*redskins* stands by itself as an exclamation. However, the Native American chief is described as "crafty," echoing modifiers from the corpus like *wily* and *skulking*. Furthermore, the comic associates an imagined "redskin" culture with extreme, unjustified violence and analogizes its collective identity to insects, to "a swarm of wasps." Again, these characterizations echo collocates from the corpus including *marauding, hostile, ferocious, dirty, cruel, bloody*, and *bloodthirsty*.

A case could be made, therefore, that the comic fits into a verifiable pattern. For a fully realized analysis, we would need to include additional artifacts and more detailed close readings. Like the computational analysis of corpus data, the qualitative analysis of evidences is grounded in a fundamental understanding of discourse as regimented. It is an understanding that is shared by Foucauldian traditions or modes of discourse studies like critical discourse analysis and colonial discourse analysis that explore the production and reproduction of semiotic routines that are conditioned by time and place. One method for exposing the cumulative effects of these routines is to demonstrate the contributions of a variety of evidences to a larger pattern. This is the approach of Shohat and Stam (1994), for example, in *Unthinking Eurocentrism*. Just as there is explanatory power in the counting of things, so too is there power in their juxtaposition. Moreover, the analysis of evidences can be connected to histories – the structuring of social relations, the workings of institutions, the policing of bodies. The comic in Figure 1.1 might be linked to the emergence of comic books as a new kind of print culture, its role in promoting mythologies of the American West, and the relationship

between those mythologies and the government's genocidal policies toward Native Americans.

Again, this is nothing new. It has long been a tenet of discourse studies that qualitative analysis can expose discursive patterns that are largely invisible and that mediate how we understand and operate in the world. Scholars like Fairclough (1995, 1992), Van Dijk (1993), Wodak (2001, 1999), and Jørgensen and Phillips (2002) have made these arguments in far more detail and far more eloquently than I have here. Neither is it a new observation that the emphasis on lexical and grammatical patterns in certain types of discourse analysis might be productively allied with computational methods like those in corpus linguistics. In fact, early critics of critical discourse analysis like Widdowson (1995), Stubbs (1997), and Toolan (1997) suggested that it suffered from a randomness of data selection that could be mitigated by corpus methodologies in much the same way that Jockers questions data selection in literary studies. Indeed, researchers have since published a wide range of studies that have married corpus linguistic and discourse analytic frameworks (e.g., Baker et al. 2008; Caldas-Coulthard and Moon 2010; Mulderrig 2011, 2012; Orpin 2005). In other discourse studies traditions like colonial discourse analysis, the application of computational methods has been less frequent, perhaps because the field is more "abstract" than "linguistically oriented" (according to Fairclough's definitions). Yet, when Said (1994: 203) announces the need to understand "the Orient" as linguistically constituted because "the Orient was a word which later accrued to it a wide field of meanings, associations, and connotations, and that these did not necessarily refer to the real Orient but to the field surrounding the word," it is not difficult to imagine how computational methods might be mobilized in pursuit of such a project. And some have suggested the potential for merging the Foucauldian "archeological" methods that Said draws upon with corpus approaches (Koteyko 2006).

What is different about this study is neither its methodological pieces nor the idea that those pieces might be brought together in useful synergies. What is different is *how* it proposes that those pieces might be brought together. One way to think about the relationship between the quantitative and qualitative analysis is to imagine each archived digital artifact as a node in a vast network. Those nodes are potentially connected by themes, tropes, and ideas all constituted though discourse. At first, however, all of those nodes and connections are darkened. The computational analysis lights up connections among a select subset of works and in those connections are clues – structural through lines and disruptions, likenesses and

contrasts. From those clues, tracing possible connections to additional nodes requires reading new artifacts. In its simplest terms, the analysis begins with a machine identification of patterns, which, in turn, informs a human one.

Such an integrated approach has a number of benefits. As I have alluded to a few times already, quantitative analysis enables us to process volumes of text that would otherwise be impossible and to see patterns that would otherwise be invisible. It also provides a check against our own biases and assumptions. Alternatively, while computers are fast and efficient, they are not smart. They perform precisely as we tell them to. In qualitative analysis, we can perceive relationships and contextual meanings that would be difficult (if not impossible) to train a machine to "see." Each kind of reading, distant and close, possesses different strengths and enforces different types of rigor.

