Demography of Literary Form: Probabilistic Models for Literary History

by

Allen Beye Riddell

Graduate Program in Literature
Duke University

Date: _____

Approved:

_____
Katherine Hayles, Supervisor

_____
Michael Hardt

_____
Timothy Lenoir

_____
Cosma Shalizi

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Graduate Program in Literature
in the Graduate School of Duke University
2013

Abstract

Demography of Literary Form: Probabilistic Models for Literary History

by

Allen Beye Riddell

Graduate Program in Literature
Duke University

Date: _____

Approved:

_____
Katherine Hayles, Supervisor

_____
Michael Hardt

_____
Timothy Lenoir

_____
Cosma Shalizi

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Literature in the Graduate School of Duke University
2013

Abstract

Digitization of library collections has made millions of books, newspapers, and academic journal articles accessible. These resources present an opportunity for historians interested in identifying patterns in cultural production that emerge over the space of decades or even centuries. For example, considerable interest has been expressed in studying the emergence, decline, and transmission across national and linguistic boundaries of literary form in the tens of thousands of novels published in Europe in the eighteenth and nineteenth centuries. Navigating such a large collection of texts, however, requires the use of quantitative methods rarely used in literary studies. Single, direct reading of even a thousand texts exceeds the time and resources available to most historians.

This dissertation demonstrates the application of probabilistic model of texts in the study of literary history. The major finding of the dissertation is that regularities previously identified by literary historians can be captured by probabilistic models. Following the first chapter, "How to Read 22,198 Journal Articles: Studying the History of German Studies Using Topic Models," which introduces representations of texts used in the dissertation, chapter 3, "Inferring Novelistic Genre in the English Novel, 1800-1836," and chapter 4, "Networks of Literary Production," illustrate the contribution probabilistic models of novelistic production are positioned to make to long-standing questions in literary history. Both chapters are concerned with the detection and description of empirical regularities in surviving nineteenth-century

English novels, such as the recurrence of novelistic genres—e.g., gothic, silver fork, and national tale novels. Chapter 3 makes use of a corpus that includes a random sample of novels published in the British Isles between 1800 and 1836. The use of a random sample and of probabilistic methods, both uncommon in literary studies, serves to develop new conceptual resources for future work in literary history and the sociology of literature.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the members of my committee, in particular N. K. Hayles, who provided encouragement and valuable criticism over the years and at every stage of the dissertation. Thanks to Timothy Lenoir, Michael Hardt, and Cosma Shalizi for supporting an unorthodox project. Thanks to Sara Seten Berghausen for her assistance during a prolonged search for surviving copies of nineteenth-century British novels. Thanks to Franco Moretti and Ryan Heuser for their invitation to the Stanford Literary Lab during the early stages of this research. I would also like to thank Matt Erlin for his encouragement over several years.

1

Introduction: Sociology of Literature, a Path Not Taken

Quantitative work in literary history has few precedents. While research using computers and statistics to study collections of works by a single author or a small number of authors emerged as early as the 1950s, research considering a wide range of texts from diverse authors is rare. Franco Moretti has a claim to be among the first to think seriously about what computational and quantitative methods might offer to the study of literary history. In 2003 Moretti published the first part of *Graphs, Maps, Trees* in *New Left Review* in which he proposes studying literary history with methods borrowed from fields unfamiliar to literary studies: quantitative history, geography, and evolutionary theory (Moretti 2005, 1–2). In "Graphs," Moretti proposes a literary historical project that would concern itself with the bulk of novelistic production, the tens of thousands of novels published in the eighteenth and nineteenth centuries which have been ignored by researchers and university curricula. "Graphs" outlines an ambitious project for literary history and comparative literature, in which those fields figure as potential collaborators in the social, economic, and political history of the eighteenth and nineteenth century. For instance,

literary historians might track the global traffic in cultural works and literary forms over decades in a manner not dissimilar to economic historians studying the development of global trade flows.

At the same time as "Graphs" appeared in the pages of *New Left Review*, the company Google was finalizing agreements with libraries at Harvard, University of Michigan, New York Public Library, Oxford, and Stanford that would permit the scanning of collections encompassing fifteen million volumes. A similar, more public-spirited and less commercially-oriented effort would begin shortly thereafter in association with the Internet Archive. Scanning of library collections has continued ever since. For example, during the last three years, Duke University, like many universities across the globe, has had a scanner designed by the Internet Archive, a "Scribe," in continuous operation.

As the years passed, Moretti's proposal for a research project that would have literary studies and literary history consider the tens of thousands of novels published in the nineteenth century began to look more and more feasible. Nineteenth century British novels in particular were precisely what library digitization efforts were making available: they were out of copyright and located—by virtue of being written in English—in libraries among the first to digitize their collections.

1.1 Demography of Literary Form

The word "demography" in the title is intended to reference previous work concerned with the statistical study of surviving written and printed texts, such as research on survival patterns of medieval manuscripts and regularities among copies of *The Canterbury Tales* (Barbrook et al. 1998; Cisne 2005; Howe et al. 2001).[1] The word is also intended to make explicit an interest in bringing together research in literary history

1. The title also references the exchange between Franco Moretti and Cosma Shalizi that serves as a starting point for chapters 3 and 4 (Shalizi 2011, 130).

and sociology. In literary history research interested in the potential contribution of rigorous quantitative methods "has been a path not taken"—in the words of James F. English (English 2010, xiv).[2]

As examples of recent research that has failed to attract followers, English references Janice Radway's *Reading the Romance* and Moretti's *Graphs, Maps, Trees.* As for why these efforts have failed to attract followers in literary studies—despite occasioning considerable and recurrent discussion—English suggests that the "great divide" between literature and sociology is simply too formidable; the methodological departure implied by, for example, Moretti's "distant reading," is too radical even for the present interdisciplinary moment:[3]

> Academic disciplines ... are relational entities; they must define themselves by what they are not. And what literary studies is not is a "counting" discipline. This negative relation to numbers is traditional—foundational, even—and it has not been challenged by the rise of interdisciplinarity (xii).

Institutional barriers notwithstanding, the arguments that one can offer in favor of incorporating new methods into the practice of literary-historical research are persuasive. Even as material barriers to research on surviving literary works are diminished through library digitization, Moretti's observation remains valid: the vast majority of literary production in the nineteenth century is ignored in research

2. Sociology of literature as practiced from within sociology does not appear to be marginalized. For example, Larry Isaac's "Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem novel" was awarded the Clifford Geertz Prize for best article in 2010 by the Section on the Sociology of Culture of the American Sociological Association (announcement available at `http://www.ibiblio.org/culture/?q=node/46`).

3. The interest in quantitative methods in historical research in the United States has a different history. Social history, characterized by, among other things, an interest in statistics and methods such as regression analysis, thrived in the 1960s but was later displaced by cultural history and a suspicion (in some cases well-founded) of quantitative methods (Eley 2005; Sewell Jr. 2005; Haskell 1975).

and in curricula without much justification and particularly so in the United States—"the country of close reading" (Moretti 2000a, 57). Expanding literary history to study the tens of thousands of surviving literary works requires quantitative methods. Even ten thousand novels, a small fraction of novelistic production in Europe in the nineteenth century, is too much for an individual researcher to read. (I review Moretti's justifications for not ignoring more than "nine tenths" of literary production in chapter 3.)

1.2 Colletti, Della Volpe, and Falsifiability

There are more general arguments for methodological pluralism and, in particular, for entertaining the use of statistical inference in literary history and, more generally, in the interpretive social sciences and humanities. Explaining his choice of quantitative history, geography, and evolutionary theory as fields literary history might learn from, Moretti writes,

> The distant reason for these choices lies in my Marxist formation, which was profoundly influenced by Galvano Della Volpe, and entailed therefore (in principle, if not always in practice) a great respect for the scientific spirit. And so, while recent literary theory was turning for inspiration towards French and German metaphysics, I kept thinking that there was actually much more to be learned from the natural and social sciences (Moretti 2005, 2).

This work shares the sentiment that literary history stands to learn something from greater interaction with the methods of the natural and social sciences. That literary studies in the United States has not had much amicable contact with quantitative methods in past decades is not a claim that need search hard for evidence.[4]

4. For example, according to JSTOR, between 1970 and 2000 the phrase "random sample" occurs in only one article (out of 2,566) published in *PMLA* (Proceedings of the Modern Language

The desirability of greater interaction or even modest methodological bilingualism is by no means self-evident; Moretti's suggestion has met with considerable resistance.[5] Jane Gallop expresses the concern that neglecting close reading and borrowing methods of other disciplines (in particular, a turn towards historical research) risks compromising the distinctive contribution (and continued existence) of literary studies (Gallop 2007, 184). Gayatri Chakravorty Spivak insists that the purview of the humanities is precisely the unquantifiable, the "singular and unverifiable" (Spivak and Caruth 2010).

I claim that there is at least one justification for greater interaction that deserves broad support. Any intellectual community benefits from occasional engagement with different perspectives, methods, and materials. Such engagement pushes against the tendency of groups to move "towards self-affirming structures of ideas" and the "risk of stagnation from taken-for-granted assumptions and habitual practices" (Smith 2006, 124). Whether or not these tendencies represent a real rather than an imagined danger for those in the humanities in the United States, a body of theoretical and empirical work testifies to the benefits of the presence of diverse viewpoints in a variety of situations (see, for example, Burt (2004)). The primary contribution of the subsequent chapters is the exploration of the theoretical and practical benefits of using methods drawn from outside the traditional disciplinary—and interdisciplinary—boundaries of the humanities.

Pleas for methodological pluralism are of course not new. C. P. Snow in his 1959 Rede Lecture, "The Two Cultures and the Scientific Revolution," lamented the divide between "the literary intellectuals" and natural scientists (Snow 1993). In his view, mutual incomprehension hampered work on shared goals, such as the

Association). By contrast, it occurs in 332 articles (out of 7,652) in the *American Journal of Sociology*.

5. Bérubé (2011) offers a general discussion of (the aftermath of) the "science wars" of the 1990s.

alleviation of poverty (an earlier title of his lecture was "The Rich and the Poor").[6] More recently, Bruno Latour has signaled concern about an intellectual monoculture in the humanities and that critique, as a defining method of the community, has "run out of steam." Latour calls for a "renewed empiricism," encouraging scholars in the humanities to "add reality" to inquiry often divorced from widely-shared concerns (Latour 2004, 232).[7] Latour's recent work may give some indication of what this "renewed empiricism" entails; in recent articles he considers sympathetically the contribution of network analysis (Latour 2010; Latour et al. 2012).

Moretti contextualizes his inquiry with references to his formation and to the influences of Galvano Della Volpe and Lucio Colletti (Moretti 2006, 71). Briefly discussing these figures serves as a reminder, helpful given the preceding discussion, of a diversity of perspectives on the relationship between the human and the experimental natural sciences.[8] Colletti and Della Volpe were known in the 1950s for their belief, not uncommon at the time, in a scientific Marxism according to which Marx figures in the science of society, much as Galileo figures in physics (so the slogan went). Colletti, in a significant 1974 interview with Perry Anderson in *New Left Review*, distances himself from this earlier position, but insists on the importance of putting hypotheses in contact with observation: ideas "must be checked, verified or falsified, by confronting them with data of observation, which are different in nature from any logical notion" (Colletti 1974, 12). Without committing to any specific theory of knowledge, Colletti refuses the idea that the methods of inquiry and the sort of knowledge sought in the human and social sciences share nothing with those

6. Collini reminds us that Snow thought economic development of other societies as requiring only the importation of a sufficient number of scientists and engineers from industrialized countries (Collini 1993, lxviii).

7. Latour identifies the work of Donna Haraway as an important influence in this context.

8. A number of commentators have also remarked on the contemporary resonance of Colletti's views in the wake of the 2008 financial crisis, in particular his insistence on the need for Marxism to return to works such as of Rudolph Hilferding's *Finance Capital* (Redhead 2010; Mann 2009).

in the natural sciences.

An interest in falsifiable hypotheses appears often in Moretti's work (Moretti 1982, 1988, 2000a, 2005). There are a variety of reasons why falsifiability might be seen as desirable (Godfrey-Smith 2007). The suggestion that literary historians might seek to make testable claims does not imply an endorsement of naive realism or covering-law positivism. It is worth recalling in this context that the Popperian desire for falsifiable arguments does not carry with it any firm claim about whether adhering to such a practice will get one closer to truth—a corroborated claim is one that has survived an attempt to falsify it (including extremely weak attempts) (Godfrey-Smith 2003).[9]

The suggestion that literary history might make testable claims animates at many of the explorations present in the subsequent chapters. That there are not more examples of falsifiable claims (chapter 3 features the most) is principally due to lack of a larger sample of surviving literary production and related data, rather than a lack of interest in putting forward such claims.

1.3 Publishing History

In addition to demonstrating the productive use of unfamiliar methods, this work contributes to publishing history by demonstrating the possibility of assembling a random sample of novels published in the British Isles between 1800 and 1836. That little is known about the precise contours of literary production in the nineteenth century is not, I think, widely appreciated. In 1988 John Sutherland called publishing history a "hole" at the center of literary sociology. Moretti mentions in "Graphs" that nobody has an estimate about how many novels were published in the British Isles in the nineteenth century—"twenty thousand, thirty, more, no one really knows" (4). It

9. While scientists often appeal to Popper as a model for their practice, philosophers of science more often than not will point to different habits and practices as the hallmarks of science (Godfrey-Smith 2007).

is also sobering to recall that the widely-held belief that male writers overtook women writers as authors of English novels around 1840 could not be verified until relatively recently (Tuchman 1989; Garside 2000). Tuchman, writing in 1989, describes the lack of information available at the time:[10]

> To prove that men invaded the novel, we must first establish that before 1840 at least half of all novelists were women. Many literary historians claim that well into the nineteenth century novelists were mainly women. But their evidence is not definitive because it is impossible to learn how many novels were published, let alone what proportion of them women wrote (45).

In the interest of supporting further empirical work on the history of British and European publishing, the corpus used in chapters 3 and 4 will be made publicly accessible along with supporting bibliographic "metadata" and documentation of the random sampling procedure.

1.4 Chapter Outline

What follows is a brief description of the three principal chapters in the dissertation.

Chapter 2, "How to Read 22,198 Journal Articles: Studying the History of German Studies Using Topic Models," serves as an introduction to methods used in

10. "Edging Women out" is an important example of research from the early 1980s that made use of both quantitative and qualitative evidence in literary history. Tuchman used the archives of the Macmillan publishing house, analyzing submissions and reader reports from a sample of 2,861 manuscripts between 1866 and 1917. While women wrote the majority of novels published in the early part of the 19th century, the situation had changed dramatically by the end of the century, and Tuchman uses the Macmillan Archives to investigate the change. One development in particular, the monopolization of the "high-culture novel" by men, figures prominently in Tuchman's work and the Macmillan Archives, among other sources, provide quantitative evidence that is woven into the historical and sociological investigation. For example, at Macmillan the acceptance rate of novels written by men increased between 1866 and 1917, surpassing the rate of acceptance for women (Tuchman 1989, 63). Tuchman cautions the reader to interpret the decline carefully; Macmillan accepted very few novels for publications (63). John Sutherland also raises important concerns about Macmillan's representativeness in his review of Tuchman's book (Sutherland 1989a). Similar concerns about the extent of the decline in women's share of published novels during the period are expressed in Casey (1996).

later chapters. Academic journals record the development of German Studies in the United States over the twentieth century. Reading through tens of thousands of journal articles presents a challenge. Chapter 2 considers alternative ways of reading more than twenty thousand articles published between 1928 and 2006 in *Monatshefte*, *New German Critique*, *The German Quarterly*, and *German Studies Review*. One approach, a probabilistic topic model captures major trends, including the relative decline in articles about language pedagogy and the rise of literary history and criticism.

Topic models, which also date to roughly the same period and geographic location as Moretti's "Graphs" and library digitization, serve as a useful tool for summarizing collections of texts. They go some way toward overcoming an obstacle that is pervasive in text analysis, the problem of polysemy. That the same word can mean different things in different contexts poses a problem for most methods of text comparison and summarization. Topic models provide a principled way of dealing with the challenge and, in practice, have proved reliable guides to the contents of large corpora.

Chapter 3, "Inferring Novelistic Genre in the English Novel, 1800-1836," and chapter 4, "Networks of Literary Production," are both concerned with the contribution probabilistic models of novelistic production stand to make to long-standing questions in literary history. Both are concerned with the detection and description of empirical regularities—such as the recurrence of novelistic (sub)genres including the gothic, silver fork, and national tale novels—in the nineteenth-century English novel. These chapters make use of a corpus that includes a random sample of novels published in the British Isles between 1800 and 1836. To the best of my knowledge, this is first time that a random sample of nineteenth-century novels has been assembled and used in literary-historical research.

Chapter 3 and chapter 4 are both concerned with modeling the history of the

novel—but they consider distinct strategies. They ask, if there are indeed empirical regularities in novelistic production that would position literary history to contribute to social, economic, and political history, how might one go about detecting these regularities and checking that the patterns are not illusions or statistical noise. Chapter 3, "Inferring Novelistic Genre" is the chapter most directly connected to Moretti's "Graphs." Novelistic genres are something Moretti discusses in "Graphs" and they are a well-studied feature of novelistic production. They serve as a means of testing whether or not the tools of quantitative text analysis can be brought into conversation with existing research in literary history. I find that they can be, and show that the characterizations of a collection of novels arrived at by a topic model are indeed similar to those arrived at by literary historians. I also show that thinking about modeling genres with probabilistic models prompts useful deliberations about what is meant by "novelistic genre" and "morphological similarity."

Chapter 4 branches out in a different direction and considers the consequences of putting relationships among authors and publishers at the center of research on novelistic genres and the history of the novel. This chapter also deals with a question left unaddressed in chapter 3, namely why it is that we observe novels by different authors using very similar vocabulary, such as is certainly the case with novelistic genres in the nineteenth century (and today in the twenty-first). Briefly, the hypothesis put forward is that many novels are similar because writers are reading similar works and "copying" what they find, where copying includes the borrowing of words and phrases from other novels. I lack the comprehensive corpus of novelistic production to test this theory, so I lay out a model that I believe serves as a basis for future work.

These final two chapters make no claim to have discovered a superior framework for the analysis of literary history at scale. Rather they show that while (probabilistic) abstraction may "reduce" the complexity of cultural artifacts, this is not the

same as reductionism. These chapters identify a number of cases where quantitative methods are profoundly useful and in a position to concretely change how cultural and literary historians work.

2

How to Read 22,198 Journal Articles: Studying the History of German Studies Using Topic Models

In the past decade, research libraries have digitized their holdings, making a vast collection of scanned books, newspapers, and other texts conveniently accessible. While these collections present obvious opportunities for historical research, the task of exploring the contents of thousands of texts presents a challenge. This chapter provides a practical introduction to a family of methods, often called topic models, that can be used to explore very large collections of texts. Researchers using these methods may be found not only in computer science, statistics, and computational linguistics, but also increasingly in the human and social sciences, in fields such as women's history, political science, history of science, and classical studies (Grimmer 2010; Block and Newman 2011; Hall 2008; Mimno 2012a). This introduction uses a topic model to explore a particular corpus, a collection of 22,198 journal articles and book reviews from four US-based German Studies journals: *The German Quarterly*, *New German Critique*, *German Studies Review*, and *Monatshefte*. As this is the first time this corpus has been explored using quantitative methods, this introduction

also presents a new perspective on the disciplinary history of German Studies.

This chapter has three parts. First, I review existing methods that researchers, often historians, have used to explore very large collections of texts. Then I introduce a topic model—a probabilistic model of words appearing in a collection of texts—as an alternative way of reading a corpus. I aim to show that a topic model of the German Studies journals reveals disciplinary trends that would be immensely time-consuming to document otherwise. Finally, I discuss prospects for using topic models in nineteenth-century research generally and in intellectual history specifically.

2.1 Existing Approaches: Direct and Collaborative Reading

The early 2000s witnessed the emergence of several library digitization efforts (Open Content Alliance and Google Books, to name two examples). During this period, observers asked what historians might plausibly do with such vast digital collections. Gregory Crane, a classicist and editor-in-chief of the successful Perseus Digital Library, put the question succinctly in 2006, asking, "What you do with a million books?" (Crane 2006). As a practical matter, however, Crane might as well have asked what to do with a thousand books, since carefully reading a thousand volumes already involves more time than many researchers are willing to devote to a single project.

For the sake of brevity, I will refer to any collection of texts as a "very large collection" if it contains more texts than a single researcher would be expected to digest in a year's worth of dedicated reading. 22,198 journal articles would count as a very large collection, as would the proceedings of the British Parliament in the nineteenth century or all articles published in an established regional newspaper (Mimno and Blei 2011; Nelson 2011). What options are available to researchers interested in such collections? If they look to past efforts, they have two strategies available: "direct reading" and "collaborative reading."

Direct reading is familiar. Regardless of the size of the corpus, researchers may invest the required time to read and digest its contents. There are many examples of scholars reading through enormous collections of texts in the course of their research. The American historian Laurel Thatcher Ulrich spent years reading and re-reading the nearly 10,000 diary entries of Martha Ballard, a midwife in Maine around 1800 (Ulrich 1990). Examples of studies requiring extensive reading from German cultural and intellectual history include Fritz Ringer's *The Decline of the German Mandarins,* which involved his reading a significant fraction of all books written between 1890 and 1933 by German full professors in the human sciences, and Kirsten Belgum's *Popularizing the Nation,* which took among its objects the ca. 2,500 issues of the weekly magazine *Die Gartenlaube* printed between 1853 and 1900 (Ringer 1969; Belgum 1998). Familiarity with a very large collection may also be gained over the course of years of research and teaching. There are many scholars of the nineteenth-century European novel—such as Katie Trumpener or John Sutherland—who, I suspect, have read a significant fraction of all European novels published in the eighteenth and nineteenth centuries.

A second option, collaborative reading, involves dividing up the task of reading among a number of participants. This approach brings with it the challenge of coordinating among readers. There are many examples of this approach (Simon and Rabkin 2008; Isaac 2009; Moretti 2005; Unsworth 2006). One effort that managed the problem of coordination particularly well is the Genre Evolution Project, led by Carl Simon and Eric Rabkin at the University of Michigan (Rabkin 2004; Simon and Rabkin 2008). Simon and Rabkin gathered a team of faculty, graduate students, and undergraduates together to read the ca. 2,000 short stories published in major US science fiction magazines between 1929 and 1999. The team was interested in studying how the science fiction genre changed over time and in testing existing claims about the genre against the evidence provided by the short stories corpus.

No participant read all the stories, but participants did overlap in their reading assignments. To coordinate their efforts the team focused on gathering information about a range of discrete "features," including the genders and ages of authors as well as characteristics of the narratives, such as whether a story was set in the past or whether uses of technology led to a "bad outcome." As each story was read by at least two participants, any reader's judgment could be checked against the readings of others. In this fashion, cases of disagreement could be identified and discussed. In the social sciences, this kind of checking is known as assessing inter-rater reliability.

Another example of collaborative reading is Larry Isaac's study of the "labor problem novel" in nineteenth- and early twentieth-century American fiction (Isaac 2009). Isaac considers a novel a labor problem novel if it contains one of four specific representations of labor union activity (typically, a labor strike). The time frame for his study covers nearly fifty years, 1870-1918. Since thousands of novels were published in the United States during this period, reading through all of them for mention of a strike would have been an epic undertaking. Instead, Isaac made use of existing studies and bibliographies of novels from the period and divided up the task of reading candidate labor problem novels between himself and graduate students. His team eventually arrived at a list of around 500 novels fitting the definition.

Both direct reading and collaborative reading may be combined with random sampling. If researchers are interested in investigating trends in book publishing in France between 1800 and 1900, and they happen to have a list of publications from the period, they may take a random sample and work with that corpus. If the sample is random and sufficiently large, the researchers may be confident that significant trends in the larger body of books will be identifiable in the smaller sample.

My description of these two approaches, direct reading and collaborative reading, is intended not only as a contrast with the computational and probabilistic methods that will be introduced shortly. It is also a reminder that there are many ways of

exploring a very large corpus. Researchers should not be intimidated by quantity. Even a million books could be studied by gathering a large random sample and using collaborative reading.

2.2 Machine Reading: Latent Dirichlet Allocation and Topic Models

Other ways of reading a very large collection of texts exist. A range of alternative approaches might be labeled, following N. K. Hayles, "machine reading" (Hayles 2012, 55-80). In this section, I will introduce one of these alternatives, known informally as a topic model.

Readers need an object, and machine readers are no different. The corpus used here consists of 22,198 "articles" published between 1928 and 2006 from the following four US-based German Studies journals (book reviews and editorial announcements are included):

1. *Monatshefte*, published since 1899,
2. *The German Quarterly*, published since 1928,
3. *New German Critique*, published since 1974,
4. and *German Studies Review*, which first appeared in 1978.[1]

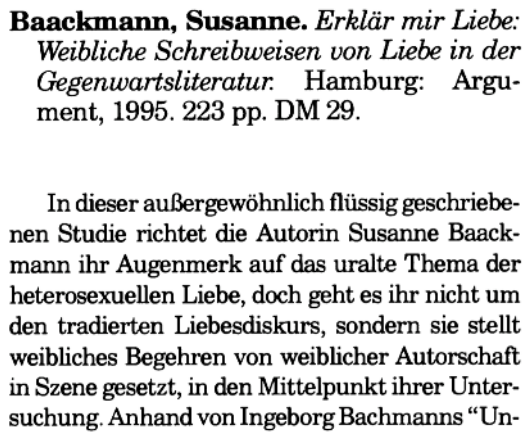
Machine readable text versions of all the articles were gathered using JSTOR's Data for Research service (DFR), which is open to the public. JSTOR is a US-based

1. The original size of the corpus provided by JSTOR was 26,104 documents. From this initial corpus, I removed articles flagged by JSTOR as "misc," typically front matter and advertisements, as well as documents having fewer than 200 words. This yielded the corpus of 22,198 . To facilitate computation, rare words (those occurring in fewer than ten documents) were removed along with extremely frequent words in German and English (so-called "stop words") and words with only one or two characters. The number of words remaining was 15,680,621 , of these 74,158 were unique words. *Monatshefte* changed its name three times between 1899 and 1946. While referred to simply as *Monatshefte* in the United States, its full title since 1946 has been *Monatshefte für deutschsprachige Literatur und Kultur*.

online repository for academic journals. These four journals are the most prominent journals dedicated to German Studies available on JSTOR.

It is worth discussing the format JSTOR uses to make these articles available. Not only are there important limitations that must be mentioned, but the format itself provides an entrée to the history and basic concepts of computational linguistics. As a preliminary step, JSTOR uses optical character recognition (OCR) to turn page scans into machine-readable text. While this is a remarkably accurate process in the sense that nearly all printed words are recognizable in the machine-readable version, OCR is not a neutral process. Lost in the procedure is information about page layout, typography, paper color, and so forth. This process is best illustrated with an example. Figure 2.1 shows a page scan of a book review, chosen at random from the corpus. The review, written by Karin Herrmann and published in 1997 in *The German Quarterly*, discusses Susanne Baackman's book *Erklär mir Liebe*. OCR stores this text in a computer file, a text document. In this case, the first line in the text document corresponding to image in figure 2.1 reads "Baackmann, Susanne. Erkldr mir Liebe:". The error ("Erkldr" instead of "Erklär") is typical; JSTOR's OCR mangles umlauts: "ä" becomes "d," "ü" becomes "ii," and so forth. In most cases, such errors are not a problem, since the confusion is consistent and there is, for example, no English word "fiir" for which the converted "für" might be mistaken. There are also difficulties, some intractable, in resolving end-of-line hyphenation (e.g., the final word "Baack-" of the second line of the review). In studies of large numbers of documents of reasonable length such issues of hyphenation prove only a minor inconvenience. Even though the OCR process cannot resolve a single world from the hyphenated "Baackmann" that spans two lines, the word occurs many times throughout the text without hyphenation.

After OCR, JSTOR discards word order, makes all words lowercase, and removes



Baackmann, Susanne. *Erklär mir Liebe: Weibliche Schreibweisen von Liebe in der Gegenwartsliteratur.* Hamburg: Argument, 1995. 223 pp. DM 29.

In dieser außergewöhnlich flüssig geschriebenen Studie richtet die Autorin Susanne Baackmann ihr Augenmerk auf das uralte Thema der heterosexuellen Liebe, doch geht es ihr nicht um den tradierten Liebesdiskurs, sondern sie stellt weibliches Begehren von weiblicher Autorschaft in Szene gesetzt, in den Mittelpunkt ihrer Untersuchung. Anhand von Ingeborg Bachmanns "Un-

FIGURE 2.1: Scan of the first page of a review of Susanne Baackman's *Erklär mir Liebe* by Karin Herrmann, published in *The German Quarterly* (Summer 1997).

all numbers (fig. 2.2).[2] Discarding word order means there is no way anyone can reconstruct the original review. Since all articles published after 1924 are "protected" by US copyright law, it is this feature that shields JSTOR from liability and facilitates public access to the DFR service. Having access to the full text of these articles and reviews would be preferable. It would, for example, enable researchers to correct idiosyncrasies like the mangling of umlauts. That this is not possible—that US and international law blocks the non-commercial use of the full text of journal articles from the 1950s and 1990s in historical research—is a consequence of the current international copyright regime (Boyle 2008; Lessig 2005).

It is not only copyright law that prompts JSTOR to provide articles in this format. The format is also one extremely familiar to computational linguists. It is called the bag-of-words representation or the vector space model.

2. This final step—removing all numbers—creates a special problem with this corpus. Since the Eszett, ß, is mangled by JSTOR OCR into "l3," all words containing ß are removed. Given the nature of this present inquiry—the concern for clear trends visible across many articles—this does not present a serious problem: any easily detectable trend in the corpus will be the product of many words systematically co-occurring.

```
<article id="10.2307/408237" >
  <wordcount weight="6" > baackmann </wordcount>
  <wordcount weight="1" > mir </wordcount>
  <wordcount weight="3" > liebe </wordcount>
  <wordcount weight="15" > der </wordcount>
  <wordcount weight="2" > susanne </wordcount>
  <wordcount weight="1" > weibliche </wordcount>
  <wordcount weight="1" > schreibweisen </wordcount>
  <wordcount weight="1" > ist </wordcount>
  <wordcount weight="13" > die </wordcount>
  <wordcount weight="5" > sie </wordcount>
  .
  .
  .
</article>
```

FIGURE 2.2: JSTOR XML for review of Susanne Baackman's *Erklär mir Liebe*. Lines have been reordered to enable comparison with figure 2.1.

2.2.1 Bag-of-Words and Vector Space Representations

The moniker "bag-of-words" captures what is left after discarding word order: an unordered list—or "bag"—of words.[3] A convenient way of organizing these lists is in a table of word frequencies. If I collected the bag-of-words for each book review in the 1997 issue of *The German Quarterly*, a small part of that table would be Table 2.1 (with the first line corresponding to the review of *Erklär mir Liebe*). This kind of table is easy to construct given the format used by JSTOR (fig. 2.2).

Those encountering this representation for the first time may be puzzled as to why this representation is used. To understand its origins, it is helpful to consider a smaller set of documents. Imagine for a moment that our corpus consists of the thirty-six chapters of Theodore Fontane's novel *Effi Briest* (1894). Each chapter is considered as a separate text document. If our vocabulary were limited to two

3. Formally, we might consider a bag in the context of the following three concepts: set, bag, and sequence. A set is an unordered list of elements that ignores order and duplicates, $S = \{4, 4, 5\} = \{4, 5\}$. A bag is an unordered list that takes into account repeated elements, $B = \{4, 4, 4, 5\} = \{5, 4, 4, 4\}$. A sequence considers both order and repeated elements, $Q = \{4, 4, 5\} \neq \{5, 4, 4\}$.

Table 2.1: Word frequencies for book reviews in *The German Quarterly* (Summer 1997).

	baackmann	mir	liebe	der	the	⋯
review1	6	1	3	15	0	⋯
review2	0	0	1	28	1	⋯
review3	0	0	0	6	91	⋯
review4	0	1	0	4	85	⋯
review5	0	1	0	43	2	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

solitary words: “effi” and ”innstetten”—the names of the two main characters—the resulting table of word counts would be Table 2.2. This table provides a compact, if impoverished, representation of each chapter. Each row of counts (each chapter) may also be considered alone, as pair of numbers—e.g. $(21, 7)$. These pairs may be interpreted as vectors—specifically, vectors in two-dimensional space (fig. 2.3). This is where the name vector space model originates. And just as each chapter of *Effi Briest* has a representation as a vector in a vector space, so too does each journal article in the corpus.

Table 2.2: Word frequencies for chapters of *Effi Briest*

	effi	innstetten
Chapter 1	21	7
Chapter 2	14	3
Chapter 3	32	9
Chapter 4	8	6
⋮	⋮	⋮
Chapter 27	1	28
Chapter 28	2	17
Chapter 29	1	13
⋮	⋮	⋮
Chapter 34	14	2
Chapter 35	9	12
Chapter 36	20	4

The advantages of using the vector space model are best understood in the follow-

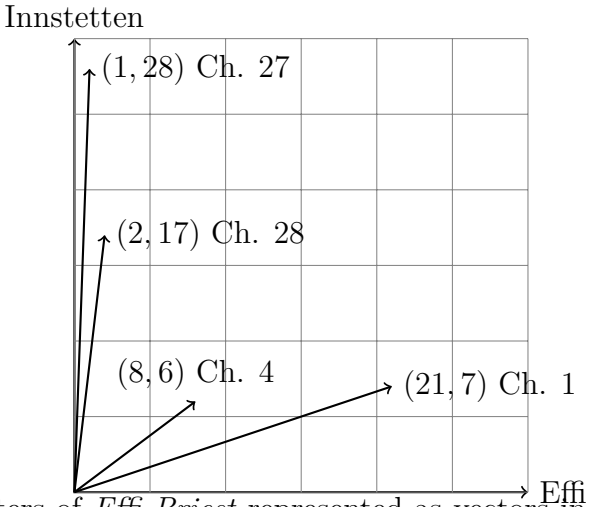


FIGURE 2.3: Chapters of *Effi Briest* represented as vectors in the two-dimensional plane.

ing context: mathematicians have spent nearly 200 years developing machinery for manipulating, comparing, and creating vectors (Crowe 1967). If we can represent our chapters or articles as vectors, we can make use of these tools. For example, we can compare the chapter vectors from *Effi Briest*. In our Effi-Innstetten space it is easy to see that the vectors reflect how much Effi and Innstetten feature in each chapter. Chapters in which Effi interacts with Innstetten point in a different direction from that of chapters in which they do not interact. In this manner we can compare two chapters without much interaction, the first chapter, before Effi marries Innstetten, and the final chapter (fig. 2.4). This notion of "pointing" in the same direction can be made precise by referring to the angle between vectors. When the angle is used to compare two vectors, it goes by the name "cosine distance" (Manning and Schüzte 1999).

Returning to the vector of the review of *Erklär mir Liebe* in *The German Quarterly*, we can use cosine distance to ask what other articles in the corpus are most similar to the review—where similar here means "having the smallest angle between the word count vectors." Dissimilar articles, those whose vectors form the largest angle with the book review's vector of word frequencies, may also be located. Table 2.3

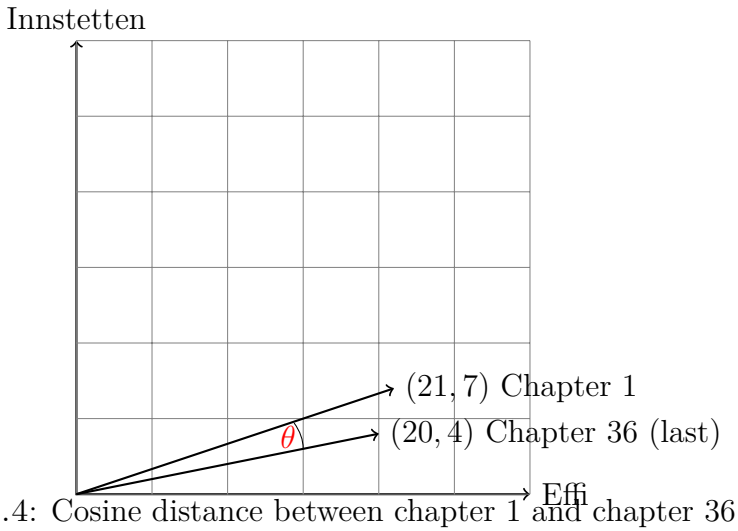


FIGURE 2.4: Cosine distance between chapter 1 and chapter 36.

lists these articles.

Like any abstraction, the vector space model obscures important aspects of texts, word order chief among them—e.g., "The child ate the fish" and "the fish ate the child" are indistinguishable. It fails spectacularly when confronted with polysemy: "Mann" in "Ein junger Mann" is counted the same as the "Mann" in "Thomas Mann." And many measures used to compare word count vectors are maddeningly opaque. For example, while it is tempting to characterize cosine distance as a measure of similarity, this similarity has no interpretation familiar to human readers. Moreover, when dealing with roughly comparable texts, experiments have shown that cosine distance and related measures are only loosely correlated with human judgments of similarity (Lee, Pincombe, and Welsh 2005).

Another objection to the use of the vector space model is that readers often do not care about individual words per se; rather, they are interested in groups of related words. For example, if we really wanted to capture how much each chapter of *Effi Briest* featured Effi, we would want to consider all the words associated with her. She is called "Effi" by her parents and by Innstetten, but she is called "gnädige Frau" by others. We would also be interested in the possessive form "Effis" along

Table 2.3: Articles similar and dissimilar to Herrmann's review of Susanne Baackmann's *Erklär mir Liebe*. Similarity measured by cosine distance.
Similar articles

- Annegret Pelz, "Karten als Lesefiguren literarischer Räume," *German Studies Review* 18 (February 1995): 115-29.
- Sigrid Kellenter, "Geertje Suhrs Märchengedichte: Grimms Heldin mündig?" *German Studies Review* 18 (October 1995): 393-418.
- Hans-Jürgen Bachorski, "Per antiffrasin: Das System der Negotionen in Heinrich Wittenwilers Ring," *Monatshefte* 80 (Winter 1988): 469-87.
- Roland Berbig, "Ein Fest in den Hütten der gastlichen Freundschaft: Überlegungen zum Verhältnis von Freundschaft und Heimat bei Hölderlin," *Monatshefte* 88 (Summer 1996): 157-75.
- Barbara Becker-Cantarino, "Lessing, 'Der Misogyne'. Sexualität und Maskerade in Lessings frühen Lustspielen," *Monatshefte* 92 (Summer 2000): 123-38.

Dissimilar articles

- William G. Meyer, "Nutley High School's Plan of Language Teaching," *The German Quarterly* 18 (November 1945): 172-73.
- Elizabeth Weitman Gelber, review of *Herrn Schmidt sein Dackel "Haidjer"* by Bruno Nelissen-Haken, *The German Quarterly* 11 (November 1938): 223.
- "Correspondence," *The German Quarterly* 9 (May 1936): 130.
- John L. Martin, "The Veteran as a Student of Modern Languages," *The German Quarterly* 20 (January 1947): 5-6.
- Walter Wadepuhl, review of *Pocket Dictionary of the German and English Languages* by K. Wichmann, *The German Quarterly* 12 (May 1939): 171.

with the inflected forms of "gnädige Frau." These are all distinct vocabulary items in the vector space model. Similarly, with our corpus of journal articles, if we were interested in identifying the proportion of articles devoted to a certain topic, such as the study of German folktales, we would be interested in a set of words, such as: "tale," "tales," "fairy," "grimm," "folk," "wilhelm," and "brothers." If we were interested the rise of feminist criticism, we would be concerned with tracking the occurrence of a cluster of words, such as "women," "woman," "male," "feminist," "gender," "patriarchy," and "social." Whether we are working with the chapters of a novel or with journal articles, it would be convenient to relax the vector space model somewhat and instead represent texts in terms of these distinctive constellations of words.

Remarkably, human readers need not specify what words belong to these clusters

of words. Given a large corpus of texts, these groups of related words can often be inferred from their patterns of occurrence alone. In a limited sense, the data—here, the corpus—can "speak for itself." Making use of a topic model is one way of achieving this feat.

2.2.2 Latent Dirichlet Allocation and Topic Models

"Topic model" is an informal label for a member of a family of probabilistic models developed over the last ten years. These models trace their roots to a model described in 2003 by David Blei, Andrew Ng, and Michael Jordan (Blei, Ng, and Jordan 2003). The authors named this model Latent Dirichlet Allocation or LDA. "Latent" refers to the model's assumption that the aforementioned clusters of words exist and are responsible in a specific sense for all the word frequencies observed in the corpus. As these groups of words are themselves hidden, their distribution in the corpus needs to be inferred. "Dirichlet" refers to the probability distribution that does this work. The distribution is named after the nineteenth-century German mathematician Peter Gustav Lejeune Dirichlet (1805–59).[4] The name "topic model" was retrospective. In practice, the model successfully finds groups of related words in a large corpus of texts, groups of words that readers felt comfortable calling topics (Blei 2012).[5] Strictly speaking, these topics are probability distributions over the unique words (vocabulary) of the corpus; those words to which the distributions assign the highest probability are those I will refer to as associated or linked with the topic. While new topic models have appeared in the intervening years, I will use LDA to model the

4. Dirichlet was a contemporary of Carl Friedrich Gauss and Carl Gustav Jacobi. Alexander von Humbolt supported his candidacy to the Prussian Academy of Sciences. Through Humbolt he met his future wife, Rebecka Mendelssohn, sister of the composer Felix Mendelssohn and granddaughter of Moses Mendelssohn. Dirichlet played a vital role in the development of modern mathematics, the modern definition of a function being credited to him (James 2002).

5. Blei's commentary is worth repeating: "Indeed calling these models 'topic models' is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA" (Blei 2012, 79).

journal article corpus.[6]

To understand how LDA works it is easiest to start with the end result.[7] LDA delivers a representation of each document in terms of topic shares or proportions. For example, assuming that thirty topics are latent in the corpus, the words in the article by Catherine Dollard, "The *alte Jungfer* as New Deviant: Representation, Sex, and the Single Woman in Imperial Germany" are associated with topics in the following proportions: 47% topic 25, 17% topic 19, and 9% topic 20 (with 27% distributed with smaller shares over the remaining twenty-seven topics) (fig. 2.5). The plurality of the words are associated with topic 25, which in turn is characterized by its assigning high probability to observing the following words: "women," "female," "woman," "male," "sexual," "feminist," "social," "gender," "family," and "mother."

Dollard, Catherine. "The *alte Jungfer* as New Deviant: Representation, Sex, and the Single Woman in Imperial Germany," *German Studies Review* 29 (Feb 2006): 107-26.

	share
Topic 25	.47
Topic 19	.17
Topic 20	.09

	top words
Topic 25	women female woman male sexual feminist social gender family
Topic 19	german political social history austrian national studies germany
Topic 20	life time people death love little story world father day left

FIGURE 2.5: Catherine Dollard's article in *German Studies Review* in terms of its prominent topics. Shares and words are based on a topic model (LDA) with thirty topics. Considered separately, each of the remaining topics contributes less than 5 percent of the words in the article.

How does LDA arrive at this representation? Should readers trust its description of articles in the corpus? The first question has a ready answer. LDA and other topic models add an interpretive layer on top of the vector space model. These models look at word frequencies through the lens of probability, permitting considerable flexibility in the interpretation of the counts. (I work through the details of a simple topic model

6. For subsequent developments, see Blei and Lafferty (2006); Teh et al. (2006); Wallach, Mimno, and McCallum (2009); Williamson et al. (2010).

7. Other introductions to LDA include Blei (2012); Blei and Lafferty (2009).

in appendix A.) Recall that when we are thinking in terms of cosine distance (which is not probabilistic), observing that two documents share a word (e.g., "weimar") counts immediately as evidence of similarity. With probability added, judgment of similarity can be postponed and made in the context of other evidence (i.e., other shared words). This flexibility is advantageous when we are dealing with the fact of polysemy in human language—a single word frequently has a diversity of meanings. For example, consider two articles that both use "weimar", one concerning Goethe (who lived in this city) and one about the Weimar Republic. Seeing the word "weimar" in both documents should not necessarily count as evidence that the two documents concern similar subjects. If the observed word frequencies justify the inference, the addition of probability to the model permits the association of the same word "weimar" with two different topics.

Should we trust that the description of documents in terms of topics corresponds at all with what our own judgments would have been had we read the 22,198 articles? The titles of journal articles provide a quick validation of the model. Recall that the topic model only uses the text of the article; words in the title are given no special status. That there is an alignment between what the topic shares suggest an article concerns and what the article title suggests provides a convenient check as to whether the model aligns with human judgments.[8]

2.2.3 Four German Studies Journals, 1928–2006

To explore the corpus of journal articles using LDA, I fixed the number of topics at 100.[9] As described above, LDA infers the distribution of the 100 topics across all the

8. The validation of topic models is an area of research in its own right. For a discussion of the issue see Chang et al. (2009).

9. The specific number of topics has no meaning itself apart from the particular probabilistic model used. In practice, however, varying the number of topics tends to vary how "finely grained" the resulting topics are. For further discussion, see Wallach, Mimno, and McCallum (2009). The `R` software package was used to model the data in conjunction with the `tm` and `topicmodels` packages; visualizations were made using `ggplot2`, see R Development Core Team (2011); Feinerer, Hornik,

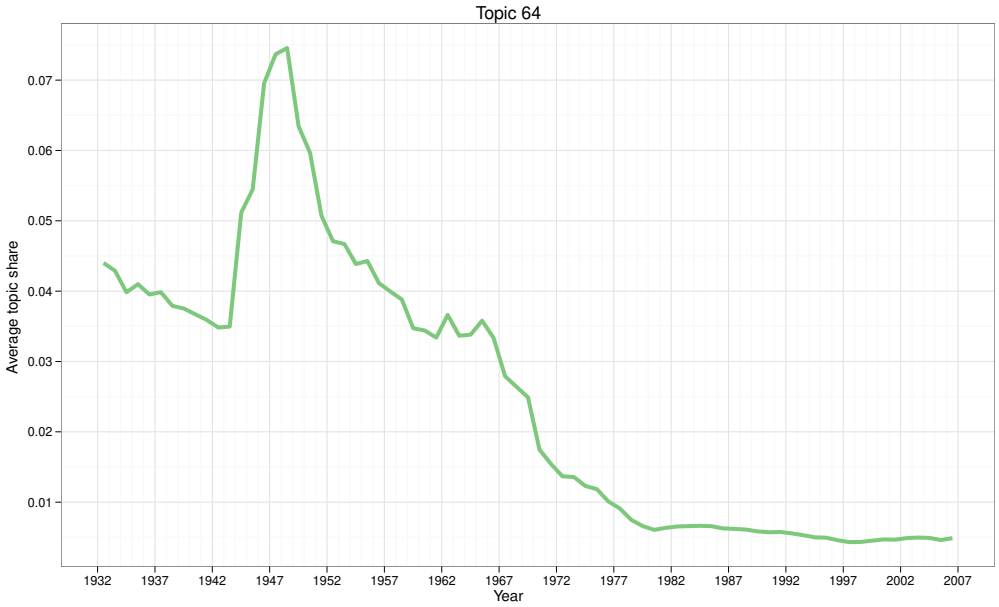
articles in the corpus as well as words characteristic of each topic. When we examine the inferred topics and plot their prevalence over the twentieth century, two dominant trends emerge. The first trend is a decline in articles on language pedagogy. Topic 64 captures this trend neatly. Its characteristic words include "students," "language," "course," and "teaching"; the titles of its associated articles confirm that the topic is linked with language pedagogy (fig. 2.6). While some of the decline in articles on language instruction is surely an artifact of the corpus (in 1968 *The German Quarterly* split off a separate journal for language instruction, *Die Unterrichspraxis* which is not included in the corpus), the decline in the share of these articles is also visible well before 1968. The second trend is the gradual rise in articles concerned with literature and literary criticism (fig. 2.7). This trend is connected with a topic characterized by words such as "literature," "literary," "writers," and "authors."

The recent history of US universities offers context for these two trends. Both are characteristic of an expansionary period, the "Golden Age," of higher education in the United States. During this period—roughly between 1945 and 1975—the number of graduate students increased nearly 900 percent. In the 1960s the number of doctorates awarded every year tripled. The Cold War is often cited among the factors contributing to the expansion of higher education generally and of graduate education in particular. In this period research displaced teaching as the defining task of the professor. Research for scholars in the humanities was associated with literary history and, eventually, literary criticism (Menand 2010, 64-66, 74-77).

In addition to the decline of articles on teaching and rise of articles on research, two other topics exhibit distinctive trends (fig. 2.8). The first topic I associate with feminist criticism. Articles connected with this topic appear much more frequently after 1975. The second topic tracks the arrival of the journal *New German Critique* in 1974. Words strongly associated with the topic include "social," "bourgeois,"

and Meyer (2008); Grün and Hornik (2011)

students language german student reading course class time teacher teaching read foreign method college material



- Eugene Jackson, "Testing for Content in an Intensive Reading Lesson," *The German Quarterly* 10 (May 1937): 142-44.
- Edwin F. Menze, "The Magnetic Tape Recorder in the Elementary German Listening Program," *The German Quarterly* 28 (November 1955): 270-274.
- H. J. Meessen, "The Aural-Oral Sections at the University of Minnesota, 1944-45," *The German Quarterly* 19 (January 1946): 36-41.
- C. R. Goedsche, "The Semi-Intensive Course at Northwestern," *The German Quarterly* 19 (January 1946): 42-47.
- D. S. Berrett et al., "Report on Special Sections in Elementary German at Indiana University," *The German Quarterly* 19 (January 1946): 18-28.

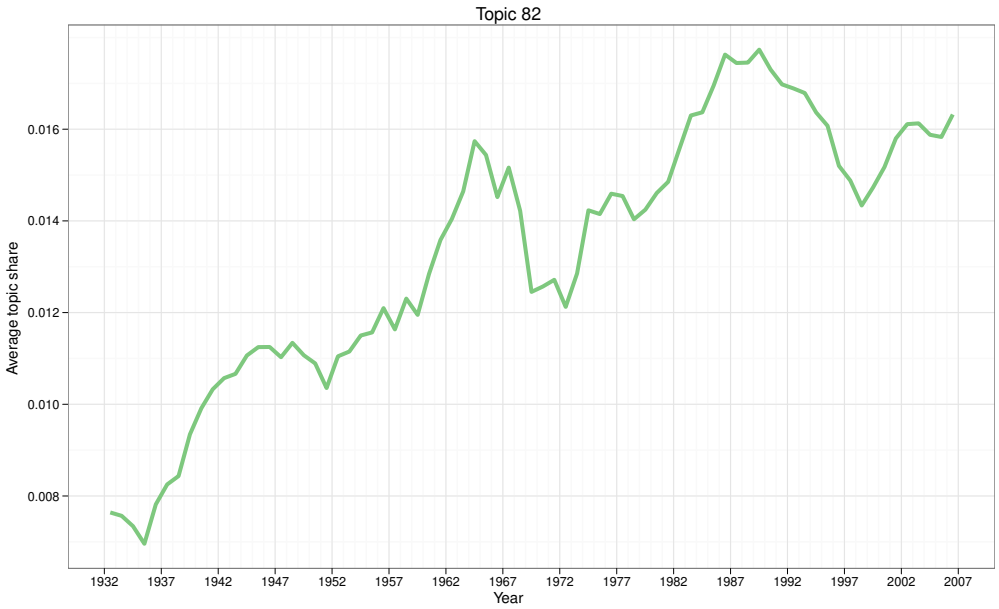
FIGURE 2.6: Topic 64 characteristic words, five-year moving average, and representative articles.

"political," "class," and "society". Herbert Marcuse's "The Failure of the New Left" numbers among the articles most strongly associated with the topic. None of the words comes as a surprise to those familiar with the journal. Its publisher describes the journal as having "played a significant role in introducing US readers to Frankfurt School thinkers . . ."[10]

All of the topics mentioned so far appear in different proportions in the corpus. Figure 2.9 shows the frequency of several topics over time on the same scale. Recall

10. This description comes from the journal's page on its publisher's website (`http://www.dukeupress.edu/Catalog/ViewProduct.php?viewby=journal&productid=45622`).

literature literary german writers authors century writer writing author period book contemporary texts novels

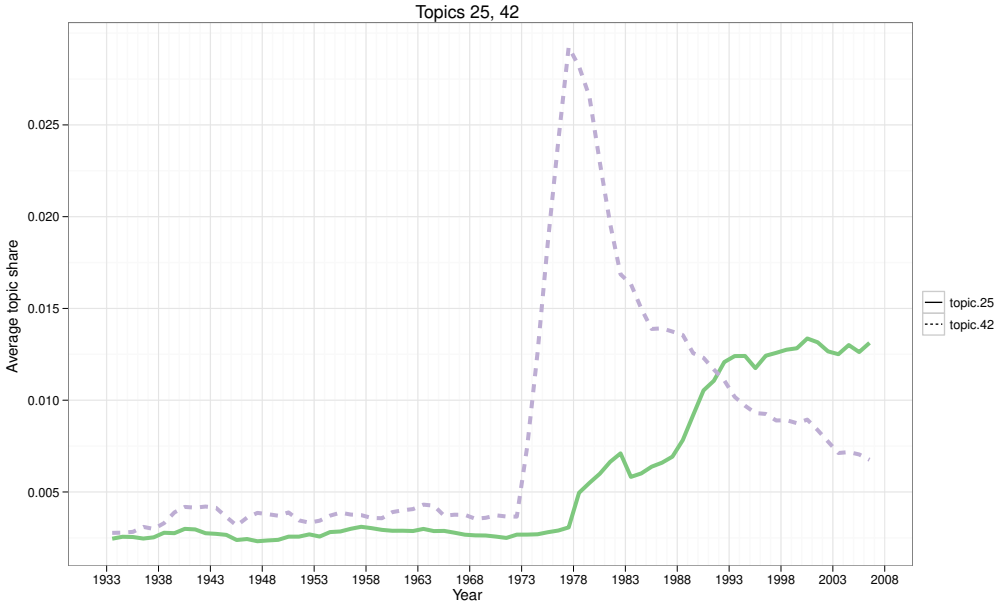


- Leland R. Phelps, review of *The Emergence of German as a Literary Language* by Eric A. Blackall, *Monatshefte* 52 (April-May 1960): 213-14.
- Andreas Kiryakakis, review of *Dictionary of Literary Biography: Volume 66: German Fiction Writers, 1885-1913 Part I: A-L* by James Hardin, *German Studies Review* 13 (May 1990): 331-32.
- Marianne Henn, review of *Benedikte Naubert (1756-1819) and Her Relations to English Culture* by Hilary Brown, *The German Quarterly* 79 (Fall 2006): 532-33.
- Stephen Brockmann, review of *German Literature of the 1990s and Beyond: Normalization and the Berlin Republic* by Stuart Taberner, *Monatshefte* 98 (Summer 2006): 318-19.
- Willa Schmidt, review of *German Fiction Writers, 1885-1913* by James Hardin *Monatshefte* 85 (Spring 1993): 99-101.

FIGURE 2.7: Topic 82 characteristic words, five-year moving average, and representative articles.

that what is being counted on the vertical axis is the average topic share among all articles in a given year (or the average proportion of all words in a given year associated with a given topic). If we accept for a moment the analogy between subject matter and topic, it would mean that a year with ten articles published and a 0.1 average share for the topic associated with language pedagogy might have two articles with half their words associated with the pedagogy topic. Or it might be the case that for all ten articles, one tenth of their words were associated with the pedagogy topic. In either case, the average topic share is 0.1. It is also worth

Topic 25: women female woman male feminist gender sexual feminine social role patriarchal movement sex roles masculine
Topic 42: social bourgeois class political critique society theory historical capitalist production marxist marx revolutionary capitalism economic



Topic 25

- Elizabeth Heineman, "Gender Identity in the Wandervogel Movement," *German Studies Review* 12 (May 1989): 249-70.
- Agatha Schwartz, "Austrian Fin-de-Siècle Gender Heteroglossia: The Dialogism of Misogyny, Feminism, and Viriphobia," *German Studies Review* 28 (May 2005): 347-66.
- Maria Dobozy, "Women and Family Life in Early Modern German Literature," *Monatshefte* 98 (Spring 2006): 133-35.
- Meredith Lee, "Der androgyne Mensch: 'Bild' und 'Gestalt' der Frau und des Mannes im Werk Goethes," *The German Quarterly* 71 (Spring 1998): 186-87.
- Ursula Mahlendorf, "Frauen und Gewalt. Interdisziplinäre Untersuchungen zu geschlechtsgebundener Gewalt in Theorie und Praxis," *Monatshefte* 98 (Spring 2006): 141-43.

Topic 42

- Karl Korsch, "The Crisis of Marxism," *New German Critique*, no. 3 (Autumn 1974): 187-207.
- Rainer Paris, "Class Structure and Legitimatory Public Sphere: A Hypothesis on the Continued Existence of Class Relationships and the Problem of Legitimation in Transitional Societies," *New German Critique*, no. 5 (Spring 1975): 149-57.
- Herbert Marcuse, "The Failure of the New Left?" *New German Critique*, no. 18 (Autumn 1979): 3-11.
- Paul Piccone, "Korsch in Spain," review of *Karl Korsch o el Nacimiento de una Nueva Epoca*, ed. Eduardo Subirats, *New German Critique*, no. 6 (Autumn 1975): 148-63.
- Paul Piccone, "From Tragedy to Farce: The Return of Critical Theory," *New German Critique*, no. 7 (Winter 1976): 91-104.

FIGURE 2.8: Topics 25 and 42 characteristic words, five-year moving average, and representative articles.

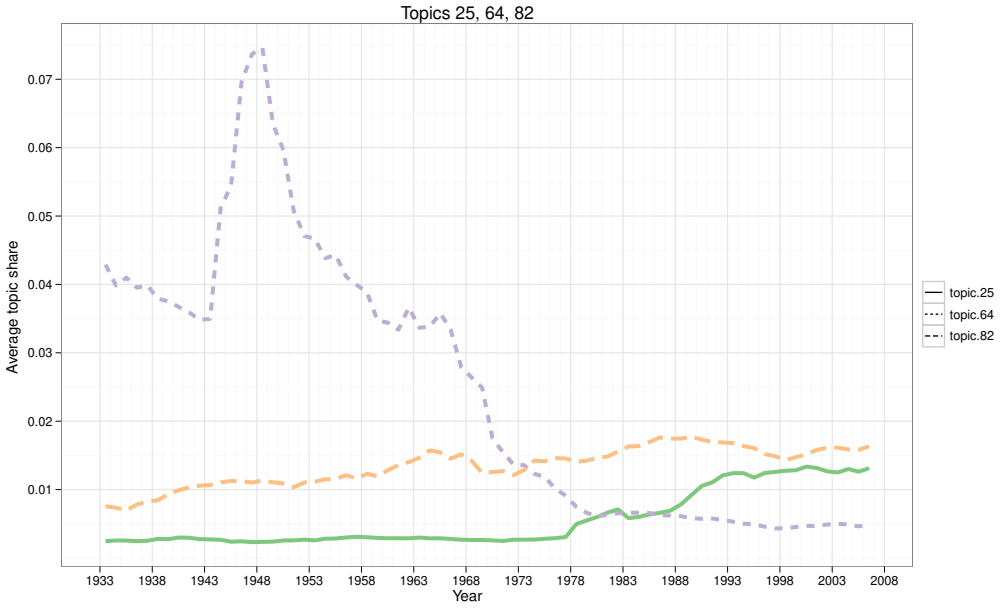


FIGURE 2.9: Comparison of topic 25 ("women …"), topic 64 ("students …"), and topic 82 ("literature …").

emphasizing that the LDA model makes use of relative rather than absolute word frequencies. That is, a 500 word review that is 20% topic 64 is treated the same, in certain important respects, as a 9,000 word article that is 20% topic 64, even though the number of words and share of space in the journal are different. Infrequent topics also bring with them their own set of concerns. As the arrival of *New German Critique* shows (fig. 2.8), the addition of a handful of articles with distinctive features leaves its mark. With topics associated with only a few articles a year, such as the "folktales" topic discussed below, selection bias becomes a concern. It is possible that some trends are not real in the sense that a rapid decline might reflect a certain kind of article migrating elsewhere—perhaps to a European history journal—rather than any decline in research on the subject in German Studies generally.

2.2.4 Long Nineteenth-Century Topics

Two topics that track specific areas of nineteenth century scholarship are worth mentioning as their trajectory over the period reveals predictable rhythms of scholarly publishing.

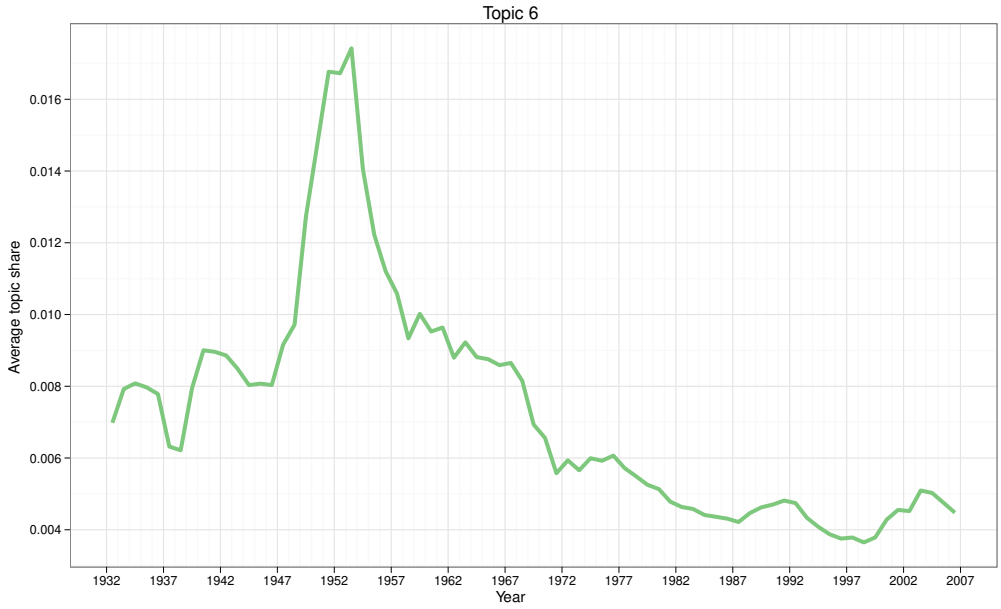
A single topic is associated with articles on the life and works of Goethe (fig. 2.10). A rapid increase in articles associated with this topic begins around 1947. This surge of articles coincides with the bicentennial of Goethe's birth (1749). *The German Quarterly*, for example, devoted the entire November 1949 issue to the bicentennial. That the topic model reflects this as well as it does offers additional validation that it is capable of capturing the gross features of the corpus.

Another topic identifies scholarship connected to folktales (fig. 2.11). With peaks around 1955 and 1990, there is a temptation to think that interest in folktales may rise and fall in a regular cycle. Yet further reflection yields a simpler explanation for the second rise: the anniversary of the births of Jacob and Wilhelm Grimm (1785 and 1786 respectively). The fluctuations in the topic's prevalence before 1970 may be due to a number of factors. For example, the arrival of new journals emphasizing scholarship on twentieth-century subjects seems likely to have contributed to the decline in the relative share of articles concerned with scholarship on folktales.

2.2.5 Topic Modeling Pitfalls

While LDA has proven an effective method for exploring very large collections of texts, it has important shortcomings, some of which are shared by other topic models. First, topics lack an interpretation apart from the probabilistic model in use. Articles may be compared in terms of their topics—one such measurement is called the Kullbeck-Leibler divergence—but this metric suffers from problems of interpretation familiar from the discussion of cosine distance. Moreover, recent work has shown that automatic measures of the fit between a topic model and a corpus (e.g.,

goethe faust goethes wilhelm werther weimar iphigenie ottilie gretchen charlotte meisters mephisto meister dichtung wahlverwandtschaften

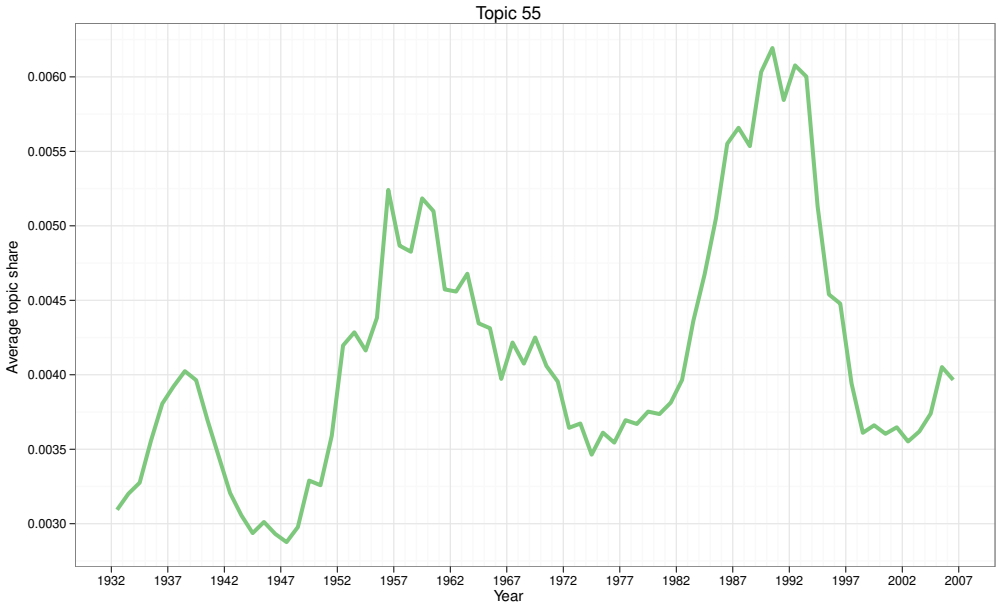


- L. M. Price, "Goethe Bibliography for 1939," *Monatshefte für deutschen Unterricht* 32, no. 2 (February 1940):83-88.
- Heinz Bluhm, "Goethe Bibliography for 1942 to 1944: German Non-Periodical Publications," *Monatshefte* 39, no. 2 (February 1947): 126-33.
- J. A. Kelly, "Goethe Bibliography for 1938," *Monatshefte für deutschen Unterricht* 31, no. 8 (December 1939): 400-06.
- Heinz Moenkemeyer, "Zum Verhältnis von Sorge, Furcht und Hoffnung in Goethes Faust," *The German Quarterly* 32, no. 2 (March 1959): 121-32.
- Hellmut Ammerlahn, "Mignons nachgetragene Vorgeschichte und das Inzestmotiv: Zur Genese und Symbolik der Goetheschen Geniusgestalten," *Monatshefte* 64, no. 1 (Spring 1972): 15-24.

FIGURE 2.10: Topic 6 characteristic words, five-year moving average, and representative articles.

held-out likelihood) do not always align with human readers' assessments of the coherence of inferred topics, suggesting a mismatch at some level between the "topics" of topic models and topics familiar to human readers (Chang et al. 2009, 288-96). Given this shortcoming, it becomes essential that those using topic models validate the description provided by a topic model by reference to something other than the topic model itself. Fortunately researchers familiar with the period, documents, and writers associated with a corpus typically have the expertise to devise appropriate checks.

tale tales fairy grimm folk wilhelm stories jacob brothers tradition grimms folklore magic story popular



- Maria M. Tatar, review of Breaking the Magic Spell: Radical Theories of Folk and Fairy Tales by Jack Zipes, *The German Quarterly* 55, no. 2 (March 1982): 231-32.
- Ruth B. Bottigheimer, review of One Fairy Story Too Many: The Brothers Grimm and Their Tales by John M. Ellis, Fairy Tales and the Art of Subversion: The Classical Genre for Children and the Process of Civilization by Jack Zipes, The Trials and Tribulations of Little Red Riding Hood: Versions of the Tale in Sociocultural Context by Jack Zipes, and Die Geschichte vom Rotkäppchen: Ursprünge, Analysen, Parodien eines Märchens by Hans Ritz, *The German Quarterly* 58, no. 1 (Winter 1985): 144-47.
- Ruth B. Bottigheimer, "Sixteenth-Century Tale Collections and Their Use in the 'Kinder- und Hausmärchen,'" *Monatshefte* 82, no. 4 (Winter 1992): 472-90.
- Ruth B. Bottigheimer, "Tale Spinners: Submerged Voices in Grimms' Fairy Tales," *New German Critique*, no. 27 (Autumn 1982): 141-50.
- Donald P. Haase, review of The Trials and Tribulations of Little Red Riding Hood: Versions of the Tale in Sociocultural Context by Jack Zipes, *Monatshefte* 78, no. 3 (Fall 1986): 385-86.

FIGURE 2.11: Topic 55 characteristic words, five-year moving average, and representative articles.

An additional complication is the fact that the number of topics in a model is arbitrary. In this chapter, I made use of a thirty topic fit (fig. 2.5) and a 100 topic fit to characterize the same corpus of journal articles. While many of the topics of the thirty topic fit resemble those of the 100 topic fit, the topics are distinct. That the number of topics and the composition of the inferred topics can vary in this manner should reinforce the idea that an individual topic has no interpretation outside of the particular model in use. Blei and his coauthors are admirably clear on this point

(Blei, Ng, and Jordan 2003, 996n1).

LDA and other topic models also make assumptions known to be incorrect (Wallach, Mimno, and McCallum 2009; Williamson et al. 2010; Blei and Lafferty 2007, 2006). For example, LDA assumes that association of words with a topic does not vary over time. In other words, LDA assumes scholars are using the same collection of words to talk about folktales in the year 1940 and the year 2000. We know this is wrong. That LDA works as well as it does is due to the fact that many words are used consistently over time. That is, regardless of the decade in which the articles were written, articles about Goethe's life will tend to use words like "Goethe" and "Faust." For other kinds of inquiry, especially those concerned with less conspicuous trends, changes in language use are a significant concern. Changes in terminology in particular—for example, if writers systematically begin using "folklore" in a context where they previously would have used "folktales"—present a potential problem for LDA. For all these reasons, the assumptions made by topic models require close and careful reading.

2.3 Prospects for Topic Models

Long nineteenth-century materials, in particular, are unusually hospitable to the use of machine reading and probabilistic models. A staggering amount of printed material survives to the present day. Moreover, these texts are all unencumbered by copyright in the United States. Contrast this with the disposition of materials published in the twentieth century. Scholars working with printed material from the twentieth century are hamstrung by copyright law, unable to share text collections freely if the collections contain works published after 1923.

For researchers in the humanities and interpretive social sciences, learning how to use and reflect critically about models such as LDA is growing easier. Leading universities such as MIT and Stanford have announced a number of freely accessible

online courses that cover probability and computational linguistics. These courses discuss the bag-of-words model and probabilistic models of text collections. One such course is taught by Andrew Ng, the third author of the original LDA paper.

This chapter has made no attempt to use topic models to investigate existing accounts of the history of German Studies. Beginning with specific hypotheses, however, often makes for compelling research. Perhaps unsurprisingly, it has been computational linguists who have pioneered using topic models to ask specific questions about the history of their own discipline (Hall, Jurafsky, and Manning 2008; Hall 2008; Sim, Smith, and Smith 2012). For example, David Hall takes up an hypothesis inspired by Thomas Kuhn's account of the historical trajectory of science as one punctuated by periodic "revolutions" in dominant methods. Hall observes that there have been widely acknowledged shifts in the prominence of certain methods within computational linguistics over the past twenty years.[11] If these methodological shifts represented a revolutionary change of "paradigm" in Kuhn's sense, then Hall anticipated that the researchers associated with "insurgent" method would have arrived recently in the field. In other words, these researchers would not be established scholars who had abandoned their prior methodologies in favor of new ones (Hall 2008, 5-6). A topic model of journal articles allowed Hall to identify significant methodological shifts in the discipline and those authors associated with the changes. This general line of inquiry—with or without the guiding Kuhnian perspective—could be adapted to any number of other disciplines, including German Studies. As this chapter has demonstrated, there are a number of changes in method and subject matter that are visible in the discipline's journals since 1928. Future research might use quantitative methods to identify the scholars associated with these shifts.

My aim in this chapter has been to show that a topic model reveals disciplinary trends that would otherwise be prohibitively time-consuming to document. Used

11. The rise of statistical machine translation is a prominent example of such a shift.

alongside direct and collaborative reading, topic models have the potential to offer new perspectives on existing materials and novel accounts of the dynamics of intellectual history.

3

Inferring Novelistic Genre in the English Novel, 1800-1836

> Because form is precisely the repeatable element of literature: what returns fundamentally unchanged over many cases and many years. This, then, is what formalism can do for literary history: teach it to smile at the colorful anecdote beloved by New Historicists …and to recognize instead the regularity of the literary field. Its patterns, its slowness (Moretti 2000b, 225).

3.1 Introduction

Gothic, epistolary, and historical novels flourished in the British Isles during the late eighteenth and early nineteenth century. The share of literary production claimed by these and other novelistic genres is considerable.[1] During its peak year, for instance, gothic novels accounted for thirty percent of all new novels published (Figure 3.1) (Lévy 1968). Literary historians have documented the rise and fall of these and other novelistic genres. Other familiar categories from the nineteenth century include

1. Following Moretti (2005), I will refer to these categories as "novelistic genres." If context makes it clear the discussion is limited to novels, the qualifier "novelistic" may be dropped. In discussions of eighteenth- and nineteenth-century literature "genre" is used in a variety of ways by different authors—e.g., to distinguish epic and tragic narratives, or among poetry, plays, and novels.

the national tale, silver fork, Bildungsroman, and Newgate novels. (Lévy 1968; Adburgham 1983; Hollingsworth 1963; Trumpener 1998). Recently, Moretti (2005) has revived interest in these categories by aggregating information about genres' periods of popularity and looking for regularities in the arrival and disappearance of genres.

While there is no consensus among literary historians on a general definition of novelistic genre, many genres were recognized by readers, writers, and publishers at the time. The clearest evidence comes from novels' (sub)titles, which often signaled a generic affiliation—e.g., *The Baron's Daughter: A Gothic Romance*, *Durston Castle; Or, the Ghost of Eleonora: A Gothic Story*, *The Wild Irish Girl: A National Tale*, and *Caledonia; Or, the Stranger in Scotland: A National Tale*. Literary historians do provide detailed descriptions of individual genres, frequently making reference to shared features or "codes" (Moretti 2005; Cohen 2002). For instance, Abrams and Greenblatt describe gothic novels as "a group of novels, set somewhere in the past, that exploit the possibilities of mystery and terror in sullen, craggy landscapes; decaying mansions with dank dungeons, secret passages, and stealthy ghosts; chilling supernatural phenomena; and often, sexual persecution of a beautiful maiden by an obsessed and haggard villain" (Abrams and Greenblatt 2000, 19). The features characteristic of a novelistic genres need not be limited to settings, or indeed to anything found in the text of a novel. A book's binding may be an important signal to readers, as it was for gothic novels, earning them the moniker "bluebooks" (Koch 2002). As is still the case today, publishers are often associated with genres (Adburgham 1983; Trumpener 1998). Narrative voice and plot structure have also been suggested as distinguishing morphology (Elson, Dames, and McKeown 2010; Allison et al. 2011).

Scholars in the humanities and social sciences have made considerable use of novelistic genres. Countless monographs and journal articles have been devoted to

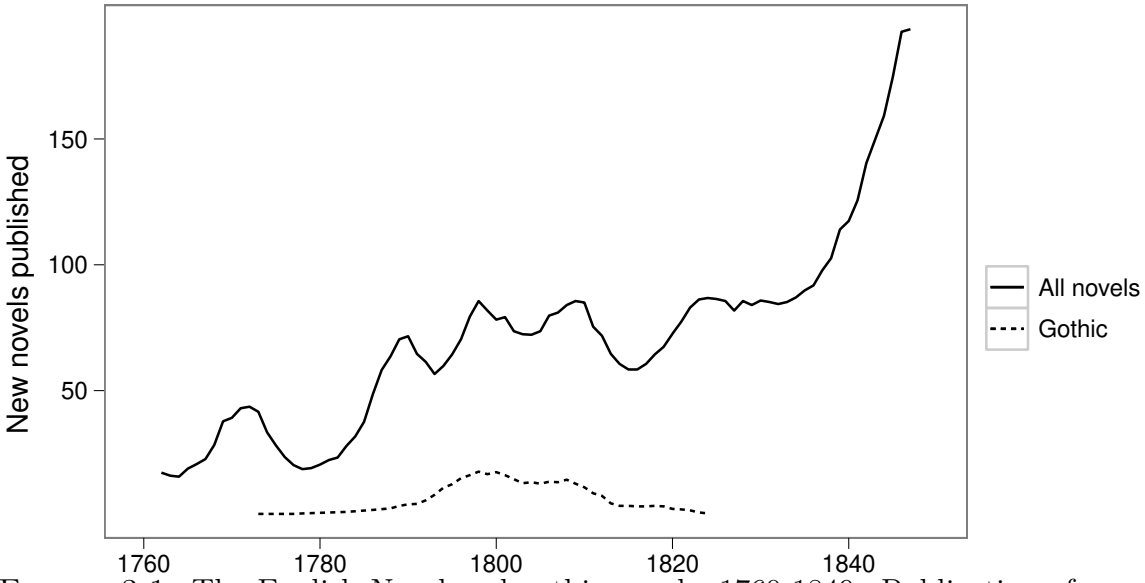


FIGURE 3.1: The English Novel and gothic novels, 1760-1849. Publication of new novels and those classified as gothic novels (five year moving average). Sources: 1770-1836 from Garside and Schöwerling (2000) and Garside et al. (2006); 1837-1849 from Block (1961); gothic novels from Lévy (1968).

the genres of the eighteenth and nineteenth-century English novel. The association of a period of popularity of a specific novelistic genre with political and social events is not uncommon. For example, the *Bildungsroman*, with its concern for youth and the process of development, has been read as symptomatic of the period following the upheaval of the French Revolution (Moretti 2000c). The silver fork novels have been connected with Regency Era social aspirations (Adburgham 1983). Novelistic genres have also been interpreted as offering a sweeping record of social relations. That is, by identifying a novel with a genre, writers situate their work in relation to existing novels, writers, codes, conventions, and institutions (Cohen 2002; Bourdieu 1988, 1996). Thus novelistic genres provide insight into the internal dynamics of novelistic production and a window into the literary field. Novelistic genres have also been used in sociology. Isaac (2009) investigated the relationship between publication of early 20th century "labor problem novels" and the historical record of labor militancy in the United States.

While the claims made using the historical record of novelistic genres have occasioned considerable debate, the salient facts about novelistic genres—including their periods of popularity and lists of associated novels, authors, and publishers—are often difficult to pin down. Researchers interested in a genre must rely on the work of one or two literary historians who have studied the genre in depth. (I will refer to these historians as "genre experts.") For a number of reasons, it would be desirable to have an alternative means of collecting (or corroborating) the vital details of a genre, rather than having to depend on the judgment of one (or two) genre experts. First, for the vast majority of novels published in the nineteenth century, no expert classification exists. When describing a genre, experts often mention only a handful of novels regarded as exemplary. In the infrequent event that a literary historian does provide a list of all novels belonging to a genre, the list is rarely exhaustive. A list may, for instance, only cover the genre's period of popularity and omit titles published after the genre ceased to be prominent.[2] Having an alternative means of finding novels with characteristics similar to those found in a set of novels already identified as members of a genre would support researchers working with novelistic genres who wish to use a sample of novels larger than that provided by an expert's list of exemplary novels.

Second, an alternative approach to identifying the genre membership of a novel would be valuable when confronted with cases where experts disagree on the membership of individual novels or independently claim a novel as a member of more than one genre—e.g., Lady Morgan's *Florence Macarthy* (silver fork and national tale), Bulwer-Lytton's *Paul Clifford* (silver fork and Newgate), and Roche's *Tradition of the Castle* (gothic and national tale). Disagreements about the genre membership of novels published at the beginning and end of a genre's period of popularity are par-

2. Adburgham (1983) stops listing silver fork novels after 1842 even though there are a small number of novels published after that date that are uncontroversial members of the genre, such as *Castles in the Air* (1847) by Catherine Gore.

ticularly problematic as they are likely to affect the periodizations of genres (Moretti 2005; Shalizi 2011). When the "disagreement" takes the form of competing classifications made by genre experts, an additional perspective would be potentially useful as a means of understanding the underlying reasons for competing classifications.[3] Finally, and most importantly, it is desirable to have an alternative to relying on the authority of one or two experts for the list of novels associated with a genre. Because expert classifications rely on background knowledge and familiarity with a broad range of novels, it is difficult for other researchers to reproduce existing classifications.

Having an alternative means of grouping novels together, particularly one that is readily reproducible, would inspire more confidence in the comprehensiveness and accuracy of any classification. The desire for reproducibility need not be understood as calling into question the work of the original expert, rather it can be seen as an interest in building on existing work. If literary historians articulate their reasons for classifying novels in a given genre such that others can follow them, it becomes easier to add to their work when new information comes to light. And new information about novels published two hundred years ago does arrive. In the past twenty years novels published in the British Isles that were thought to have been lost have been located and evidence of the existence of novels previously unknown to literary historians has come to light (Garside, Belanger, and Mandal 2001).

In this chapter I consider one such alternative means of identifying novels belonging to a novelistic genre. This method relies on a probabilistic topic model of the texts of a large collection of novels to provide a representation that associates novels with one another based on shared latent features inferred from novels' word frequencies. The representation of the novels is provided by the Hierarchical Dirichlet Process, a

3. Scholars have identified novels that they believe borrow morphology from novels in a genre but do so in a way that obscures their origins (Garside 1991). An alternative method of classification might help substantiate claims about such "cryptic" novels.

non-parametric latent feature model (Teh et al. 2006; Blei, Ng, and Jordan 2003). While word frequencies provide an impoverished representation of text, they have the advantage of being widely accepted. Readers are more likely to disagree about the characterization of a plot structure as episodic or a setting being "gothic" than they are to disagree about the words appearing in the pages of a given edition of a novel. Whereas any reader can check whether the word "trapdoor" occurs seven times in a novel, assessing whether or not a novel has an episodic plot requires considerably greater background knowledge and agreement as to what "episodic" means. I focus on three genres (gothic, silver fork, and national tale) for which extensive bibliographies exist. To the extent that classifications are uncontroversial—many novels are formulaic and derivative—the ability of a topic model to independently generate a description of a corpus that resembles existing expert classifications provides a check of the assumptions of the topic model and gives us a reason to believe that such probabilistic models can be used to study larger collections of scanned novels and locate candidate novels for inclusion in recognized genres. As has been described above, the need for such an alternative is obvious. It is likely that well over 30,000 novels were published in the British Isles in the nineteenth century alone. Bibliographies associating these novels with existing genres are frequently not available. A credible model of similarities among novels would allow researchers to corroborate received classifications and find novels that may have been missed by existing expert studies.

This chapter proceeds as follows. First, I review ideas behind existing approaches to grouping novels based on shared morphology. Second, I describe the proposed model and the corpus of gothic, silver fork, and national tale novels. Third, I will assess the success of the model by comparing its predictions about the clustering of novels to the judgments of literary historians. Finally, I will consider the practical and theoretical implications of having a statistical model that approximates expert classifications.

3.2 What are Novelistic Genres?

Literary historians have considered the challenge of inferring novelistic genre without relying on expert classifications. While studying the titles of novels published between 1740 and 1850, Moretti (2009) observed regularities in title word frequencies and phrases that correlated with novels being classified as gothic. Allison et al. (2011) took up the problem of unsupervised classification explicitly and examined whether or not patterns in selected word and punctuation frequencies might be associated with specific genres. Allison et al. studied a small collection of novels and used principal components analysis (PCA) and visual inspection of multidimensional scaling plots to characterize differences among novels.[4] Allison et al. found that novels from certain genres did separate visually whereas others did not. Allison et al. also discuss challenges facing unsupervised clustering of eighteenth- and nineteenth-century novels on the basis of word frequencies alone, noting that certain genres may be marked by narrative structure rather than lexical features. The *Bildungsroman*'s episodic structure is the example provided: "discussions with old mentors and young friends, false starts, disappointments, the discovery of one's vocation …" (15). Second, authors may switch genres (or write in several), making the lexical "signature" of the genre difficult to distinguish from authorial style. Instances of authors writing in multiple genres include one author whose works appear in the corpus considered in this chapter: Lady Sydney Morgan (née Owenson) wrote national tale novels and silver fork novels (Trumpener 1998; Adburgham 1983).

4. Allison et al. cite Cavalli-Sforza, Menozzi, and Piazza (1994) as influential in their approach to the problem and their choice to use of PCA. The traffic between population genetics, cultural evolution, and quantitative literary history deserves attention. The afterword to the widely discussed *Graphs, Maps, Trees* is written by Alberto Piazza, a coauthor of Cavalli-Sforza, Menozzi, and Piazza (1994). That work, in turn, is in conversation with the paper by Pritchard, Stephens, and Donnelly, which independently developed the mixed-membership model of allele frequencies that is essentially identical to Latent Dirichlet Allocation (Blei 2012). Novembre and Stephens (2008) is also a notable point of contact that concerns the use of PCA.

3.2.1 A Preliminary Definition

If we assume for the moment that the familiar novelistic genres—e.g., gothic, historical, and *Bildungsroman*—are not retrospective inventions of historians, then we might do worse than begin with the following provisional definition of a novelistic genre: a group of novels that share common morphology and are recognized by writers, publishers, and readers as belonging to the same category.[5] We should hesitate, however, before generalizing about the nature of the shared morphology and how it is shared. It may be nothing more than a label (the subtitle "a gothic tale" is shared by many English novels). Publishers may have affixed the label to novels quite independently of the content of the novel. The shared morphology may involve something more substantial, such as "a set of codes" that readers and writers are able to identify (Cohen 2002, 18). We have already seen an instance of the latter in the definition of the gothic novels provided by Abrams and Greenblatt. Here the codes are characteristic settings, including "sullen, craggy landscapes; decaying mansions with dank dungeons, secret passages, and stealthy ghosts; chilling supernatural phenomena."

Such a minimal definition does not, however, offer an account of why a group of novels share morphology. A complementary definition providing such an account understands novelistic genre first as a social relation. By identifying a novel with a genre, writers and publishers take a position, situating a work in relation to existing novels, writers, codes, conventions, and institutions—including existing novelistic genres.[6] The association of a work with a genre frequently had foreseeable economic consequences and, to the extent that a position was recognized by contemporaries,

5. The novel itself is usually referred to as a literary genre, so novelistic genres are on this account subgenres. Moretti also suggests the phrase "market category," which I believe is apt.

6. I am in general agreement with the account of novelistic genres given by Cohen (2002). Cohen also considers genre as a position-taking, writing that "[e]vidence for a position is primarily textual and established through analysis: proof of its existence is that the critic finds a number of texts sharing a set of codes" (p. 18).

was also bound up with assessments of symbolic prestige. Any association, however, arises in a specific context. Existing institutions and conventions condition and orient writers as they write. Pierre Bourdieu's metaphor of the "field" is helpful here. Field does not refer to a field of expertise but rather to a field in the sense used in physics (electric, gravitational, and so forth). The metaphor aids in conceptualizing how agents may act under constraints—a field impinges on objects subject to it—but still resist forces exerted upon them. In the literary field, while generic conventions and institutions structure a writer's practice, they do not determine it (Eastwood 2007). This characterization of genre and of literary creation is particularly satisfying because it moves us beyond weaknesses in received approaches to literary history. Adopting this perspective—a writer as a creative actor simultaneously conditioned by external forces—makes it difficult to consider literary creation as governed only by internal aesthetic imperatives or a creative genius. Likewise, it makes it difficult to naively read the content of a literary work as symptomatic of prevailing social conditions. Thinking about literary production in terms of novelistic genre already inclines us towards this perspective because genre is frequently both a literary category and a market category. As literary categories, genres are in constant flux insofar as they are characterized by an changing ensemble of identifying morphology. The plots, settings, devices, vocabulary, and other codes of genres are not static. For example, the codes characteristic of gothic novels written in 1795 are not those of gothic novels written in 1815. On the other hand, novelistic genres need also to be thought of as "market categories," often dominated by a small set of publishers and marketed consistently as a stable and well-defined group.

The probabilistic model considered in this chapter makes no attempt to account for or evaluate possible mechanisms by which novels might share morphology. This is a significant shortcoming of the model and one to which I return in the conclusion.

3.2.2 Why Infer Genre?

Not having to rely on the authority of a single expert for classification of novels is the primary motivation for developing an independent model of morphological similarity among novels. Even absent such motivation, having an independent means of inferring groups of related novels would be useful. For example, it would permit those interested in a given genre to identify novels that may have been overlooked by literary historians, or to which historians may not have had access. Copies of novels from the early nineteenth century continue to be located in library holdings, including novels previously thought not to have survived.[7]

Independent corroboration of classifications (or even associations posited by studies of novelistic genres) would be particularly valuable in cases where classifications are contested. Moretti (2005) identifies periods of popularity for forty-four British novelistic genres between 1740 and 1900. These are periods during which novels associated with each genre were actively circulated (fig. 3.2). Moretti notes that competing periodizations exist for several of the forty-four genres, giving one example: industrial novels. In deciding on a periodization for industrial novels (1832–1867), Moretti opts for the periodization provided by Gallagher (1985) over that of Cazamian (1973). Moretti justifies his preference on the basis of Gallagher's "more convincing morphological argument" (Moretti 2005, 18n8). One obvious source of disagreement about a periodization is a disagreement about the inclusion (that is, classification) of individual novels. If one historian includes novels published before 1832 in their bibliography of a genre and another historian does not include those novels, then the two experts are likely to disagree about the genre's period of popularity.[8] An alternate method for identifying novels exhibiting similar morphol-

7. See, for example, section "D: Titles Previously not Located for Which Holding Libraries Have Subsequently Been Discovered" in (Garside, Belanger, and Mandal 2001, 16).

8. Other discordant periodizations appear in Moretti (2005). For example, Adburgham (1983)

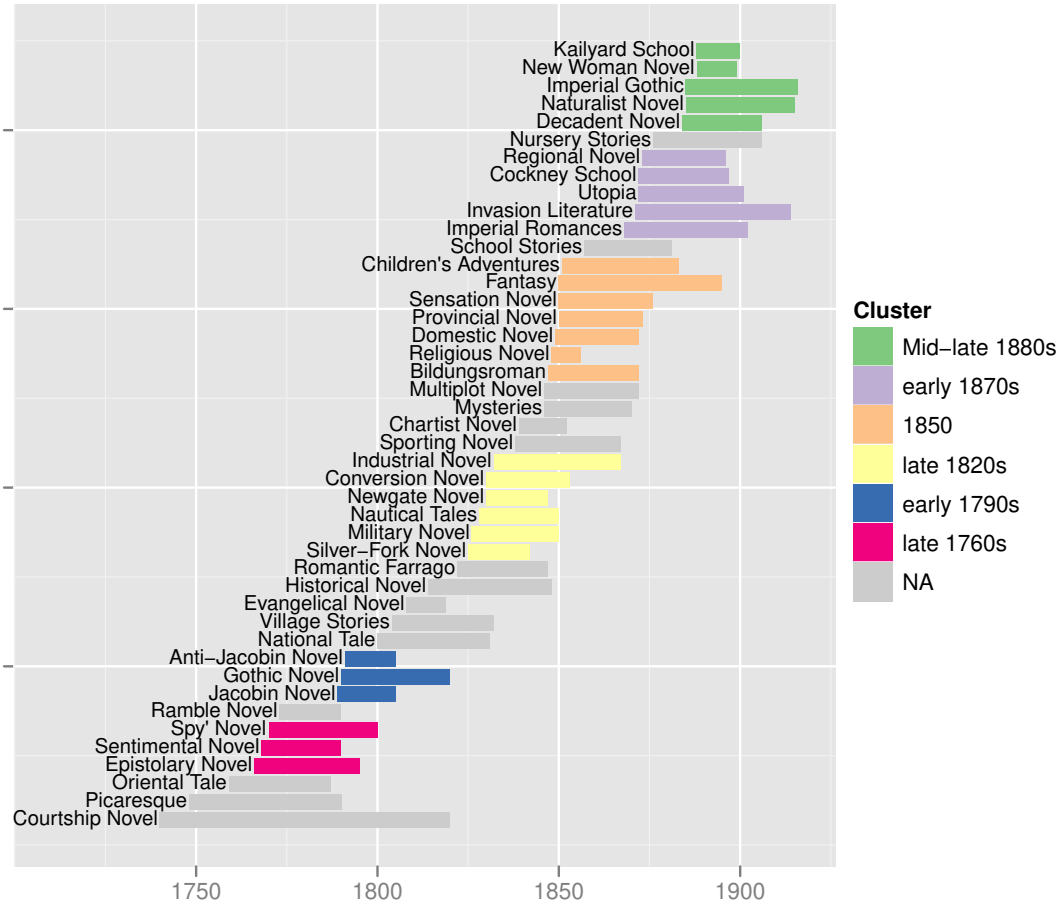


FIGURE 3.2: Periodizations of forty-four British novelistic genres given in Moretti (2005). Moretti also identifies "six major bursts" of genre creation and estimates of these clusters are shown in the coloring of the periods. Figure reconstructed from data in Moretti (2005), 31–33.

ogy or deploying the same convention could help to identify and even resolve such disagreements.

Classifications and periodizations may be called into question even in the absence of scholarly disagreement. One possibility is that experts themselves may consciously or unconsciously adjust the period during which novels admissible as members of a genre are found towards convenient or historically significant "focal dates" (Shalizi

gives 1814-1840 as the period for the silver fork novels and Kelly (1976) gives 1780-1805 for the Jacobin novels, but Moretti (2005) reports 1825-1842 and 1789-1805 respectively.

2011, 118). These focal dates may be years ending a decade or historically important years such as 1789, 1848, and so forth. This kind of adjustment could account for the surprising number of the fourty-four genres identified by Moretti that end on a decade boundary (fig. 3.3).[9] Adburgham's work on the silver fork novels provides one example of why such adjustments warrant our attention. While Adburgham provides reasons for ending her bibliography and the period of popularity of the silver fork novel in 1840—such as many of the genre's principal authors having ceased writing and the definitive end of the Regency Era with the marriage of Queen Victoria—Adburgham admits that a handful of silver fork novels were published after 1840 (Adburgham 1983, 309-319). Given this continued activity, it seems possible that other novels missed by Adburgham may have merited the classification and that the decline of the silver fork novel may not have been as rapid as described. That Adburgham's periodization ends so neatly on a decade boundary heightens this concern.[10]

Researchers may also be interested in studying groups of novels exhibiting similar morphology but that do not strictly match existing categories. Having a somewhat more general method of identifying novels with similar features that does not require human readers to sift through thousands of novels—or hundreds of thousands in the case of novelistic production in the twentieth century—would be useful. Consider Larry Isaac's recent work with the late nineteenth- and early twentieth-century American "labor problem novel," defined by the presence of specific representations of labor militancy (typically, a labor strike). The time frame for his study covers

9. Rather than around four or five occurrences of a genre's period ending on a decade-boundary, there are ten such cases. There is no reason why any ending digit should appear more often than any other, so the frequency of each digit should be distributed uniformly, with an expected value around four or five occurrences (44 / 10 final digits = 4.4).

10. Moretti (2005) departs from Adburgham's periodization and uses 1842 as the final year of the silver fork novel. Given the activity of Catherine Gore, 1842 seems an improvement on 1840 but it does leave unresolved the question of what it means for a genre's period to end as silver fork novels still appeared after 1842. For example, Catherine Gore's *Castles in the Air* was published in 1847.

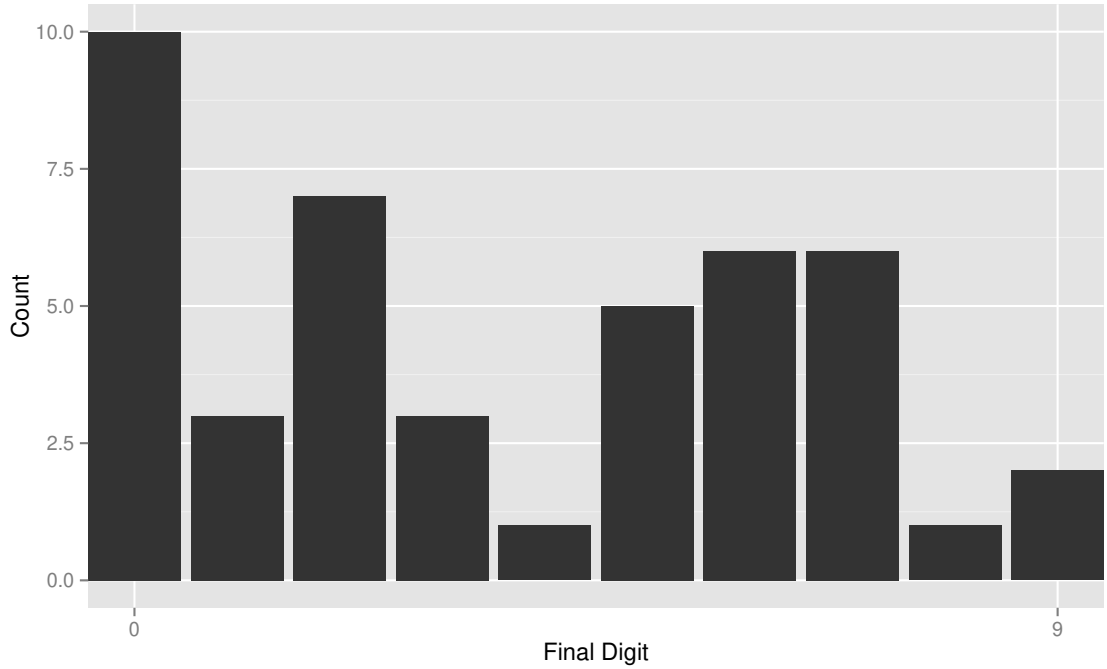


FIGURE 3.3: Histogram of final digit in the ending year of the forty-four British novelistic genres identified in Moretti (2005). The final digit "0" indicates that the periodization ended on a decade boundary, such as 1790-1820.

nearly fifty years, 1870-1918, and the scale of novelistic production during this period makes reading through all novels for mentions of labor strikes prohibitively time consuming. Isaac made no use of quantitative text analysis, relying instead on existing bibliographies concerned with novels featuring representations of labor movements in the United States. Yet having some means of reading through all surviving (and scanned) novels published during that period for characteristics similar to the novels already identified as labor problem novels would be desirable because without such a comprehensive survey of novelistic production or random sampling there is no way to say what novels may have been missed. That is, it is difficult to say how many more novels might have been identifies as "labor problem novels" had all novels published in the period been considered.

There may also be theoretical insights to be gleaned from a convincing model of

the similarities among novels. Airplanes do not achieve flight by the same means as birds, but understanding how airplanes generate lift helps us understand how birds fly as well. Analogously, if a model of similarities of novels in terms of word frequencies yields the same judgments of genres as literary historians, inspecting how the models work may yield insights valuable for historians. In this respect, rather than being seen as "reductive" models might be thought of as an additional kind of "thick description," providing an abstract account of common features while always making reference to the words observed in each individual novel.

3.2.3 Characterizing Novelistic Genres with Shared Morphology

The idea that the novels in a genre are characterized by shared morphology is common. For instance, Moretti appeals to shared morphology as a working definition of novelistic genre generally: "morphological arrangements that *last* in time, but always for *some* time" (Moretti 2005, 14). And we have already seen a definition of the gothic novels relying on shared features—"sullen, craggy landscapes; decaying mansions with dark dungeons, secret passages, and stealthy ghosts…." Moving from morphology inferred by a human reader to individual words (or word frequencies) should not be done without an abundance of caution. Readings happen inside people's heads; it is by no means self-evident how paragraphs or individual words relate to readers' identification of a particular feature or morphology in a text, such as a particular setting or plot device. For the moment, however, I make the provisional assumption that it is possible to move between human-perceived features—what I take to be referenced by Moretti and Abrams and Greenblatt—and word frequencies for the specific genres under consideration. For example, given the description of the gothic novels, we anticipate a set of words—e.g., "ghosts," "dungeon," "cell," "manor"—being more likely in gothic novels than non-gothic novels. This does not mean that the only way for a novel to feature "stealthy ghosts" is for the novel to

contain a word referring to ghosts, such as "ghost" or "projection." A novel's narrative may feature ghosts in its storyline without any synonyms of "ghost." But if a novel does make extensive use of words and phrases strongly associated with the characteristic features mentioned, we should anticipate literary historians classifying a novel as a gothic novel. Considerable uncertainty about the relationship between human readings and word frequencies does not entail that we give up any attempt to reason about the former. For the three genres considered in this paper, the descriptions provided by genre experts and a passing familiarity with a handful of novels associated with the genres warrant a provisional assumption that individual words (unigrams, or 1-grams), in addition to being features of the novels in their own right, provide information about morphology that may be described more generally.

Defining a group of novels by reference to explicitly shared morphology, however, has important limitations. Consider a small collection ten novels: five gothic novels and five randomly selected non-gothic novels. Comparing the two sets of novels, we find that the presence of a small number of characteristic words does indeed distinguish gothic novels from non-gothic novels. "Depraved," "inhuman," "monstrous," "mouldering," and "turbulent," are unique to the gothic novels and these words come as no surprise given descriptions of the genre (fig. 3.4). Attempting to generalize an approach relying a fixed list of words, however, runs into two difficulties. First, what counts as relevant morphology is in important respects arbitrary and, second, even when those doing the classifying agree on relevant features they may disagree on how to measure them. One group of literary historians may believe plot structure is more relevant than vocabulary for determining a novel's genre. Another group may put weight on "paratext"—e.g., frontispieces, illustrations, binding, paper, and typeface. Yet another group may stress particular aspects of the narrative, such as focalization, presence of indirect discourse, or absolute number of characters (Elson, Dames,

and McKeown 2010; Moretti 2005).[11] And even within these groups there may be disagreement about how to measure features. If one regards the number of distinct characters in a novel as relevant— 紅樓夢 (*Dream of the Red Chamber*) has more than 400 characters—does one count the total number of distinct characters or does one account for the novel's length and consider the average number of characters per thousand words?[12] Different measurements may give rise to different categorizations. These two challenges are not purely theoretical. If we consider the same group of ten novels and randomly assign half to one group, it is not difficult to find words that distinguish the first group as a distinct category: "balmy," "frowns," "hushed," "nothings," and "trance" (fig. 3.4). With countless features available to describe any given novel and countless interpretations of those features, it will be possible to locate properties that, taken in isolation, support almost any classification.[13]

Figure 3.4 offers a succinct account of why thinking about genre in terms of a limited range of morphology will not provide us with a reliable way to identify novelistic genres in the nineteenth-century novel. Additional assumptions about what counts as relevant morphology and what "sharing" or "arrangement" means are required. The comparison with classification efforts in biology is helpful. Consider the category of "warm blooded animal," a grouping that includes both birds and mammals. Since the nearest common ancestor of birds and mammals is a cold blooded organism, focusing narrowly on one shared feature like warm-bloodedness as a sign of similarity will give rise to incorrect classifications (in terms of ancestry). Ideally, a method of grouping organisms or objects together will rely on a theory about why similar groups

11. Of course, specific elements of narratives may be of interest, as in Vladimir Propp's *Morphology of the Folktale.*

12. In computational linguistics and other contexts, procedures like this often are referred to as "normalization."

13. There are no guarantees of agreement on relevant features and measurements. It is unlikely that radically different conceptions of morphology could result in shared categories. The range of morphology one might consider is endless: number of vowels, chemical composition of the ink, month of publication, and so forth.

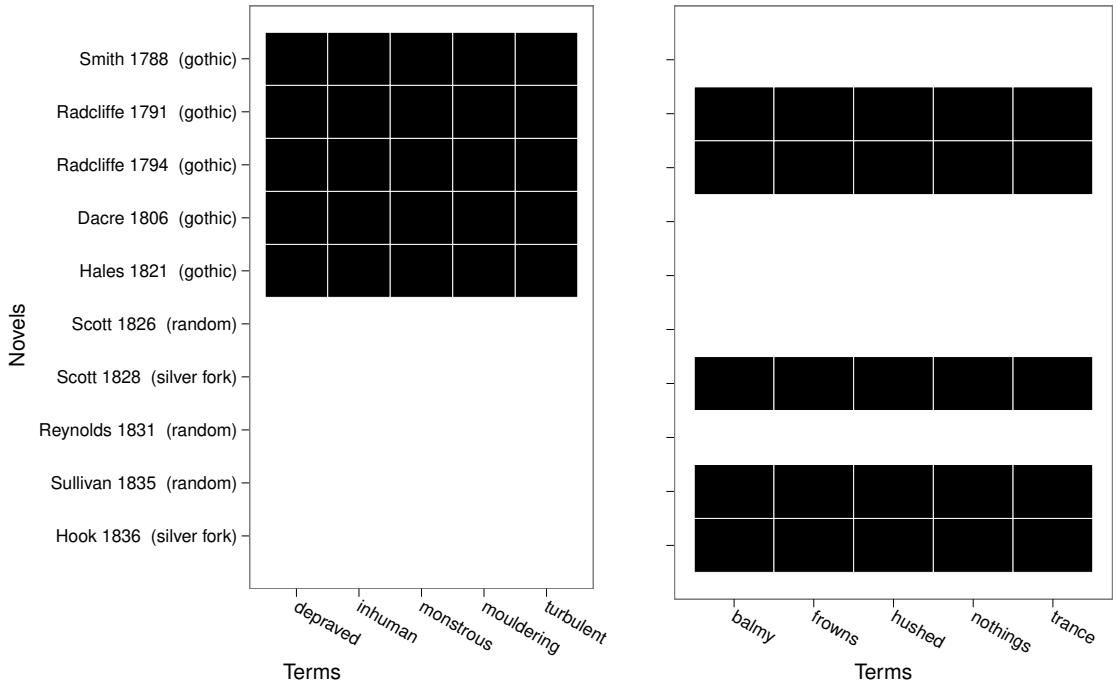


FIGURE 3.4: Words characteristic of five gothic novels in a collection of ten novels (left) and words characteristic of a random partition of the same ten novels (right). A solid block in the grid indicates the presence of the word indicated at the bottom of the figure.

emerge in the first place. Biologists who group species together based on similarities in allele frequencies have a precise theory about why organisms have different allele frequencies. It is this theory that motivates biologists' attention to allele frequencies rather than superficial morphology such as an organism's size or coloration. It is this theory that biologists appeal to when confronted with someone who wants to make coloration a primary consideration in classification—who would insist, for example, that male and female mandarin ducks are different species.[14] With the novel, although literary historians have general accounts about why novels might resemble other novels—see, for example, the preceding discussion of Bourdieu—their theories are difficult to apply when discussing specific features of individual novels. The absence of such a precise theory warrants an abundance of caution when making

14. The coloration and plumage differs dramatically depending on the duck's gender.

guesses about the connections among novels based on features such as vocabulary, plot structure, and setting. Even direct paraphrase is no guarantee of a connection between two novels.[15]

An alternative strategy for defining a novelistic genre would welcome a wide range of morphology, define a way to measure similarity among novels given that range of features, and insist novels of the same genre will tend to be more similar to each other than to novels not associated with the genre. We might describe this approach as looking for "family resemblances" among novels. Just as it is often possible to guess at familial relationships in the absence of any one trait that all family members share, such as eye or hair color, it may be possible to group novels together despite there being no single feature that all members of the group share. Such an approach would avoid problems associated with fixing a set of features. Without a fixed dictionary of features, such an approach would also trivially accommodate new features, something that should be reassuring. That is, a method of grouping novels together that maintains its groupings even when new features are added—e.g., binding, city of publication, writer's social connections—seems more reliable than a method that ignores or cannot accommodate new features.

One class of methods that approaches clustering in this fashion defines the similarity—or, equivalently, distance—between every pair of entities (novels, artifacts, biological organisms, etc.) in a population. These methods are frequently labeled "distance-based." Clustering biological populations by calculating the distances between organisms based on measurements of morphology is especially common in cases where genetic information is lacking. With distance-based methods, much depends on how similarity is measured: different measurements often yield different clusters. In

15. Imagine finding a contemporary novel that begins with phrase "It is a truth universally acknowledged, that ..." There is no guarantee that the phrase is borrowed from Austen's *Pride and Prejudice* rather than from some intermediate source. A similar argument would follow when considering isolated paraphrase or quotation of portions of the Bible.

the case of novels, a comparison of the word frequencies of two novels might use measurements of similarity such as Jaccard similarity or cosine similarity. Jaccard similarity focuses on the number of vocabulary elements shared by two novels and cosine distance considers the cosine of the angle between two vectors that contain the word frequencies of the two novels being compared. (Comparing novels using cosine similarity is described in chapter 2.) These measurements of similarity do not always agree. Jaccard similarity makes no consideration of how many times words occur beyond the first occurrence. In a corpus where documents are distinguished by the concentration of certain words (rather than their presence or absence) Jaccard similarity may yield considerably different measurements of similarity than cosine similarity.

A different class of methods for clustering novels is "model-based." These begin with the assumption that the novels originate (in a sense to be specified) from a fixed but unknown number of groups. A model-based approach then infers the group membership of each novel based on its features, as well as inferring the number of groups present in the corpus. Model-based methods tend to be associated with probabilistic clusterings as a novel's assignment to a group is expressed as a probability (Pritchard, Stephens, and Donnelly 2000, 2–3). A model-based approach to clustering novels will be used in this chapter because such models provide for the best resolution the problem of polysemy—a particular challenge whenever word frequencies are used as morphology. Jaccard and cosine distance fail to distinguish among, for example, the "hook" in "Theodore Hook" (a silver fork novelist), "coat hook," and "right hook." More importantly, model-based approaches also permit the prediction of words in an additional unseen (or imagined) novels. It is frequently not possible to make these predictions with distance-based methods. Assessing models' predictive performance is a convenient way to compare the accuracy of competing models because the measure is so readily understood. Given a corpus of a hundred

novels, one novel may be "held out" and competing models asked to predict the words that occur in the held-out novel.[16]

The model-based approach used in this chapter is based in large part on Latent Dirichlet Allocation (LDA), which has been used in a variety of settings, including in the humanities and interpretive social sciences (Block and Newman 2011; Mimno 2011, 2012a). The specific model used in this chapter is a non-parametric extension of LDA which uses the Hierarchical Dirichlet Process (HDP) as the means of inferring the association of novels with a number of latent groups, where the number of latent groups is not specified in advanced (Teh et al. 2006).

Before describing the corpus and the model in detail, the leap from discussions of morphology in general to discussions of word frequencies alone deserves additional description. As mentioned above, being set in a haunted manor is a feature of novels about which words ("manor" or "haunted") may give us some information. Determining whether or not a novel describes action in a given setting requires a trusted human reader: word frequencies alone cannot distinguish between discussions of a haunted manor around a table in London and action being set in a haunted manor. To what extent models based on word frequencies reliably predict the presence of features described by readers is an empirical question about which the experiment pursued in this chapter will indirectly answer. If a model based on word frequencies predicts clusterings of novels that align with expert classifications better than random chance, such a result should count as evidence that word frequencies are informative about morphology identified by human readers.

16. Even absent a held-out novel, a model may be used to generate the words of a fictitious novel. The semantic coherence (or other anticipated properties) of these words may be measured against an appropriate standard (Mimno and Blei 2011).

3.3 Data: Three Novelistic Genres

The corpus used in the analysis consists of a random sample of novels published between 1800 and 1836 and a representative sample of gothic, silver fork, and national tale novels. A bibliography of novels published in England and the British Isles between 1770 and 1836 is available to support the random sampling of novelistic production from library collections (Garside and Schöwerling 2000). The bibliography by Garside and Schöwerling is the product of decades of work and has a solid claim to be comprehensive.[17] The characteristics of gothic novels have been mentioned above. Silver fork novels are known for their portrayal of fashionable society and are often set in London. (An additional small silver fork is a culinary accessory found on dinner tables among the wealthy.) Adburgham lists the "essential facets" of a silver fork novel (referring to Lister's *Granby*): "there are some politics, some gambling scenes and a duel; there are dazzling balls in the London season, and country-house parties in the winter; the characters include a dandy, a toad-eater, a scheming high-society villain, a pair of lovers ill-starred until towards the end of the third volume. There are social climbers clambering towards Almack's [a social club], provincial belles at a race meeting ball in Doncaster Assembly Rooms; there is satire at the expense of the middle class and the rich roturiers. But above all, there are semi-flirtatious drawing-room conversations and dinner-table repartee" (Adburgham 1983, 92-3). National tale novels are a varied group but include many bestsellers. National tale novels were known for featuring a protagonist who travels to Scotland or Ireland and for sharing a similar narrative structure. Trumpener describes the basic plot shared by early national tale novels as follows: "[A] young hero or heroine,

17. Even novels of which no (known) copies survive are included in the bibliography as their existence may be inferred on the basis of publisher advertisements, book reviews, and related sources. Based on a random sample from the bibliography, I found that scans of the majority of novels published between 1800 and 1836 are available in some form from consortia devoted to library digitization (Internet Archive, Hathi Trust, and so forth).

raised in England or on the continent, travels to Ireland or Scotland expecting to find barbarism. Instead, the protagonist falls in love with his or her new surroundings and with the aristocratic native guide who has helped him or her understand the region's beauty and cultural interest. The novel ends with the marriage of the lovers—and thus also with the allegorical union of Britain and its constituent 'national characters"' (Trumpener 1998, 910). The corpus contains 35 gothic novels, 22 silver fork novels, 18 national tale novels, and 18 randomly selected novels. These novels are listed in appendix C.

The random sample of novels is drawn from the exhaustive survey of novelistic production in Garside and Schöwerling (2000). The genre-specific samples are drawn from two different types of sources: random samples from the genre-specific bibliographies of Adburgham (1983), Lévy (1968), and Trumpener (1998), as well as the collection of well-known novels associated with the three genres used in Allison et al. (2011). (An example of a well-known gothic novel would be *The Mysteries of Udolpho* by Ann Radcliffe.) Scans and machine-readable text versions of the novels were gathered from a number of repositories, including the Internet Archive (in particular, the University of Illinois at Urbana-Champaign's nineteenth-century novels collection), Project Gutenberg, University of Adelaide, and the Corvey Collection. The random sample originally included twenty-four novels. Scans of two novels falling in the random sample could not be located. Four of the novels in the random sample were also listed in the bibliography of silver fork novels of Adburgham and are counted among those novels.[18]

From the corpus I removed a selection of frequent words (stop words), words

18. There are 99 silver fork novels mentioned in the bibliography of Adburgham and the population of novels published during this period is 2,903. The probability of finding four or more such novels in a sample of twenty-four is quite low, roughly 1 in 100. To verify that nothing had gone wrong during sampling, I counted the number of novels appearing in the first 100 novels in the sample that also appeared in the silver fork bibliography. Six silver fork novels appeared in the first 100 sampled novels. Finding six or more in 100 trials is expected to occur more than ten percent of the time.

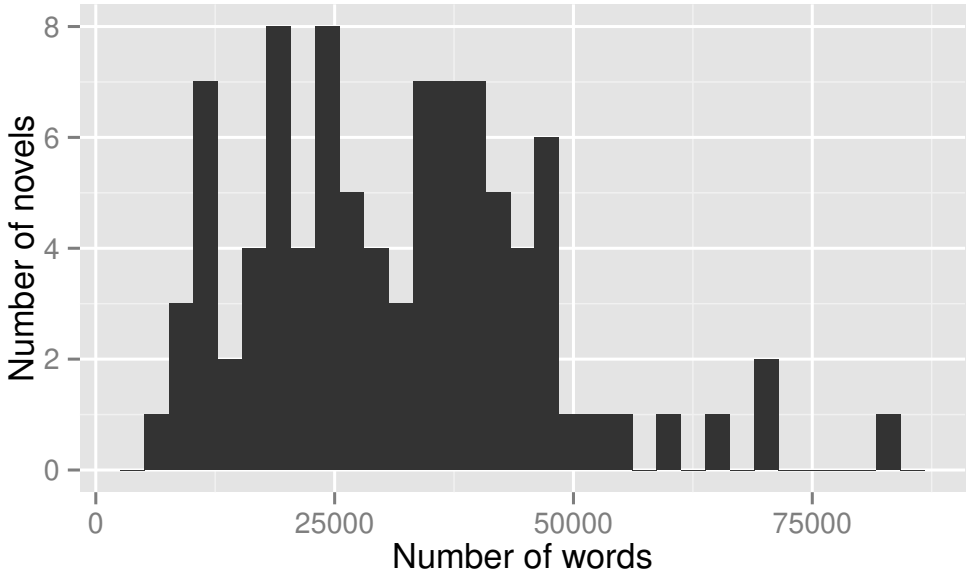


FIGURE 3.5: Word counts for the 93 novels in the corpus.

having fewer than three characters, words occurring in fewer than five novels, and words corresponding to character names their capitalized forms of address ("Mr," "Miss," "Captain," etc.). The final corpus includes 93 novels, 31,808 types, and 3,806,014 words. After removing the words mentioned above, the median length of a novel in the corpus is 40,178 words. Figure 3.5 shows the lengths of the novels after preprocessing. The shortest and longest novels in the corpus are both by Maria Edgeworth. The shortest is *Castle Rackrent* (1800) and the longest is *Tales of Fashionable Life* (1809), a collection of stories.

For the gothic and silver fork novels, we can be precise about their share of novelistic production. The gothic novel was hegemonic in its heyday, accounting for 30% of new novels during its peak year (Moretti 2005; Lévy 1968; Garside and Schöwerling 2000). Between 1825 and 1836 Adburgham identifies 60 silver fork novels, 5 percent of the 1,024 new titles published during those years. There has been no attempt to collect a list of all candidate national tale novels. An informal estimate based on the number of journal articles mentioning the genre in *Nineteenth-Century*

Literature suggests it is roughly comparable to the silver fork novels: 27 articles mention "silver fork" and 23 mention "national tale" between 1986 and 2011 (Vols. 41-66).

Several reasons inform the choice of these three genres. First, having a range of genre sizes seemed desirable. The gothic novels represent a large "market category" during the late eighteenth- and early nineteenth-century (Moretti 2005). The other two genres are smaller. Second, an adequate representation of the challenge of inferring novelistic genre based on minimal information (word frequencies) required two genres that were perceived to be "similar" in the sense of appearing in roughly the same period and sharing publishers and authors. The silver fork and national tale genres fit this description as many of the novels in both categories were published by Henry Colburn. Also decisive was the illustration by Allison et al. that these two genres, unlike others, could not be easily distinguished using punctuation and word frequencies of frequent words (Allison et al. 2011, 19).

3.4 Modeling Novelistic Genre

Introduced in chapter 2, Latent Dirichlet Allocation (LDA) offers a representation of a corpus of texts in terms of latent features. These latent features are frequently referred to as "topics." Recall that LDA sets forth a story about how the texts in the corpus were generated. The model posits that each word in a text derives from one of a number of latent topic distributions (distributions over the vocabulary of the corpus). Such an assumption allows for the description of a document in terms of the proportion of its words associated with each topic. For example, if there are three topics latent in a corpus of novels, an individual novel may be described in terms of the latent topic assignments associated with its words—e.g., 0.8 topic one, 0.1 topic two, and 0.1 topic three. These proportions may be thought of as topic "shares" or "weights." In order to use LDA to model an existing corpus of novels,

the generative story is run in reverse, in a sense. That is, making the assumption that there are a number of latent topics responsible (in a manner specified by LDA) for the words in the corpus, it becomes possible to infer the association of words with topics. The assumption that there is a fixed number of topics (specified in advance) may be relaxed by using a non-parametric version of LDA, which uses the Hierarchical Dirichlet Process (HDP) to infer the number of latent topics (Teh et al. 2006). Since the HDP may be used in a variety of models, including those that have very little in common with LDA, I will follow Teh et al. and use "HDP-LDA" to refer to the non-parametric extension.[19]

While topic shares are often used to summarize the contents of documents in terms of constituent "themes"—as was the case in chapter 2—the topic shares may also be interpreted as a form of classification. Indeed, it was with classification in mind that the probabilistic model now familiar as LDA was first developed in Pritchard, Stephens, and Donnelly (2000). To understand the distinction, it is helpful to replace the word "topic" with "population" and think of each document in a corpus as deriving from an admixture of the characteristics of a number of distinct types. With this conceit, HDP-LDA characterizes a novel as a mixture of distinct types. In this case we are less interested in the words associated with the types than with the distribution of types in the corpus. If, for example, HDP-LDA characterizes all the gothic novels (and only the gothic novels) as roughly 0.8 population 1, 0.1 population 2, and 0.1 population 3 then the model's characterization roughly matches the expert classifications for those novels. This way of comparing the expert classifications and HDP-LDA's representation of the corpus will be made more precise in the following section.

While it is possible (and easier) to use a model that assigns each novel, based on its word frequencies, to one and only one of a number of populations, such a model is

19. HDP-LDA is the HDP with a Dirichlet base distribution.

inconsistent with what is believed about literary genre. There are countless examples of novels that borrow from more than one genre. A contemporary example in wide circulation is *Blade Runner* (1982), a film based on Philip K. Dick's *Do Androids Dream of Electric Sheep* (1967), which borrows from conventions in both science fiction and detective stories (Kerman 1997). An example closer to our period would be Bulwer's bestselling *Paul Clifford* (1830), a story of a prosperous gentleman who also leads a life as a criminal.[20] *Paul Clifford* has been justifiably classified as both a silver fork and a Newgate novel (Adburgham 1983; Hollingsworth 1963). A mixed-membership model allows for the modeling of each novel as a mixture of multiple populations of novels.[21]

3.4.1 Alignment with Expert Classifications

This section compares the description of the novels in the corpus provided by HDP-LDA with the classifications provided by genre experts. It is worth recalling that the model makes no use of the expert classifications or indeed anything other than the words found in the texts of the digitized novels. In the nomenclature of machine learning, the model is "unsupervised." Save for the decisions about tokenization and what elements of the vocabulary to include, classifications based on this model are made independently of the judgments of human readers.

As has been discussed above, the assumption here is that the classifications provided by the experts are, at least for this subset of the genres, accurate and comprehensive. As the corpus includes only 93 novels, this is an assumption that is significantly easier to check than verifying a periodization of a single novelistic genre, which requires familiarity with the breadth of novelistic production. That the three

20. The novel is also famous for its opening line: "It was a dark and stormy night."

21. Modeling the words of each document as associated with a single latent type is undesirable for other reasons as well. Mentioned above, one important measure of how well a model works is how well it predicts held-out portions of a corpus. By this measure, the mixed-membership model performs much better for a wide range of texts (Blei, Ng, and Jordan 2003).

genres in question have accessible lists of novels associated with them also makes checking this assumption feasible.[22]

In order to compare the model's description of the corpus with the expert classifications, we need to put them in terms that are comparable. There is not a straightforward way of doing this for the simple reason that the pooled judgments of the genre experts give us three categories (four if we count the random selection as its own category) whereas the probabilistic model loosely classifies the novels into as many as 50 categories. The difficulty in comparing these classifications is not difficult to see. An analogous situation would be one where two people partition a collection of articles from several newspapers into categories and one person uses categories that are considerably finer than the other person. That is, one person groups articles into categories such as "sports," "politics," and "business," and a second person makes an initial partition identical to the first person's but then further divides the categories based on the newspaper in which the story originated—e.g., "Guardian-sports," "Neue Züricher Zeitung-business," and "人民日报-politics." While it is plain that these two clusterings are similar in some sense, it is not obvious how to formalize this notion of similarity. At minimum it seems desirable that any measure characterize a finer and a coarser clustering of objects as more similar to each other than two clusterings that have been made at random. A family of measurements of the similarity of clusterings that satisfies this requirement is based on mutual information (Meilă 2002). Mutual information is a measure of the relationship between two random variables (MacKay 2003, 138–40). (Correlation is another, perhaps more familiar, measure of the relationship between two random quantities, such as the height and weight of an organism selected at random from a population.) Moreover,

22. By accessible I mean that the literary historians in question have provided a list of novels they associate with the genre. This list often appears in an appendix or separate section, as it does in Adburgham (1983) and Lévy (1968). The list of national tale novels provided by Trumpener (1998) is accessible in the sense that it comes in the form of a short encyclopedia entry dense with references to specific novels.

it is possible to adapt a mutual-information-based metric to the problem at hand: a topic model such as LDA or HDP-LDA provides a probabilistic or "soft" classification of the novels, whereas the experts assign each novel in the corpus to a single category.

The comparison of a soft clustering and a firm clustering is easy to understand when the number of clusters is the same. If we were dealing with a topic model that described this corpus in terms of three topics, we could treat the proportions of words associated with each topic as the probability that the novel comes from one of three types. So if the model described the novel as originating from a mixture of topic 1, topic 2, and topic 3 in the proportions 0.8, 0.1, 0.1 respectively, then we would treat this description as implying that the document is in group 1 with probability 0.8, group 2 with probability 0.1, and group 3 with probability 0.1, where probability in this context is an individual's characterization of uncertainty about the classification (Kadane 2011, 1–8).[23] The expert classifications have a similar interpretation: if there are three groups (gothic, national tale, and silver fork) a classification of a novel as a gothic novel corresponds to the assignment of the novel with group 1 with probability 1, group 2 with probability 0, and group 3 with probability 0—written more concisely as $(1, 0, 0)$. Mutual information permits us to ignore concerns about the cluster labels or indices "lining up." We can see that a model's judgment that all gothic novels likely belong to group 1 with probabilities $(0.8, 0.1, 0.1)$ is close to the expert classification $(1, 0, 0)$. And if the model's judgment for all gothic novels were $(0.9, 0.05, 0.05)$ that clustering would be even closer. A final adjustment to mutual

23. It is something of a leap to go from talking about proportions and shares to talking about probability of group membership. Strictly speaking, HDP-LDA and LDA (and the model offered in Pritchard, Stephens, and Donnelly (2000)) are mixed-membership or admixture models. Words in novels, under these models, originate from different sources. It is not faithful to the models to interpret the topic proportions as classification probabilities. With this in mind, the mutual information comparison finds some justification if the topic shares are viewed from the perspective of a human classifier who must assign a single label to a population containing a mixture of types. That this imagined classifier might assign labels with probabilities that relate to proportions seems reasonable.

information will permit the comparison of clusterings with varying numbers of categories. Normalized mutual information adjusts the mutual information calculation onto a standard scale between 0 and 1, where 1 corresponds to perfect alignment between two clusterings.[24]

Figure 3.6 shows how well the HDP-LDA model aligns with expert classifications. HDP-LDA reliably performs better than a random assignment of the novels to the four categories (gothic, national tale, silver fork, and other).[25] Figure 3.7 shows the agreement between the expert classifications and the model for each genre separately. That is, the clustering of the model is compared with an expert classification of gothic and non-gothic novels, silver fork and non-silver fork novels, and national tale and non-national tale novels. Again, taking the expert classifications as authoritative and based on detectable features, the higher mutual information between the expert classification of the gothic novels and the model suggests that gothic novels are share features more consistently or that features particular to the gothic novels are better identified by the model—or some combination of these two states of affairs. This result seems consistent with what is known about the three genres. National tale and silver fork novels already proved difficult to separate in the study by Allison et al.

In order for the model to agree with expert classifications better than chance, the HDP-LDA model must be picking up on features that characterize the genres

24. Normalized mutual information is defined in terms of mutual information between two clusterings and the respective marginal entropies of the clusterings. Let C be the class of expert classifications ("gothic," "silver fork,""national tale," and "other") and let C' be the set of topics inferred by a topic model. The mutual information $MI(C, C')$ between the expert classifications and the topic model is given by $MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}$, where $p(c_i)$ and $p(c'_j)$ denote the probability that a novel selected at random from the corpus falls into c_i and c'_j respectively. $p(c_i, c'_j)$ is the probability that a document selected at random falls into both c_i and c'_j. Normalized Mutual Information is given by $NMI(C, C') = MI(C, C')/\max(H(C), H(C'))$, where $H(C)$ and $H(C')$ are the respective entropies of the two clusterings being compared.

25. That the alignment is not stronger is due to, at least in part, the fact that the "classification" provided by HDP-LDA is so fine. The finer classification yields a higher entropy, which figures in the denominator in the NMI calculation.

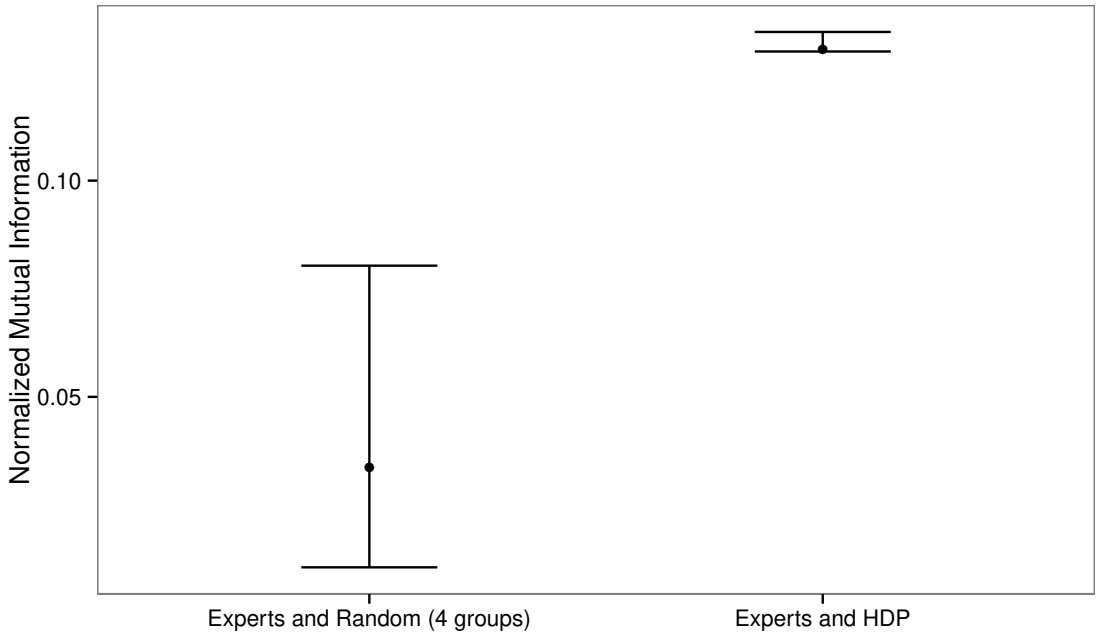


FIGURE 3.6: Agreement between model and expert genre classifications measured by normalized mutual information. Higher scores indicate clusterings are closer to expert classifications. Error bars indicate 95% credible intervals based on simulation in the case of random clusterings and based on sampling from the posterior distribution in the case of the HDP-LDA model.

in question. Inspecting the topic distributions can give us some sense of what these features are. For instance, one topic that occurs frequently among novels identified as gothic contains words such as "convent," "castle," "bosom," "melancholy," "cavern," and "cell." Topics likely to be found in silver fork novels assign high probability to words such as "ambition," "opera," "society," "marriage," "season," and "fashionable." Finally, the two most prominent topics among national tale novels feature words including "irish," "national," "revolution," "foreign," "ancient," "influence," "pure," and "missionary" (fig. 3.8). This ad-hoc inspection of the topic distributions, while it yields results that appear to confirm the accuracy of the model, is in other respects like "reading tea leaves"; the topic distributions also assigns high probability to words less consonant with our preconceptions, such as "artist" for national tale,

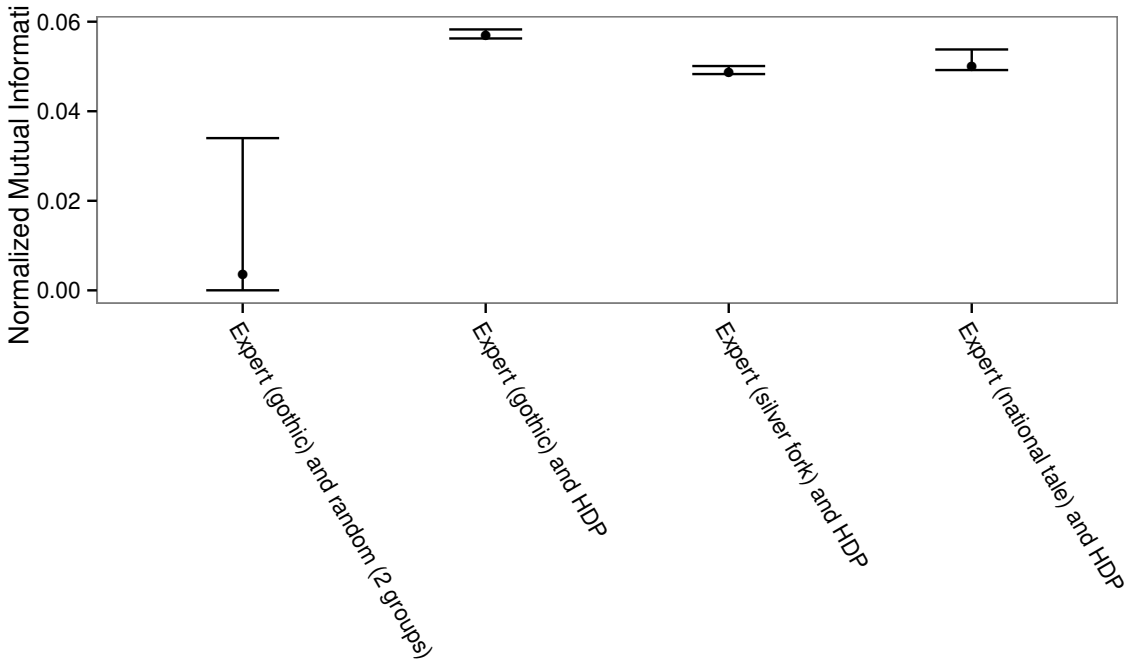


FIGURE 3.7: Agreement between model and expert genre classifications measured by normalized mutual information. Agreement shown for each genre separately. Higher scores indicate clusterings are closer to the expert classifications. Error bars indicate 95% credible intervals based on simulation in the case of random clusterings and based on sampling from the posterior distribution in the case of the HDP-LDA model.

"astrologer" for silver fork, and "permitted" for gothic. But such skepticism should be tempered by the foregoing evaluation of the model in terms of mutual information, which offers evidence that the model is indeed picking up on characteristics across the entire corpus that align with expert judgments better than chance.

While the evaluation of the model in terms of mutual information confirms that a probabilistic model resembles the classifications of literary historians, the degree of the resemblance is difficult to interpret. Mutual information does not have a ready analog in the experience of readers. Yet there is also no standard against which an assessment of the similarity between novels in terms of their topic distributions might be compared. As has been discussed in the previous section, literary historians' definitions of genre do not tend to be precise or easily reproducible. Would a group

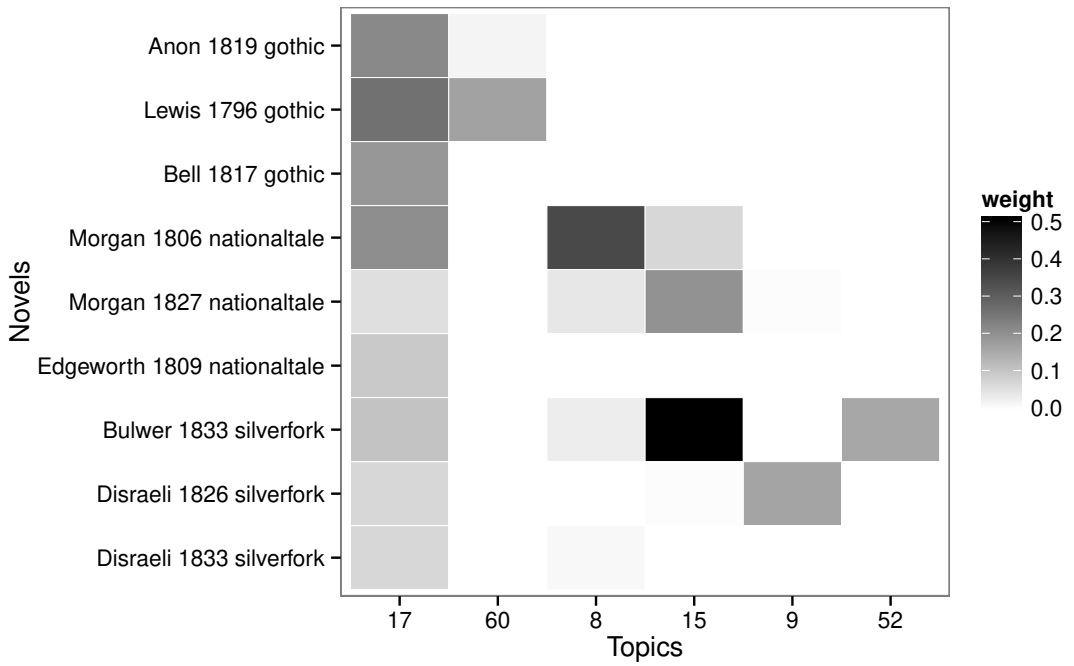


FIGURE 3.8: Selected novels and their topic shares in the HDP-LDA fit of the corpus. Novels and topics have been chosen in this figure to illustrate areas of strong agreement between the model and the expert classifications (indicated to the left of the author and year of novels).

Topic 17	entered	affection	presence	remained	youth	deep	object
Topic 60	monk	nuns	madrid	hastened	chamber	bed	friar
Topic 8	religion	influence	pure	object	missionary	genius	priest
Topic 15	brien	brussels	honoria	church	bog	french	ancient
Topic 9	highness	court	fane	hero	squib	bravo	political
Topic 52	ambition	genius	political	pause	wisdom	actress	deep

of literary historians, agreeing to the definition given by Abrams and Greenblatt (2000) and provided with access to and time to read all surviving novels published between 1790 and 1836, arrive at the same list of gothic or national tale novels? I suspect there would be some variation in the resulting lists. The variability of the lists, were it available, would provide some standard by which one could compare the topic model's description of the corpus.

The larger importance of this result lies in the generality of this method used. That HDP-LDA is able to recover—even in a limited way—groupings of novels iden-

tified by human readers gives us reason to expect that it will deliver similar results on related corpora. The present corpus was not chosen in any way to make it particularly amenable to analysis by the model. The result gives us reason to believe that HDP-LDA may be of use in cases where minimal prior bibliographic work exists, as is the case for many of the forty-four novelistic genres identified by Moretti (2005).

3.5 Implications for Literary History

For literary historians interested in identifying groups of novels sharing similar vocabulary, probabilistic topic models are invaluable. The similarity of any pair of novels may be compared by looking at their topic weights. (This notion of (dis)similarity between topic distributions has a convenient measure, the Kullback-Leibler divergence.) In this fashion, researchers may take a large number of texts and find groups of novels that are similar to each other. This is the "find more like these" functionality familiar from information retrieval contexts, here adapted to literary history. Critically, when researchers observe similar topic weights, they have an interpretation of the weights available in the topic distributions associating words with topics—such as figure 3.8.

To make the practical benefits plain, consider the following two examples. First, imagine the steps literary historians must take to establish the period of popularity for national tale novels. One step would require collecting a list of all candidate national tale novels published in the nineteenth century.[26] If they were to start with Trumpener's article on the genre, they would find it mentions only 25 national tale novels. Many national tale novels go unmentioned. Even a cursory inspection of the

26. For those interested in novelistic genre for what it might contribute to social history or sociology of culture, such periods are of interest (Moretti 2005). And researchers need not agree precisely on the list of novels in a genre for there to be agreement about the genre's period of popularity. Scholars may disagree on particular cases, such as whether a novel is a precursor or a full-fledged instance of a gothic novel. Despite such differences, they may agree on the years during which 95% of all gothic novels were published.

list of English novels published between 1800 and 1836 reveals a number of additional candidates with the phrase "national tale" in their titles—e.g., *Caledonia; or, the Stranger in Scotland: a National Tale* (1810) (Garside and Schöwerling 2000).[27] Using the model described in this chapter to model the similarities among novels would allow the literary historians to quickly identify novels that were similar to the 25 national tale novels mentioned in Trumpener (1998) and to calibrate their beliefs about additional national tale novels.

Another practical application involves a case of "cryptic" genre membership. Garside (1991) argues that Scott's *Waverley* (1814), commonly thought of as the first "historical novel," borrows significantly from previously published national tale novels. This comparison is of interest since *Waverley* is often judged more "literary" than the more popular national tale novels—hence Garside's title's suggestion of its "hidden origins." A fitted model of a large range of texts from the period would offer an additional perspective on this case. Novels from which *Waverley* might have borrowed could be proposed by considering those novels with similar topic weights.

3.6 Population Thinking

This chapter has focused on demonstrating that there are exist groups of novels characterized by shared morphology and detailing a method of associating novels with others that share features. The question of why novels might emerge that share morphology has been left largely undressed. Using a probabilistic model of the latent structure of a corpus of novels makes it difficult not to reflect on this question, not least because Bayesian models like LDA and HDP-LDA explicitly propose a generative story about the origin of the words observed in a collection of texts (see

27. Three novels contain "national tale" (or some variant) in their title: *National Tales* (1827), *Bleddyn; A Welch National Tale* (1821), and *Caledonia; or, the Stranger in Scotland: a National Tale* (1810). There is also a novel, *The Scottish Chieftains* (1831), whose title recalls a novel mentioned by Trumpener, *The Irish Chieftain, and His Family* (1809).

chapter 2).

The phenomenon of shared morphology in novels admits a variety of explanations. The Bourdivian perspective discussed in section 3.2.1 offers one account, but it is one among many. Another perspective would be an account drawing inspiration from evolutionary theory that emphasizes how novelistic morphology varies over time and is subjected to selective pressures (Winthrop-Young 1999; Moretti 1988). And nothing in the preceding analysis precludes an essentialist perspective on the origins of recurring morphological arrangements, one that attributes shared features to a relation between writers and, for example, an ideal "gothic" aesthetic. Another perspective would situate itself somewhere between the Bourdivian and the Darwinian outlook and would search for an explanation of regularities in novelistic morphology in the material circumstances surrounding the writing and publishing of novels. It would aim for a reconstruction of the literary field in terms of "population thinking" (Shalizi 2011).

The analogy between novelistic genres and biological populations is fruitful because the central challenge is the same. Just as every organism is unique, so too is the text of every novel. Even "identical" texts, such as subsequent editions or outright copies (e.g., an American edition of a British novel), differ in that works at least have distinct title pages. Printers also introduce countless differences, including variations in layout, paper, ink, and type. Even novels from the same printing are distinguishable by virtue of the variability of the printing process.[28] If printed works—like biological organisms—are unquestionably unique, how can genres, as collections of the "same" sort of thing, be said to exist? In biology, a similar challenge gives rise to population thinking, what Ernst Mayr calls "one of the most important concepts in biology" (Mayr 2001, 1976). Summarizing Mayr, Godfrey-Smith (2009) provides the following definition of the concept:

28. See Winter (1987) for a discussion of library holdings in this light.

> A population is a physical object, bound by ancestry and other causal relations, internally variable at any time and changing over time. To the extent that organisms fall into well-marked and recognizable "kinds" that we can give straightforward species names to, this is a contingent consequence of populational processes. A well-marked kind can split or dissolve, starting tomorrow, if local conditions push it that way (Godfrey-Smith 2009, 11).

Rather than representing some sort of fixed "type," a novelistic genre may be considered as just such a population. In order to consider something as a population, the relations connecting individuals need not be genetic or otherwise biological (Godfrey-Smith 2009, 147–164). Moreover, population thinking need not involve (natural) selection. Consider the example of the changes over time in the design of a musical instrument like a trumpet.[29] What connects individual trumpets available in 1900 with those available in 1850 are the copying and modification of existing designs by instrument makers. A similar kind of process can be put forward as linking new novels published with previous exemplars. While the terminology is different, this understanding of novelistic production is consonant with established perspectives on literary genres. Both Jameson (1981) and Bourdieu (1996) draw attention to the choice of (sub)genre by participants—writers, editors, publishers, and so forth—as a form of social signaling or position-taking; from a horizon of possible associations, publishers and writers link new novels with existing novels. For instance, a detective novel is written or commissioned in such a way that it signals its association with past exemplars of the same category. The signaling may take a myriad of forms, including explicit identification in a subtitle ("Detective Sketches") or the use of characteristic morphology.[30] From the perspective of population thinking, the chain

29. For changes in cornet design between 1825 and 1975, see Tëmkin and Eldredge (2007).

30. The subtitle example comes from Muddock, *Tracked and Taken: Detective Sketches* (London:

of associations may be understood as the ancestry that binds a population of individual novels into one recognized as a novelistic genre. This perspective provides a materialist account for why novels in the same genre resemble one another—and an hypothesis about why patterns in word usage might be detectable by probabilistic models. Whether or not these social relations reliably leave traces in the texts of novels is an open question. If such traces are to be found anywhere, it seems likely they reside in precisely the commodified corners of novelistic production—certainly including gothic, national tale, and silver fork novels.

Population thinking applied to novelistic production suggests—or recalls—two concerns of particular importance for future research working with large collections of novels from diverse geographical and linguistic situations. First, using population thinking to study novelistic production requires that shared morphology be interpreted in terms of a populational process. For example, were a reader to encounter a Japanese novel written in 1800 containing "gothic themes" by an author one believed never had any contact (direct or indirect) with gothic novels, the novel would not be a member of the same population as the gothic novels found in Britain. Tynyanov (1927) expresses a version of this concern about the historical novel, writing that "we may conclude that the study of isolated of isolated genres outside the features characteristic of the genre system with which they are related is impossible. The historical novel of Tolstoy is not related to the historical novel of Zagoskin, but to the prose of his contemporaries."[31] Second, population thinking suggests an avenue for future research in that it draws attention to the dynamic process—"internally variable at any time and changing over time"—potentially underlying the categorization of novels into recognizable kinds. For example, those features that characterized

Chatto and Windus, 1890). Moretti (2005) discusses characteristics of detective fiction in the chapter entitled "Trees."

31. In this context, Shalizi (2011) recalls the phenomenon of convergent evolution: sharks and dolphins share certain morphology—they are both "streamlined marine predators which live in the water all the time"—but they are members of distinct populations.

gothic novels in 1800 such that readers, reviewers, and other writers identified them as a "recognizable kind" are unlikely to be precisely the same features that characterized gothic novels in 1810 or 1820. In this sense there is no "the gothic novel," just as there is no "the domestic cat" as a static entity. Gothic novels, understood as a population, are connected not by common features per se but by a network of ancestral material relations. Shared features provide one means of guessing about that network, just as shared features in mammals can facilitate the inference of ancestry. A probabilistic model more attentive to changes in morphology over time would acknowledge that the words that characterize topics may be subtly shifting over time or that the writer of *Frankenstein* (1818) was unlikely to have encountered the text of *Vivian Grey* (1824). (Publication dates should not, however, be taken as gospel.)[32] The Dynamic Topic Model described by Blei and Lafferty (2006) is a promising base upon which such a model might be built.

3.7 Conclusion

In this chapter, I demonstrated a method for characterizing patterns in novels' vocabulary and word frequencies. I provided evidence that organizing novels based on these patterns yields groupings of novels that align with experts' classifications of novelistic genre better than chance. These patterns are of practical use when researchers are confronted with the task of gathering all texts belonging to any category whose members may be characterized by distinctive vocabulary use. Gathering such collections is an important task in existing work in literary history and sociology of culture. Where these tasks have been or are currently being undertaken, topic models offer the means to speed data collection and verify that novels have not been overlooked. Moreover, it seems plausible that in the past the expense and

32. Austen's *Pride and Prejudice* (1813) was finished around 1797. John Locke's *Two Treatises of Government* (1689) was published almost a decade after it was originally written.

time required of those undertaking quantitative literary history may have deterred researchers who might otherwise have be interested in using quantitative methods. If this is the case, having an additional method that supports the task of data collection will encourage greater participation in the sociology of literature. The tens of thousands of novels published during the nineteenth century need not remain members of "the great unread."

4

Networks of Literary Production

> A novel is that sort of thing which should be very clever or not at all and notwithstanding the name of novel they are in a great measure copied from each other...
>
> Sir Walter Scott, Letter to Charlotte Pascoe. (Quoted in Garside (1991), 52–53.)

It is tempting to take the salience of novelistic genres—or market categories—such as gothic, silver fork, and historical novels for granted. In the case of gothic novels, the assumption that the category merits its status is underwritten by countless monographs, the explicit identification of novels as "gothic stories" by writers and publishers, and discussions of the novels as a category by contemporaries in journals, newspapers, and periodicals. A similar assumption might be made by researchers discussing "science fiction" today; the existence of a category bearing that name in bookstores and libraries in Shanghai, New York, Berlin, and São Paulo testifies that it is a category relevant to contemporary readers, writers, and publishers.

Writing in a special issue of *New Literary History* devoted to genre, Hayden White articulates a common response to empirical work in literary history, although his commentary has the virtue of being specific to discussions of genre. White argues

that the project of attempting to understand a number of artistic works according to an abstract model runs the risk of what the New Critics called "the heresy of paraphrase." The price of abstraction is a reduction of the original:

> Analysis of artworks guided by models inevitably condense, thin out, and reduce them to something like a paraphrase of a poem. What you get is not the "thick description" of a patch of text or speech that has been quoted in full—so that you can check it yourself—but a description which is allowed to stand in for the text and becomes the actual object of interpretation (White 2003b, 369).

White draws out the consequences of his characterization of model-driven work, noting that in order to judge the model in question one is "dependent upon the researcher's description" of individual artworks. White concedes, however, that the simplification implied by abstract models is often necessary in order to avoid being "lost in a chaos of details" (369).[1]

My claim in this chapter is that in the case of novelistic genres, it is possible to do without an assumption of hierarchically organized categories and that, moreover, doing without the "taxonomy" that White claims is indispensable becomes easier the more information one assembles about the cultural artifacts of interest. It is precisely the lack of details rather than their abundance that makes taxonomic approaches attractive.

Deliberations about available characterizations and careful demographic accounting of novelistic production make the assumption that the history of the novel merits

1. White's analogy between the classification of cultural artifacts and taxonomy in biology appears to be informed by an outdated conception of contemporary practice in biology. Rather than being preoccupied with reducing complexity, contemporary biological systematics (a field encompassing taxonomy) is more often concerned with complicating existing classifications of organisms. More generally, the assumption that abstract models impose a framework that is strictly and simply vertical—"downward from class to order to family" is how White puts it—mischaracterizes methods of classification in widespread use in the natural and quantitative social sciences. For an example of contemporary practice, see Schuh and Brower (2009).

study. Two examples of related research projects, both mentioned by Franco Moretti, will situate the principally methodological discussions in this chapter (Moretti 2000a, 2003b, 2008). First, with a more precise accounting of the flow of novels and literary forms among national and linguistic situations, historians would be in a better position to investigate why literary forms flourished outside their original national situation—examples include the Robinsonade (originated in England but flourished on the continent) and the naturalist novel (from France to Brazil). Understanding whether and how cultural forms flow across national and linguistic boundaries—including, potentially, from "center" to "periphery"—brings literary history into contact with persistent and long-standing debates in cultural and economic history. Second, a more precise record of the history of the novel is in a position to contribute to research about regional competitions for economic, political, and cultural preeminence. Moretti mentions the puzzle of how France maintained considerable cultural hegemony despite the military dominance of Britain in the nineteenth century. A careful accounting of the traffic of novels—of literary forms, physical copies, translations, and close copies of French novels published under a different titles—would be well positioned to address this question. This kind of question is not so far removed from present concerns. The "soft" power of cultural influence (in contrast to military power) is a persistent topic of discussion, particularly in East Asia during the last several decades (Nye 1990; McGray 2002).[2]

Quantitative methods and abstract models are required to navigate any significant portion of literary production after 1800. These in turn require critical reflection about the biases of the models chosen, what they highlight and what remains invari-

2. Examples of consideration given to cultural influence are not difficult to find. For instance, in Hu Jintao's report to the 17th Party Congress of the Communist Party of China he mentions "national cultural soft power" (国家文化软实力) explicitly: "要 坚持 社会主义 先进 文化 前进 方向，兴起 社会主义 文化 建设 新高潮， 激发全 民族文化 创造 活力， 提高 国家文化软实力， 使 人民 基本文化 权益 得到更好 保障， 使 社会 文化 生活 更加 丰富多彩， 使 人民 精神 风貌 更加 昂扬 向上。"

ant.[3]

4.1 "Taxonomy" Without Hierarchies

The World Wide Web consists of a vast collection of interlinked documents.[4] In 2005 the number of publicly accessible Web pages was put at more than 11.5 billion (Gulli and Signorini 2005). Search engines routinely facilitate the navigation of this immense collection. The most frequently used search engine by traffic (Google) works in part by representing pages in terms of "incoming" and "outgoing" links. That is, pages are represented in terms of the pages to which they link and from which they are linked. Pages are ranked according to the number of links the page receives and the ranking of the pages doing the linking (Brin and Page 1998). To say that this recursive definition of a page's rank has proved useful is an understatement; at the time of this writing, Google accounts for approximately six percent of all page views by internet users.[5] By "paraphrasing" Web pages in terms of their links, users of the Web are able to locate and read pages of interest. While there are many modifications made to the ranking procedure in practice, the core heuristic makes no classifications (such as flagging certain pages as authoritative) and imposes no hierarchy on the collection of web pages. Judging by its widespread use, this flattened representation of the Web manages the chaos of details of individual pages remarkably well.

In addition to ranking pages, the particular configuration of a network of links may also be of interest. Consider the study by Adamic and Glance of a thousand

3. Moretti (2000a) provides a succinct commentary on this point: "We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor. But it's precisely this 'poverty' that makes it possible to handle them, and therefore to know. This is why less is actually more" (57–58). In a footnote, Moretti continues: "Inevitably, the larger the field one wants to study, the greater the need for abstract 'instruments' capable of mastering empirical reality" (58n7).

4. "Documents" is used here generally. Images and video, among other media, also circulate on the Web.

5. On February 27th, 2013, Alexa reports that visits to google.com and its subdomains accounted for 5.794 percent of global page views. 71 percent of visitors used google.com (the search interface).

websites ("weblogs" or "blogs") devoted to commentary about politics in the United States (Adamic and Glance 2005). In February 2004 Adamic and Glance visited over a thousand blogs whose names they had gathered from numerous online weblog directories. Adamic and Glance cataloged each site in terms of outgoing and incoming links. They then created a visualization of these data; each node in the network (or graph) represents a website and a line joining two nodes stands for a link between the two websites. Their visualization is reproduced as figure 4.1.[6] While the visualization colors the blogs red or blue according to their classifications in the online directories, this hint is largely unnecessary; two highly-connected communities emerge from the network of links. It is easy to imagine simply omitting the coloring (and the label of right- or left-leaning) and assigning, if pressed to do so, the vast majority of blogs to "community 1" or "community 2" based on the structure of the network of links. However, such a division need not be made as the network stands on its own. Its visualization offer answers to many of the kinds of queries that presupposed the existence of two distinct groups in the first place. While careful attention needs to be paid to the assumptions implicit in any definition of a "community" in a network, the analysis of political blogs in the US suggests that it is possible to identify and examine groups of documents (or websites) without relying on pre-given hierarchies.

Documents on the Web have the virtue of making some of their relationships to other documents explicit with hyperlinks (i.e., "`<a href="...">link text</a>`"). Explicit links facilitate the task of collecting details about the network of relationships among documents. Hyperlinks are, however, far from the only way a text document provides information about its potential relationships with other text documents. For blogs, as well as for other texts, citations (without links), direct quotation, and paraphrase also provide provisional evidence of a relationship between

6. A similar visualization of a large collection of French political blogs can be found in Fouetillou (2006).

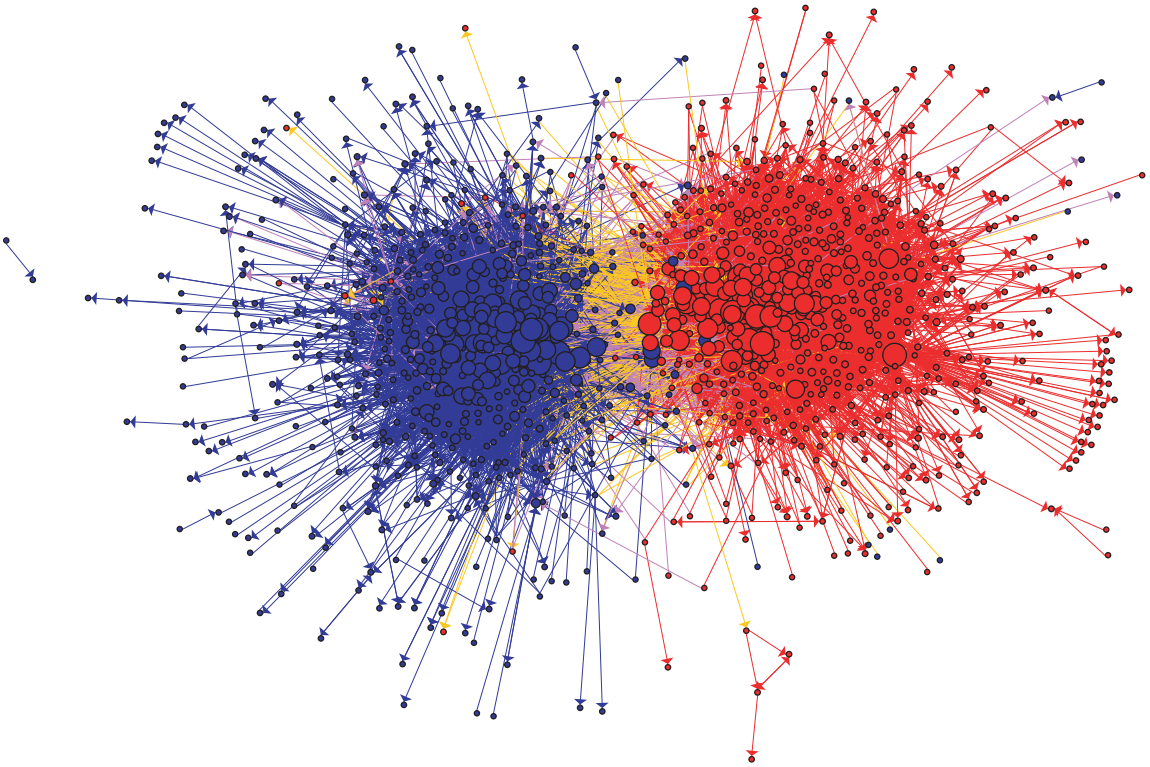


FIGURE 4.1: Visualization of community structure of United States political blogs during the 2004 presidential race. Figure appears in Adamic and Glance (2005).

two documents.[7] Information about the relationships among documents may also be supplemented by information about the relationship among writers. For example, correspondence networks among prominent figures in the eighteenth and nineteenth centuries are often well documented. Personal libraries and diaries may also suggest possible connections among authors and written works.

To bring the discussion still closer to the domain of literary history, consider the navigation of the following fictitious collection of twelve novels represented by the

7. Hyperlinks, citations, and direct paraphrase should not be taken at "face value"—i.e., as firm evidence of a relationship. Hyperlinks and citations can be and have been used in a variety of ways; they do not necessarily imply that there is a particular relationship between two documents—such as implying that the author(s) of a document had any familiarity with the text being cited. For example, a phenomenon labeled "coercive citation" has emerged in academic publishing in recent years. Because citation frequencies have been used in many rankings of academic journals, journal editors wishing to improve their journal's ranking have an incentive to see that articles appearing in the journal are cited as often as possible. This situation has been cited as an explanation for the phenomenon of article authors being prompted by journal editors to insert additional (superfluous) citations into their articles during revision (Wilhite and Fong 2012).

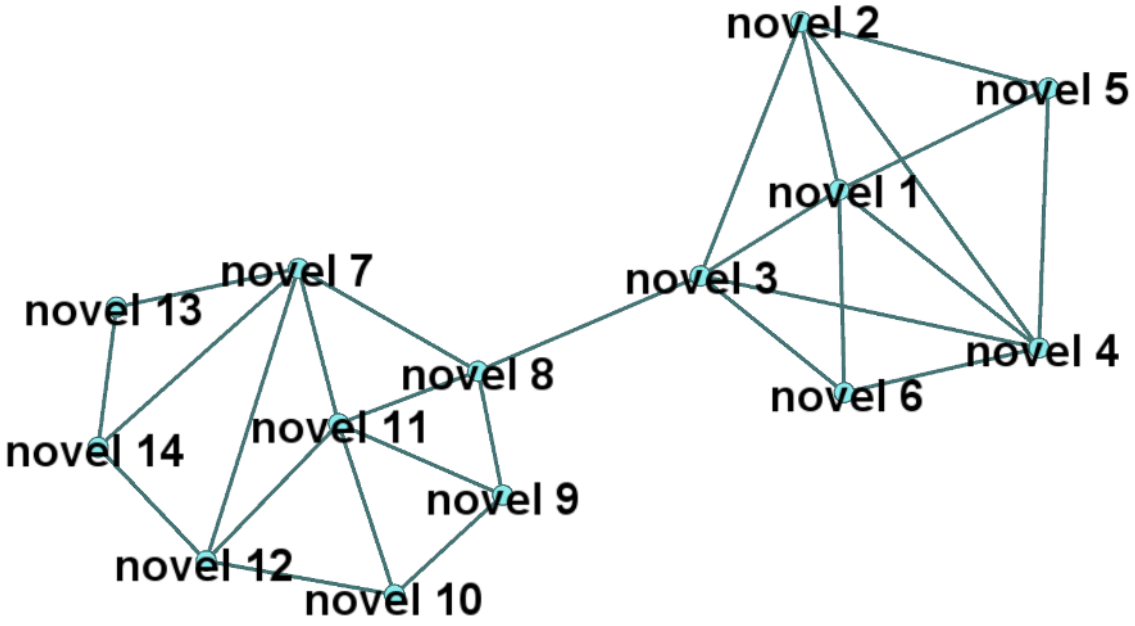


FIGURE 4.2: Visualization of relations among novels.

network in figure 4.2. Assume that connections ("edges") between pairs of novels are drawn when there is evidence of a connection, such as the presence of shared features. Features might include plot structure (episodic or not), shared publisher or author, or vocabulary (does the word "ghost" or "detective" occur in the novel more than some specified threshold). While it appears that there are two distinct "groups" of novels—and there exists a rich vocabulary to describe patterns of connectivity—whatever is implied by "group" is completely dependent on the relationships among novels (Easley and Kleinberg 2010). That this is indeed the case may be seen by imagining what would happen if an additional novel were added to the graph. Were the novel to connect disparate nodes, the division between the two subgraphs would become less distinct.

Proposing a network of relations among novels or other cultural artifacts becomes easier the more detailed the artifacts' descriptions are. While information about a

novel's publisher and author provide some indication of possible relationships with other novels, features derived from a novel's text permit speculation about a wider range of relationships. For example, a literary historian might explore the graph implied by connecting each novel in a corpus with an additional novel whose word or phrase frequencies are most similar—its "nearest neighbor."[8] Indeed, Latour et al. suggest that it is precisely in cases where details about entities are lacking that one is tempted to posit a "corporate whole." The organization of cultural artifacts into hierarchies, on this account, would be a response to the lack of a "chaos of details" rather than its presence (Latour et al. 2012, 593–94).

One argument for considering descriptions of novelistic production in terms of a network of relationships relies on the belief that the perception of salient traits of cultural artifacts depends in some manner on the perspective adopted. Justifying his "distant reading" of the history of the novel, Moretti observes that by focusing on the "small," such as the distribution of tropes across a range of novels, or on the very "large," such as genres and systems, different features of literary history come into focus, gaining one "sharper sense of their [elements of texts] overall interconnection" (Moretti 2000a, 57; 2005, 1). This point should be no more controversial than the observation that anyone who has walked around Washington, DC and also viewed an aerial photograph of the planned city stands to benefit from making connections between the two perspectives. That there are multiple valuable perspectives on a city or a collection of novels should not necessarily imply that one is privileged or even that there exists any sort of hierarchy among the different perspectives. Matthew Kirschenbaum makes a similar point in an interview discussing a project that involved studying the poetry of Emily Dickinson using quantitative methods. In the interview, Kirschenbaum speaks of the "rapid shuttling" between perspectives

8. Familiar metrics for similarity in this case include Jaccard distance, cosine similarity, or Kullbeck-Leibler divergence of novels' topic shares.

afforded by quantitative analysis and close reading (Hayles 2012, 31). Although it lacks the polemical force of "distant reading," Kirschenbaum's phrase captures better the sentiment that one is not necessarily closer or further away when one takes advantage of alternative perspectives of an ensemble of cultural artifacts. A graph of relations promises a richer description of a collection of entities than a simple partition into two categories. And examining a collection of novels in terms of a graph of relations facilitates asking certain kinds of questions, such as the position of novels within a community: Is a novel situated at the core or on the periphery of a tightly connected group?

A stronger argument for considering an analysis of literary history guided by the assumption of an underlying network of relationships is that different perspectives on literary production often bring with them—or are at least biased towards—specific explanatory strategies. For example, thinking of empirical regularities in literary production as influenced by interactions among individual writers, publishers, and texts—and representing these interactions as a network—complicates explanations that make reference to prevailing conditions or a "spirit of the times." On the other hand, attributing existence and autonomy to "atomistic" agents—e.g., writers, publishers, or texts—runs contrary to explanations emphasizing ways in which observed behavior is structured by environment and preexisting relationships.

What is meant by a "collective" is clearly bound up with how a collection of cultural artifacts is represented. Recall that Moretti (2005) defined novelistic genres as "morphological arrangements that *last* in time, but always only for *some* time" (14). And in identifying a diverse range of novelistic genres among novels published in the British Isles between 1740 and 1900 (forty-four in total), Moretti calls into question the idea that "'the' novel" is a single entity, suggesting that referring to such a thing obscures considerable diversity present in its (sub)genres. Shalizi (2006) responds to precisely this point, asking whether referring to a novelistic genre as a single entity

does not obscure the heterogeneity of the "shifting succession of individual texts" that are grouped together (Shalizi 2011, 120). This line of argumentation can be taken even further. Referring to an individual "novel" (such as *Castle Rackrent*) should not hinder any acknowledgment of the diversity among different editions of the novel. And for every edition there are likely a set of printings. Books associated with different printings may have readily identifiable typographical differences. And for each printing, physical copies, if they do survive, survive in a variety of conditions, with a variety of library stamps and marginalia. Regarding literary history as a changing network makes it easier to appreciate genres as contingent phenomenon. Adding novels to the graph may, depending on where their edges lie, reinforce, merge, or fragment existing communities.

An analogous situation is familiar from the philosophy of biology in questions about where biological collectives begin and end. Some interactions among a population of buffalo may be productively understood by thinking of the group as divided into distinct herds. In other cases it may be useful to consider the group as comprised of individual buffalo. In still other contexts, it may be important to focus on the micro-organisms and cells flowing between and inhabiting (or are they comprising?) "the" buffalo. In other cases, there may be a tendency to prematurely divide up an organism. A group of "quaking aspen" trees (*Populus tremuloides*) may appear at first glance to be comprised of distinct trees when in fact there is only one tree; the apparently separate organisms are connected by a shared root system from which all the "trees" originate (Godfrey-Smith 2009, 71).

To note that the idea of "the" novel should be called into question is not to argue that it is necessarily productive to do so. I stop with the idea of a novel—and in some cases, a single printing of a single edition—because novels published in the nineteenth century in Britain typically do not differ dramatically from one edition to another. For most novels the differences between editions and printings appear to be minor.

Opening paragraphs are usually identical, although differences in the front matter between editions are common. In the 93 novels under consideration in chapter 3, different printings and editions appeared to be quite similar. Principle, however, wars with necessity in this case, as for many of the 2,903 novels published between 1800 and 1836 in the British Isles, only a handful copies are readily accessible, making systematic research into the heterogeneity among editions or printings difficult.

Pace White, the analysis of artworks guided by models need not inevitably require a reduction of the phenomenon. Models of artworks, frequently described in quantitative terms, can add to—rather than subtract from—the resources available for the study of artistic works.

4.2 Social Networks of Readers and Writers

The remainder of this chapter is devoted to the practicalities and theoretical justification for studying a collection of novels—broadly the same group as was studied in chapter 3—in terms of an underlying network of relationships. Modeling relations among texts in novelistic production can begin with modest assumptions. That is, it is possible to approach the problem of identifying communities of connected novels—if such communities do indeed exist—in a manner analogous to that adopted by studies of political blogs in the United Sites. Having done so, it becomes possible to consider collective concepts such as novelistic genre in terms of the attributes of and connections among individual novels. This kind of project differs from the experiment described in chapter 3. The preoccupation in that chapter was on validating or corroborating the judgments of literary historians. This chapter proposes to do without static collectives like genres altogether and instead model a collection of novels in terms of the network of their relationships.

One place to begin thinking about relations among novels from the bottom-up—and to do justice, I think, to Moretti's aspiration towards a "materialist conception

of form"—is with a writer's encounter with other texts (Moretti 2005, 92). Writers are first readers; to regard literary production as a social process appropriately begins with the preconditions of writing. In many cases we have evidence of these encounters. Diary entries, private letters, commonplace books, memoirs, and other forms of self-report offer indications about writers' histories of reading. Personal libraries with copies of books (pages cut or uncut) may give some hint of contact with other written works. Other traces of contact include marginalia, teardrops, and stains left on pages. Direct quotation and paraphrase in a subsequent text can suggests that a writer has had contact with a work.[9] Defined narrowly to cover cases where a writer has had contact with a text, "influence" seems an appropriate designation.[10] While influences in this sense are potentially unknowable and in any case far more difficult to infer than the relation implied by the presence of a hyperlink in a webpage, they are attributes of a novel. Temporal constraints facilitate reasoning about potential influences on a text. Novels published in a given year tend not to be influenced by works published later. It is highly unlikely, for example, that Virginia Woolf's *Mrs Dalloway* (1925) influenced Charlotte Brontë's *Jane Eyre* (1847). Reliable information about publication date and the disposition of manuscripts limit the space of probable influences on novels.

Even this provisional outline suggests the obstacles facing such reconstruction. Diary entries, private letters, and other material that might testify to possible influences on a writer do not survive in great number. Many kinds of written texts that might have influenced writers in the nineteenth century themselves do not survive. Novels represent a minuscule fraction of written texts; they are available today be-

9. Direct quotation does not always indicate that a writer has actually read the source. Quotations may pass through any number of intermediaries—such as reviews and works by other writers.

10. The use of "writer" here is not made without an awareness of the work of Wimsatt and Monroe (1954) and Barthes (1967). Bourdieu proposes the following approach the question of intentionality: "It suffices to read literary memoirs, correspondence, personal diaries …in order to be convinced that …self-awareness, always partial, is yet again a matter of position and trajectory within the field, and that it thus varies according to agents and historical periods" (Bourdieu 1996, 272).

cause they were expensive and well-made in the first half of the nineteenth century. Moreover, even when a writer is explicitly preoccupied with a previously published work it is far from certain that this will result in anything detectable, such as paraphrase or quotation. "Readings happen inside people's heads" as Johnathan Frow and others remind us (Frow 2008, 141). No traces of influence, no heritable ensembles may exist to be found (Sperber 1996). With these challenges in mind, the project of "reconstructing the literary field" with such precision appears naively optimistic.

Hopes for this research admittedly hang entirely on the degree to which influences among novels can be detected. If paraphrase and quotation are endemic—and Walter Scott's observation quoted in the epigraph gives us some hope that they may be—then it may be possible to infer influences from the surviving texts themselves. Literary historians also offer considerable testimony to the fact that literary production became increasingly commodified in the nineteenth century, something familiar today in the "subliterary genres of mass culture" found in department stores and airports (Jameson 1981, 107).

4.2.1 Inferring Influences

While direct quotation and paraphrase appeal as indicators of influence because they are unlikely to arise by chance, unsupervised detection of paraphrase is an active area of research and requires considerable computational resources (Madnani, Tetreault, and Chodorow 2012). Quotation and paraphrase are not the only evidence available in the text of a novel that may indicate the influence of other novels. Similar word usage may appear a poor indicator of influence but there are situations where it seems likely to be valuable. In academic papers, for example, specific words and phrases—such as technical terms—may be strongly associated with specific papers and references to such papers may be inferred by subsequent papers that employ the same words (Gerrish and Blei 2010). In the context of literary production, certain

words seem likely to be useful indicators of a relationship among texts. Sequels or novels that have recurring characters or settings may use distinctive proper nouns. These words may be distinctive enough to justify the inference that one text influences another. Short epithets and formulae—such as αθηνά γλαυκώπις (bright-eyed Athena)—and other distinctive word combinations may similarly support inference of connections. Even single words may provide evidence of relationships among novels. For example, examining a random sample of novels published between 1800 and 1836 reveals that "trapdoor" occurs almost exclusively in gothic novels. Twentieth-century science fiction also exhibits distinctive words—such as astronomical (e.g., "terminator" and "perihelion") and technical terms introduced by one novel or short story and taken up by subsequent writers. As even identical phrases may occur in two texts by chance, corroboration of connections between writers or texts is welcome when it is available.

Having made the assumption that influences are detectable in novels features—an assumption to which I will return—the task of identifying likely instances of influences among novels remains. With the 2,903 novels published in the British Isles between 1800 and 1836 in mind, I assume that each novel is further characterized by attributes drawn from a finite "vocabulary" of attributes. These attributes would include all the words in the novel but could easily be extended to include indicators of paraphrase, plot structure, and so forth. A model aiming to infer influences would make the assumption that novels that share attributes are more likely to be related than those which do not share attributes (subject to chronological constraints mentioned a moment ago).

A simple model fitting this description would associate with each novel a distribution over features and relate the document-specific distributions according to an underlying graphical model. This arrangement may be summarized more prosaically. Assume that every novel has a number of "parents" by which it was influenced. These

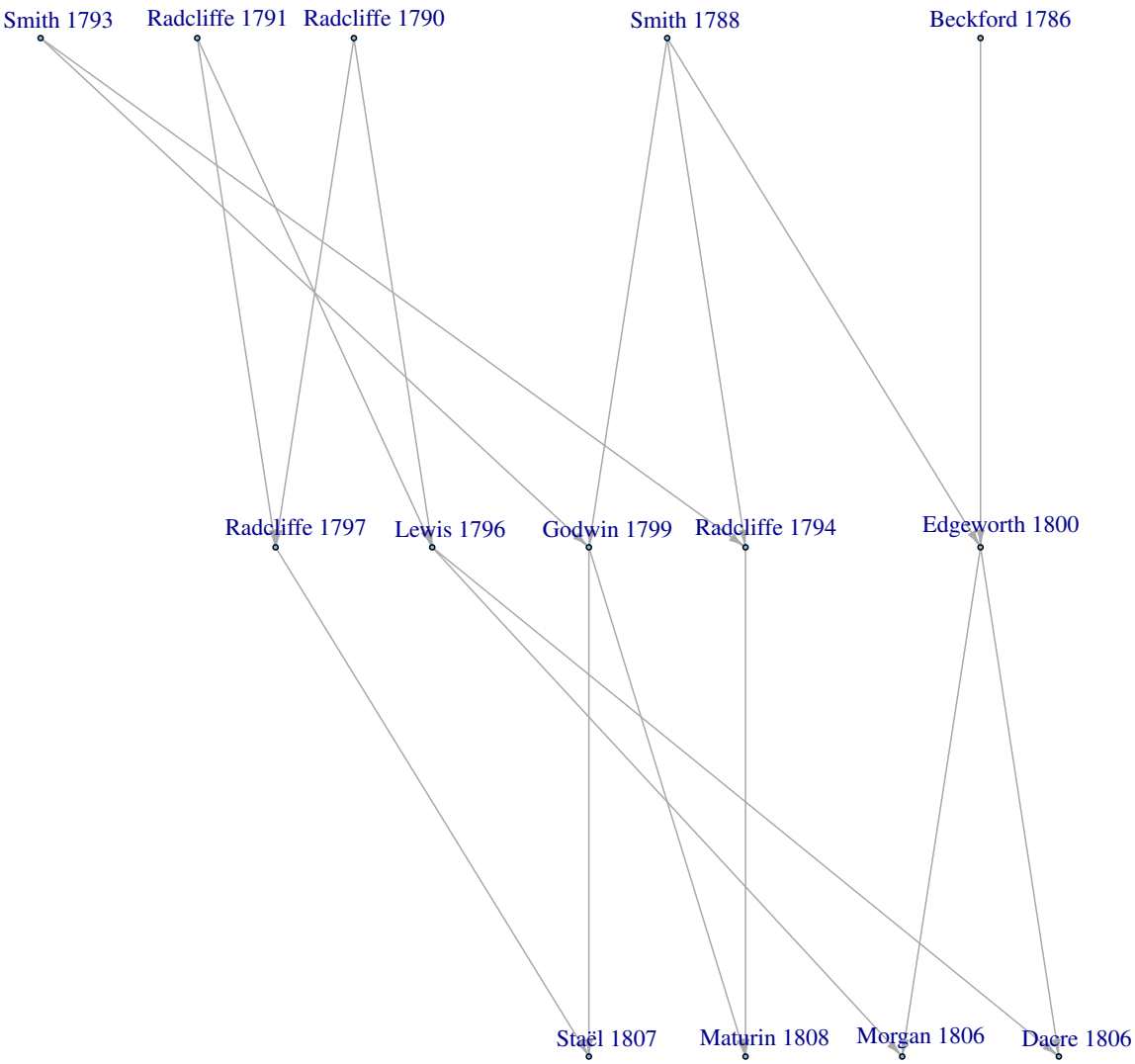


FIGURE 4.3: Randomly generated ancestral graph

influences are conveniently represented in an ancestral graph. (One randomly generated example of an ancestral graph is shown as figure 4.3.) Each novel need not be equally influenced by its parents; different degrees of influence may be expressed as weights on edges in the ancestral graph.

The computational challenge in identifying credible networks of influences arises from the fact that the number of possible networks (undirected graphs) of influence involving n texts is typically intractably large, $2^{\binom{n}{2}}$. For example, even if the universe of texts that might be influencing each other is limited to the 2,903 novels published

between 1800 and 1836, the number of possible ancestral networks is $10^{1268015}$. For comparison, there are estimated to be only 10^{80} atoms in the universe. Making additional assumptions may narrow the space of possible influences that need to be searched. For instance, it might be assumed that a novel is influenced by at most five or ten other texts. An additional solution adopted to the difficulty of searching through the space of possible relationships among a large number of entities is simply to start with a plausible initial guess at the graph of influences and evaluate as many nearby graphs as one can while staying within a given computational budget. These constraints are still compatible with the goal of finding candidate influences among some subset of novelistic production, a goal worth pursuing in light of how little is known about the vast majority of novels published in the nineteenth century. It is also worth recalling that the visualization of the collection of US political blogs is persuasive despite being made on the basis of explicit links; no attempt was made to identify other references—direct quotation and paraphrase surely among them—that are made without the use of hyperlinks.

4.2.2 Experiment

As a pedagogical experiment and illustration of the strategy described above, I will consider inferring a candidate network of influences among a small collection of nineteenth century novels familiar from chapter 3. Although this experiment uses a small corpus and a simplified model, components of the model such as the stochastic search across likely graphs and the underlying Gaussian graphical model are shared by more sophisticated models. This simplified model is also significantly easier to understand and to implement—posterior inference and calculation of the marginal probability of a model are straightforward as conjugate priors are used throughout—and may therefore be more useful as an invitation for literary historians to consider quantitative methods.

In response to the computational challenges mentioned above, this simplified experiment uses a topic model as a preprocessing step rather than as a proper piece of the model. Recall that a topic model of the corpus yields representations of each novel as a vector of topic proportions or shares. Typically these shares are either the empirical proportions of words assigned to the topics in a given document or samples from the posterior topic-document Dirichlet distribution. With these topic shares assumed to be known, the search for likely networks of influences can be made, again making the facile assumption that a novel that uses similar features as an earlier novel is more likely to be influenced by the earlier novel than one which does not use similar features. I have adopted a Gaussian vector autoregressive model as the underlying conceit of the model: each document (a vector of topic shares) is a linear combination of its parents plus a random error ("innovation error"). To keep the model as simple as possible, the parents' shares are averaged.[11]

The simplified model is described by the following equation:

$$y_t = \Phi \left(\frac{\sum_{i \in pa(t)} y_i}{|pa(t)|} \right) + \epsilon_t,$$

where y_t is a novel's vector of topic shares, $pa(t)$ is the set of parents of novel t, $|pa(t)|$ is the number parents of t and $\epsilon_t \sim N_k(0, \Sigma)$ is the random Gaussian innovation error. There are k topics and Φ is a kxk matrix of evolution coefficients.[12] Inference for Φ and Σ is familiar from vector autoregressive models—it is a special case of multivariate linear regression. These calculations are described in detail in

11. As mentioned above, it is be desirable in future developments to allow for the possibility that a novel borrows in very different proportions from its parents. A sequel, for example, might copy very heavily from the features of the first novel in the series and then copy, in lesser degrees, from other novels.

12. I will follow West and Harrison (1997) and transform each vector of proportions p_t to the log odds scale, $y_{tj} = \log(p_{tj}/\hat{p_t}) = \log(p_{tj}) - \log(\hat{p_t})$ where $\hat{p_t} = \prod_{j=1}^{K} p_{tj}^{1/K}$, the geometric mean of the topic proportions. To keep things on the log-odds scale, I model the observed vector y_t as arising from the average of its parents. (The average here being the geometric mean of the odds.)

appendix D.

The model above assumes that a graph of influences G is given while, in fact, G is not known. It is possible to reason about credible graphs of influence by using a conjugate prior on (Φ, Σ), which permits the conditional likelihood of the observed topic shares for the novels, $p(Y|G)$, to be calculated. Combining the likelihood $p(Y|G)$ with a prior distribution over possible graphs $p(G)$ allows the calculation of a posterior distribution over graphs $p(G|Y)$. Assuming a uniform prior distribution over graphs, the posterior distribution is proportional to the likelihood. In cases where the space of graphs is intractably large, an approximation of the posterior distribution can be made by evaluating a large number of candidate graphs and brazenly assuming the posterior probability of the unexamined graphs to be zero—*faute de mieux*. A stochastic search strategy may be used in which one starts at a graph that seems plausible by the standards of some simple heuristic, evaluates all neighboring graphs (those differing by one edge), and then moves to a new graph with probability proportional to its likelihood. Starting now with a new graph, the process continues. This strategy—a stochastic search—allows a great number of graphs to be evaluated and, in practice, typically settles on a local mode—a graph whose posterior probability is greater than all other graphs encountered in the search (Jones et al. 2005).

The strategy for finding plausible networks of influence among the novels may be described in less technical language. Millions of possible graphs of influence are proposed and a measure of their plausibility is calculated. Those graphs scoring highest by this measure are collected as a set of graphs of influence that, given the modeling assumptions, deserve consideration as descriptions of influences among novels in the corpus.

Corpus and Preprocessing

The corpus used is a subset of the corpus described in chapter 3. The novels in the corpus come from three familiar novelistic genres: gothic, national tale, and silver fork. Including novels associated with genres in the corpus provides a rudimentary check for the experiment: given the use of similar vocabulary between novels in the same genre, it would be surprising if the model did not infer connections among some of them. The corpus includes the following

- 19 silver fork novels from Adburgham (1983).
- 18 national tale novels from Trumpener (1998).
- 19 gothic novels from Lévy (1968).

The genre novels are a mixture of randomly selected novels and well-known novels associated with the genre. The randomly selected gothic novels are taken from the period 1815-1821 so as to enable comparison with the other two genres, both of which make their appearance only in the 1810s. The well-known novels are those used by Allison et al. (2011) in their attempt to infer novelistic genre based on frequent words and punctuation. The characteristics of novels associated with the genres are described in detail in chapter 3.

A simple topic model, Latent Dirichlet Allocation (LDA) is used as a preprocessing step to reduce the dimensionality of the data. As an initial step to verify that LDA stands a chance of preserving features of interest such as shared vocabulary, the novels were divided up into 1,000 line segments and fit with a standard LDA topic model with 100 topics (Blei, Ng, and Jordan 2003). As in chapter 3, the resulting topics do appear to capture salient features of the genres. Fitting the novel segments with a 100-topic model yielded, among others, the topics pictured in figure 4.4. Topic 41 is associated with the gothic novels and has words that literary historians would

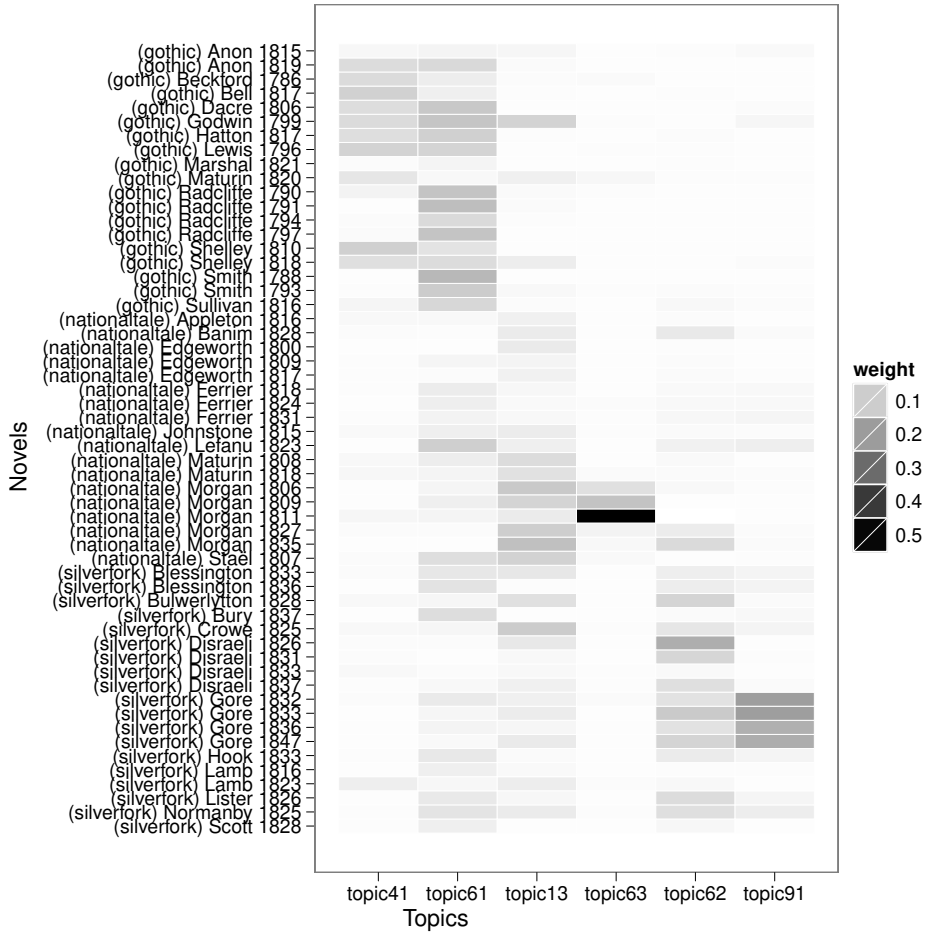


FIGURE 4.4: Gothic, national tale, and silver fork novels in terms of selected topics.

recognize as characteristic: "death," "body," "blood," "power," "crime," "dreadful," "escape." One of the topics characteristic of national tale novels does have words that one anticipates, such as "human," "missionary," "religion," "eyes," "heaven," and "religious" (topic 63). And one of the topics does match expectations for the silver fork novels, with words such as "lord," "party," "dinner," and "london" (topic 62).

In order to facilitate computation, the number of topics was reduced further and the corpus of 56 novels was refit using a fifteen-topic LDA model. Each novel now has a representation as a vector of fifteen topic shares. These shares may be used as inputs to the model described in section 4.2.2.

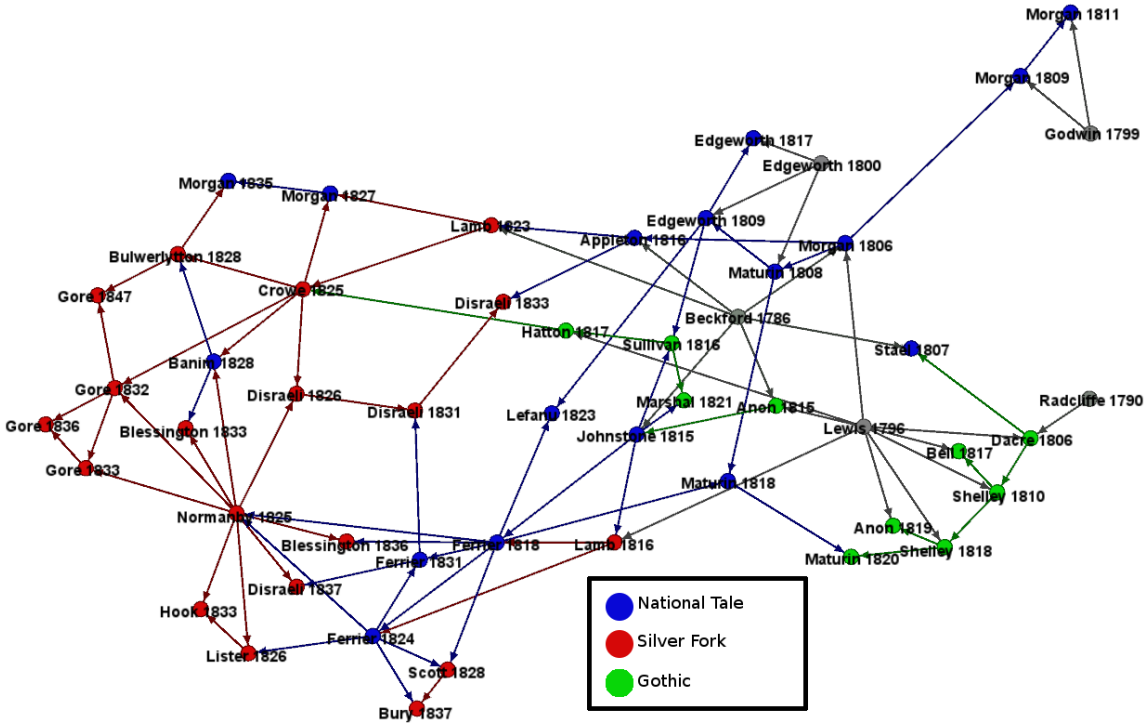


FIGURE 4.5: Modal graph for the experiment. Genres (silverfork, gothic, national tale) are reflected by label colors. The initial novels are given the color gray. Several of the initial ten novels were not connected to any other novels and are not pictured.

Results

Novelistic production must start somewhere, so in this experiment it begins with the first ten novels (in chronological order). Since there are no earlier novels in the corpus, these ten novels are assumed to have no parents and serve as the seeds for subsequent literary production. Even with this small corpus, the number of possible graphs of influence is large, so a stochastic search for plausible graphs is used. The modal graph found after this search is shown as figure 4.5.

That novels judged by literary historians to belong to the same genre tend to be connected should come as no surprise given the results of chapter 3, in which topic shares were found to align better than chance with expert classifications.

This pedagogical experiment provides a general approach that would be suitable for use on an expanded corpus. Even at this preliminary stage it is possible to

appreciate the expressiveness of the approach over the representation of novels solely in terms of topic shares. For example, time has an expression in the directed graph: arrows indicating inferred influences of novels published earlier upon those published (and presumably written) later.

In an expanded test the preprocessing step would be skipped and a topic model of novels' features would be inferred at the same time as the underlying graphical model.[13] Cases where literary historians have previously identified influences could be used to assess the plausibility of the inferred graphs of influences.

Were the graph of influences considered credible, it is not difficult to imagine how it might be put to use in literary historical research. Hans Jauss proposes a criterion for aesthetic value that looks to the distance of a work from the "horizon of expectations" (Jauss 1989). Novels breaking with convention have ready quantitative operationalizations given a graph of influences. They would be those novels difficult to connect to immediate predecessors, or those novels about which there was the greatest uncertainty as to their position in the credible set of graphs. Novels at the interstices of tightly-knit communities of novels might also merit interest, being situated at something akin to a "structural hole" in literary production (Burt 2004).

4.3 Convergent Influences

A fuller evaluation of the prospects for this kind of exploration of novelistic production awaits a larger corpus than can currently be assembled. While it is true that a majority of novels published between 1800 and 1836 have been digitized and are publicly accessible, these novels are not evenly distributed over the time period: novelistic production expanded dramatically over the course of the nineteenth century and novels published after 1815 survive at a greater rate than novels published before

13. The logistic normal distribution would be an appropriate substitute for the Dirichlet in this case. For examples of the use of the former, see Blei and Lafferty (2006) and Mimno (2012b).

1815. Given how heavily the model proposed in this chapter relies on the detection of prior influences, the absence of these texts makes assembling an appropriate corpus difficult. Improvements in optical character recognition (OCR) techniques and progress in scanning (and, inevitably, rescanning) surviving print material from the eighteenth and nineteenth century continue apace.[14]

Assuming that a broad corpus of novelistic production could be assembled, augmented with a large corpus of contemporary print materials, three significant theoretical challenges remain to be addressed. The first challenge concerns printed texts that do not survive. For every detected instance of similar vocabulary use, paraphrase, or direct quotation, there may be others that go undetected. For example, a historical novel may borrow heavily from material appearing in some issue of a periodical that does not survive. The model proposed does not allow for the possibility that certain kinds of novels may be disproportionately influenced by printed material that do not survive. The second challenge is related to the first and emerges from the fact that, even if novelistic production were nothing but paraphrase and direct quotation and conformed precisely to the Bourdivin assumptions, the knowledge that much printed material does not survive makes the graph of influences inferred only a sample of a larger unknown graph. Random sampling from graphs brings with it a range of practical and theoretical problems. The properties of the sample graph can depend strongly on the sampling procedure used (Kolaczyk 2009, 123–51). And the sample of novels and other printed material available will be anything but random. Any collection of nineteenth-century texts will be strongly biased by what libraries, individuals, and organizations were selected for preservation. Considerably caution

14. Many digitizations of novels encountered during this research will need to be rescanned. The most significant source of digitizations of British novels published between 1798 and 1834 is the Corvey library, located near Höxter in Nordrhein-Westfalen. Digitizations of novels in the Corvey collection do not, however, appear to have been made from the surviving novels themselves. Rather they are scans of the Corvey Microfiche Edition, work on which began in the late 1980s ("Corvey Introduction" 2013). The facsimiles recorded on microfiche vary considerably in their fidelity and scanning these facsimiles, of course, cannot make the text of the novels any more readable.

is therefore warranted in interpreting any proposed graph of influences. The third challenge concerns the treatment of convergent influences, something analogous to convergent evolution in biology. The model as described makes no allowance for texts that exhibit the same features but do so for different reasons. Two novels may use similar vocabulary or paraphrase the "same" material but do so via different intermediary chains of influence. Such an outcome is not difficult to imagine. Take the example of the distinctive use of the opening of a novel: "It is a truth universally acknowledged, that ..." (an example which has the virtue of occurring in the present corpus in a novel not by Jane Austen, *The Inheritance* (1824) by Susan Ferrier). Suppose this manner of beginning a sentence (or a novel) continues to be copied. If at some subsequent point, perhaps a century later, two novels are identified that both begin with the sentence but do so as the result of two writers copying from two different sources–assume the writers are ignorant of Austen's *Pride and Prejudice*—then these instances of paraphrase need to be distinguished somehow in the model, just as biologists need to distinguish between genetically unrelated birds that nonetheless exhibit highly similar morphology—e.g., flamingos and Roseate Spoonbills. Considerable research in comparative biology exists addressing precisely this modeling problem and it seems likely that models developed in that field may be able to be adapted for the study of literary production (Harvey and Pagel 1991; Rogers, Feldman, and Ehrlich 2009; Tëmkin and Eldredge 2007).

Even if these challenges reduce the promise of model-driven inquiry into the network of influences among texts to a kind of hypertrophied exploratory data analysis, considerable value would remain in the exercise. Having an additional means of navigating similarities among tens of thousands of surviving novels would support research in literary and publishing history that has a limited range of methods for exploring large archives of printed materials . Moreover, the modeling of novels as a network of influences embedded in time, even if that network is partial and incom-

plete, offers an additional illustration of the possibility of thinking about novelistic production without the use of hierarchies and external classifications. Consideration of appropriate methods for modeling the collection of surviving nineteenth-century texts need not wait until the text collection is fully assembled.

4.4 "Readings happen inside people's heads"

The most trenchant criticism of the attempt to reconstruct anything resembling a network of influences among novels—however such influence is defined—is that there is nothing in one text that can influence another because reading and writing involve humans. Reading and writing are two activities that transform representations of texts. "Reading happens inside people's heads," as Frow puts it. Or, as Dan Sperber, argues, instances of "replication" are rare for a reason: cultural artifacts are transformed in the process of reception and reproduction. In the case of novels, readers and writers ineluctably participate in the reproduction of texts:

> Constructive cognitive processes are involved both in representing cultural inputs and in producing public outputs. All outputs of individual mental processes are influenced by past outputs. Few outputs are mere copies of past inputs (Sperber 1996, 118).

That readers in particular ought to be thought of as active participants in the reception and circulation of texts is not particularly controversial. Researchers from a variety of disciplinary perspectives concede that phylogenetic models developed in biology cannot simply be imported for the analysis of cultural artifacts. Doing so would be to assume an inappropriate causal explanation for regularities observed in cultural artifacts. Tëmkin and Eldredge observe that "[w]hile it is tempting to attribute the patterns we discover in culture to the same causal processes that operate in nature, cultural systems present greater complexity than their biological

counterparts …" Tëmkin and Eldredge also emphasize that successful phylogenetic studies of cultural artifacts—projectile points and textile patterns are mentioned—tend to study artifacts from pre-industrial societies, where "traditional transmission is strong but intercultural exchange is relatively weak" (Tëmkin and Eldredge 2007, 151). Neither of these two descriptions fit nineteenth-century literary production.

Sperber does not assert that direct copying never happens—he mentions the duplication of medieval manuscripts—rather he argues that such identifiable replication is rare. Assessing Sperber's claim applied to novelistic production is hampered by the lack of scholarly interest in (and access to) the vast majority of surviving novels and the absence of computational resources required to churn through large collections of texts looking for direct paraphrase and quotation. The wager of the research program proposed in this chapter is that copying—in some form—is prevalent and that in many cases it can be identified, given a large enough collection of relevant scanned materials, improved OCR, and models like the one outlined in this chapter.

Against this optimism, Sperber would argue that humans, for a variety of reasons, "tend to generally exaggerate the similarity of cultural tokens and the distinctiveness of types" (Sperber 1996, 118). Sperber's caution is important. Returning to the Austen/Ferrier example, I made no systematic evaluation of how rare it is for a novel to begin with the sentence "It is a truth universally acknowledged" before settling on the belief it was unlikely to have been a coincidence. In retrospect, this judgment appears hasty; a cursory search reveals that the phrase "It is a truth universally acknowledged" occurs in a variety of early nineteenth-century texts—e.g., page 320 of Rollin (1804). Weighed alongside such skepticism, however, should be the fact that Ferrier expresses admiration for Jane Austen in her correspondence (*Ferrier, Susan Edmonstone (1782–1854)* 2004).

The virtue of research on nineteenth-century novels is that there is no lack of opportunities to experimentally test models aiming to detect influence. For hundreds

of authors, documentation exists in their correspondence and other sources that can suggest novels with which they had contact. As James English notes, "No other form of cultural practice has been as thoroughly subjected to academic scrutiny, as written about by scholars, or as widely promoted and disseminated by the educational apparatus as literature has" (English 2008, 126).

A different sort of skepticism about the endeavor described in this chapter asks whether attention given to patterns of influence can be reconciled with existing theories of (novelistic) genre. Does this inquiry make contact with accounts of genre as symbolic expressions of historical experience—such as that found in Fredric Jameson's *The Political Unconscious*—or as imaginary solutions to real social conflicts (White 2003a, 603). Where in a historical network of relations does space remain for the idea that a genre may offer a way of expressing certain ideas—for example, that the realistic novel was an expression of the "discovery that society was not only, or even primarily, tradition, consensus, and continuity but also conflict, revolution, and change" (White 1999, 22–23).

The compatibility of existing accounts of genre with the orientation towards literary production described in this chapter remains an open question. The working concept of genre I have had in mind is not that of Jameson but of Moretti, one in which market categories such as silver fork and nautical tales receive attention and one that begins with the more minimal starting point of recurrent "morphological arrangements." This chapter is intended to explore models that make assumptions about the ways similar morphological arrangements arise, persist, and dissolve. I find compelling the way the bottom-up study of novelistic production, in which individual editions and printings of novels figure prominently, brings attention to the heterogeneity of literary works, especially among those assigned to the same category by literary historians. The wager of this chapter is that there may be distinctive trajectories within genres. If the tradition has been to speak of the symbolic expression

of a genre, then attention to the heterogeneity within the genre—made practical by the availability of surviving literary works—may reveal that that expression is not singular but multiple.

A range of existing work on novelistic genres is more immediately compatible with the approach proposed in this chapter. Thinking of a genre primarily as a network of connected writers, readers, publishers, and texts can supplement existing accounts that focus on social circles, geography, class, gender, and nationality. These concerns occur in existing histories, including Adburgham (1983), Hollingsworth (1963), and Tuchman (1989). Elaborating a network of relations among those connected with a group of novels sharing common features would, I suspect, bring into sharper relief aspects of the history of a novelistic genre that have already been well established, as well as identifying connections that may have been overlooked.

4.5 Conclusion

The goal of this chapter has been to introduce an alternative perspective on regularities observed in novelistic production. Positing or inferring a network of relations—whatever their ontological valence—permits the navigation of novelistic production without presupposing the existence of categories like novelistic genre. If densely connected groups of novels emerge—just as two dominant communities emerge in studies of online political debate in the US—it may be convenient to refer to novels as members of a group. Yet "genre" in this sense implies nothing fixed, as genres are now defined purely in terms of connections among individual novels in a graph. Adjust one edge in the graph and the community is no longer the same. This is not the hierarchically ordered "genre" that White describes. Its members are not defined by static (or slowly changing) characteristics. If the desire to talk about the genre as an attribute of a novel persists, then even this way of talking can be nuanced by insisting that novels share an attribute modified by each sharing—something fittingly

reminiscent of the recursive definition of rank used by the search engine.[15]

15. This formulation is an adaptation of "monads share attributes modified by each sharing" (Latour et al. 2012, 609).

5

Conclusion

In this thesis I identify a number of cases where quantitative methods are positioned to address existing questions in literary history and sociology of literature. In chapter 3 I find that an abstract model prompts valuable questions about what literary historians mean by "novelistic genre" and demonstrate that a probabilistic model can reproduce the genre classifications of literary historians better than chance. Chapter 4 locates the conceptual and computational resources for a study of the tens of thousands of surviving novels that foregrounds social networks of authors and publishers.

There is considerable demand in the social sciences and humanities for methods suited for working with large collections of texts. Such work is growing more common as more surviving texts are made accessible by library digitization. For instance, researchers in cultural, literary, and intellectual history are beginning to check certain kinds of claims against the record of surviving print materials. Sometimes this can be done with keyword searches but in other cases the procedures are more involved. This dissertation identifies methods that are useful in a number of contexts and participates in discussions of how to use them skillfully. The theoretical contribution of

the dissertation is to demonstrate that (probabilistic) abstraction is able to bring into greater relief features of cultural artifacts that are relevant to historical scholarship.

That the empirical findings in chapters 3 and 4 could not be stated more firmly is due to the present state of library digitization and the small size of the random sample assembled. Many novels published before 1830 that are held in the collections of libraries in the United States appear to have been passed over by the initial wave of library scanning initiatives. Having tracked down a number of these novels, I can report that many reside in rare book collections and simply have yet to be scanned. The early days of library digitization were rough affairs; books were often transported off-site and then returned to a library. That a rare novel—of which there are often only two or three exemplars in the United States and Europe—might have been excluded from this handling makes sense. As more and more libraries bring book digitization equipment under their own roofs, the scanning of special collections seems likely to take place. The work of proposing and evaluating models of literary production can make little real progress until a substantially larger corpus of surviving volumes is assembled.

Had this thesis made use of all the surviving novels out of the 2,903 works published between 1800 and 1836 in the British Isles, there would still be lingering questions about the relevance of literary works published elsewhere, including those appearing in serialized form and those published outside the British Isles. The influence of novels published in France but which circulated across the English Channel is, for example, left unaddressed. There are also modeling choices that should be improved. The most ambitious model, put forward in chapter 4, highlights relations among novels. While this is preferable to focusing narrowly on features of texts, it would be better to model explicitly the relations among authors and publishers.[1]

1. On the other hand, modeling the novel itself as a freestanding entity in a network may have its own appeal for those who intend for the adjective "social" to include connections among non-human entities (Latour et al. 2012).

Doing so would better fit the belief that novels arise out of evolving communities of readers, writers, and publishers.

Appendix A

A Simple Topic Model

Topic models typically start with two banal assumptions. The first is that in a large collection of texts there exist a number of distinct groups (or topics) of texts. In the case of academic journal articles, these groups might be associated with different journals, authors, research subfields, or publication periods (e.g., the 1950s and 1980s). The second assumption is that texts from different groups tend to use different vocabulary. If we are presented with an article selected from one of two different academic journals, one dealing with literature and another with archeology, and we are told only that the word "plot" appears frequently in the article, we would be wise to guess the article comes from the literary studies journal.[1]

The following description of a simplified topic model (mixture of unigrams) is addressed to an audience with some background in probability and statistics, perhaps at the level of the introductory texts of Hoff (2009), Lee (2004), or Kruschke (2010).

1. Both these assumptions are inaccurate. Each article in a collection is different and every book in a library is unique—even books that are "copies" in the sense of being the same edition or from the same printing are visibly different under a microscope (although usually one need not go that far). There are no shared "sources" of texts. And every printed word is similarly unique, often visibly so if different fonts have been used; this challenges the idea of a fixed vocabulary. At their best, models are useful fictions.

The mixture of unigrams model is a close relative of Latent Dirichlet Allocation (LDA), the topic model used in chapter 2. It is, however, a less nuanced model than LDA and does not model as well the polysemy pervasive in human language.

To keep the discussion concrete, I will consider a corpus of twenty academic journal articles drawn from a larger collection German Studies journal articles. To simplify matters further, I will pretend these articles make use of a vocabulary of only eight words. (The articles have been selected so that, were their titles known, they do fall easily into two groups.) The corpus is shown as table A.1. I will show how a probabilistic model of the texts, starting from the assumption that there are a fixed number of groups, can infer (1) which documents belong to which group and (2) what words are most strongly associated with each group.

Let us assume that there are two groups ($K = 2$). We know that there are twenty articles ($N = 20$) and that each article consists of n_i words ($n_1 = 11, n_2 = 17, \ldots, n_{20} = 28$) drawn from a vocabulary of eight unique words ($V = 8$). Before considering the word frequencies in the corpus, we first specify our prior beliefs in keeping with the ideas outlined in the opening paragraph. There are three assumptions. First, if we knew which group (or topic) a document came from, then we would anticipate that words from that document are those likely to be found in other documents associated with the group. Second, since we do not have any information about the documents in advance, we will say that it is equally likely that a document comes from topic one or topic two. Finally, since we have no information about what vocabulary is associated with what topic, we will say that each word is equally likely to appear in documents associated with either topic. We can write this with symbols as

Table A.1: Word frequencies in twenty German Studies journal articles (selected words).

	literary	literature	authors	century	texts	writers	economic	critique
Article 1	0	0	0	4	0	0	2	5
Article 2	0	0	0	0	0	0	6	11
Article 3	0	0	0	3	0	0	8	0
Article 4	0	0	0	2	1	0	6	16
Article 5	0	1	0	5	1	0	3	13
Article 6	0	0	0	0	0	0	5	6
Article 7	10	3	0	4	0	1	0	0
Article 8	13	1	7	0	0	5	0	0
Article 9	7	3	0	4	1	8	0	0
Article 10	20	14	3	0	0	0	0	0
Article 11	5	6	5	0	0	10	0	0
Article 12	9	7	0	2	0	1	0	0
Article 13	3	5	3	0	0	6	0	0
Article 14	8	13	3	1	1	3	0	0
Article 15	9	3	4	0	0	6	0	0
Article 16	11	7	4	0	1	6	0	0
Article 17	2	3	0	1	1	1	0	0
Article 18	5	2	13	0	0	5	0	0
Article 19	7	3	6	1	0	11	0	0
Article 20	5	9	8	2	0	4	0	0

$$w_{ij}|z_i \overset{i.i.d.}{\sim} Multinomial(1, \phi_{z_i})\ j = 1, \ldots, n_i$$

$$z_{1:N} \overset{i.i.d.}{\sim} Multinomial(1, \theta)$$

$$\theta \sim Dirichlet(\alpha_{1:K})$$

$$\phi_{1:K} \overset{i.i.d.}{\sim} Dirichlet(\beta_{1:V})$$

where w_{ij} is the jth word of document i. z_i indicates the topic that document i is associated with (here either 1 or 2). $\alpha_{1:K} = (1, 1)$ and $\beta_{1:V} = (1, 1, 1, 1, 1, 1, 1, 1)$ are the parameters for the two Dirichlet distributions which express the prior beliefs described (note that $Dirichlet(1, 1)$ is equivalent to a $Beta(1, 1)$ distribution). The following table gives a summary of notation,

w_{ij}	jth word of document i
z_i	topic of document i
n_i	number of words in document i
α	parameter for document-topic Dirichlet
β	parameter for topic-word Dirichlet
N	number of documents
V	number of unique words (vocabulary)

How should our beliefs change once we see the words contained in each article? There are three inferential moves that, when combined, give us an answer. Making each move in succession will eventually yield an updated representation of our beliefs. These moves are easy to explain in English and the details only require familiarity with the Multinomial distribution and its conjugate prior, the Dirichlet distribution—the pair being the multivariate analog of the Binomial distribution and its conjugate prior, the Beta distribution. The first inferential move begins with an assumption. We assume that we know which documents are associated with which topics and update our beliefs about how words are associated with each topic. (Remarkably, it does not matter what topic assignments we start with.) Imagine that we have guessed the following topic assignments: the first ten articles are from topic one, the remaining ten are from topic two,

$z = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$. If these were the true topic assignments, how would we adjust our beliefs about topic one? Glancing at the table of word frequencies, I can at least say that the word "literary" should be more strongly associated with topic two than with topic one since it occurs in all of the ten articles (100%) assigned to topic two whereas it occurs in only four of ten articles (30%) assigned to topic one. The calculation of the probability of a word being associated with a group has the following expression

$$
\begin{aligned}
p(\phi_{1:K}|z_{1:N}, w) &\propto p(w|\phi_{1:K}, z_{1:N})p(\phi_{1:K}) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{n_i} Multinomial(w_{ij}|\phi_{z_i}) \times \prod_{k=1}^{K} Dirichlet(\phi_k|\beta_{1:V}) \\
&\propto \prod_{i=1}^{N}\prod_{v=1}^{V} \phi_{z_i,v}^{f_{i,v}} \times \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1} \\
&= \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{e_{k,v}} \times \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1} \\
&= \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{\beta_v + e_{k,v} - 1} \\
p(\phi_{1:K}|z_{1:N}, w) &= \prod_{k=1}^{K} Dirichlet(\phi_k|\beta_1 + e_{k,1}, \ldots, \beta_V + e_{k,V})
\end{aligned}
$$

where $f_{i,v}$ is the number of times word v appears in document i and $e_{k,v}$ is the number of times word v is assigned to topic k across all documents.

The second move swaps the position of our ignorance. Now we guess which documents are associated with which topics, making the assumption that we know both the makeup of each topic distribution and the overall prevalence of topics in the corpus. If we continue with our example from the previous paragraph, in which we

had guessed that "literary" was more strongly associated with topic two than topic one, we would likely guess that the seventh article, with ten occurrences of the word "literary," is associated with topic two rather than topic one. This would change our topic assignment vector to $z = (1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2)$. We take each article in turn and guess a new topic assignment (in many cases it will keep its existing assignment). The calculation of the probability of a document being assigned to group 1 or group 2 may be calculated with the following expression

$$
\begin{aligned}
p(z_i = k|w, \phi_{1:K}, \theta) &\propto p(w|z_i = k, \phi_{1:K})p(z_i = k|\theta) \\
&\propto \{\prod_{j=1}^{n_i} p(w_{ij}|z_i = k, \phi_{1:K})\}p(z_i = k|\theta) \\
&= \{\prod_{v=1}^{V} \phi_{k,v}^{f_{i,v}}\}\theta_k \\
p(z_i|w, \phi_{1:K}, \theta) &= Multinomial(z_i|1, q^{(i)})
\end{aligned}
$$

where $q^{(i)} \propto (\theta_1 \prod_{v=1}^{V} \phi_1^{f_{i,v}}, \ldots \theta_K \prod_{v=1}^{V} \phi_K^{f_{i,v}})$.

Finally, we update our guess about the overall prevalence of topics in the corpus—are 80% of articles topic one or are 20%?—assuming that we know the topic assignments (which we have just guessed). The reasoning here is straightforward: if 80% of the articles are assigned to topic one and 20% to topic two, then the true proportions are likely in the vicinity of 80% and 20%. In symbols