There are parallels in other disciplines to the debates going on in literary studies about the relative merits of quantitative and qualitative analysis. A useful one, I think, comes from evolutionary genetics. In that field, questions about genetic versus archeological evidence align rather nicely with those about distant versus close reading. The correspondences between traditional human-driven methods (archeology and close reading) and recent computer-driven ones (distant reading and genetics) are fairly clear and well established (e.g., Foucault's characterizing of his method as "archeology"; geneticist Alberto Piazza's afterword to Moretti's *Graphs, Maps, Trees* in which he argues explicitly for the application of evolutionary concepts in literary studies). They even extend beyond the conceptual to the operational as many of the statistical techniques used in the computational analysis of literary history like forms of cluster analysis are either borrowed directly from or are used extensively in bioinformatics. It is not difficult, therefore, to hear echoes of their debates in our own, and when the evolutionary geneticist Michael Hammer (2003) argues for an integrated analytical approach, his reasoning is an analogue for mine:

> From genetics alone we can't tell all that much. We need to have a context to work in. So, if I'm interested in the peopling of the Americas – how long ago did people move into the Americas, how many people moved to the Americas, how many times did they move in the Americas – I can get genetic data that will show me patterns of variation of the Americas and I can compare those data with patterns of variation in Asia. But I need calibration points from the archaeological record, to know when we see evidence of culture in the Americas, how does that culture relate to culture in Asia? And it's a comparative process through genetics and archaeology . . . You have to

> put the picture together with all of those pieces of the puzzle. One piece of the puzzle alone won't give you the whole picture. So we shouldn't lose sight of that. As powerful as genetics is, as a tool, to look at our own history, it can't tell us anything by itself. It has to be in a comparison framework.

Hammer's notion of a "comparison framework" captures a good deal of dynamic reading's purpose. As Hammer describes, its principal animating concern is the assimilating of different types of evidence to make sense out of patterns of variation (in his case genetic and in our case linguistic). Key mechanisms within that framework are "calibration points" – correspondences that join patterns in one data type to those in another. In my earlier analogy of a network, these are much the same as the nodes that that once illuminated can be links to new evidences. An example of a calibration point would be the collocational patterns that we saw with REDSKIN and extended from the corpus to the comic book panel. That was a useful illustration of how we can forge connections between two entirely different sets of digital data (a massive monitor corpus and a comic book archive) applying two kinds of analysis (quantitative and qualitative). However, as I noted, the scope of the hypothetical study was intentionally narrow and the calibration points intuitive. This study proposes the analysis of a much more complex set of variables and the application of more robust computational techniques. I will explore how we can calibrate data under these more challenging conditions in subsequent chapters.

As much as Hammer's quotation is helpful in setting out the broad parameters of my methodological project, I would distinguish dynamic reading from his description of work in evolutionary genetics in at least one important way. He begins his statement with an assertion that genetics by itself "can't tell [us] all that much." In the digital humanities and in corpus linguistics, I would argue that computational analysis can tell us a great deal. Furthermore, in the pursuit of particular research questions, the right kind of computational analysis may be all that we need. Rather than an argument for an approach that is universal, dynamic reading is an argument for an approach that is targeted. I see it as an alternative way to participate in digital humanities research, one that facilitates the interrogation of textual elements: narrative structures, descriptions, characterizations, and so on. And in opening up the types of data that are available to us, it invites new ways of thinking about their relationships and new ways of designing the projects we are interested in pursuing.

This latter point is critical. "Big data" research projects in the digital humanities have certainly prompted fields that have not traditionally

used quantitative methods to confront their place within those fields. In so doing, they have opened up spaces for new kinds of scholarship. However, research of this kind can also seem to foreclose participation for some. Data collection alone can prove daunting. If one wants to build a corpus of nineteenth-century novels, for example, it is possible to purchase some data and to get permission to use others. But if the goal is to amass thousands of works for a unique corpus, the labor involved in collecting the data is likely to scare off students and scholars who by choice or necessity do not have the benefit of collaborators. Moreover, digital data usually need to be cleaned and prepared before they can be analyzed, necessitating an even greater commitment. This is to say nothing of the barrier to entry (real or perceived) posed by the technical expertise needed to carry out the statistical analysis that undergirds computationally intensive approaches.

In thinking about how we triangulate among different data, we can rethink not only the role of close reading but also the shape and function of our computational analysis. Depending on our research questions and our strategies in exploring those questions, our corpora may be smaller and more specialized – "bigish" rather than "big." Specialized corpora can still be statistically robust, with the advantage that we can be better acquainted with the works that populate them, thus enhancing our ability to contextualize statistical information with close reading. In addition to occasioning a critical rethinking of the data required for our analysis, dynamic reading calls for a more transparent and explicit rationalizing of our statistical methods. Because so many of our computational tools are borrowed from other disciplines, there can be a tendency to replicate techniques and present visualizations with little discussion of the strengths and limitations of those tools. Much of our data in the digital humanities are noisier than data in bioinformatics, for example. How might that fact affect our understanding of a measure like *p*-values? Should we reflexively replicate the widely accepted standard of $p < 0.05$ as a threshold for significance? Or would we be better served in understanding a little of the debate among statisticians about the value of that threshold and consider its usefulness within the context of our own work?

This book, therefore, aims to explore a range of issues related to digital data and their analysis. How do we select and process digital data? How do we choose computational tools and measures appropriate for our research? How do we in the humanities effectively and fairly represent our statistical findings? To sum up some of the book's purposes, here are what I would consider to be the first principles of dynamic reading: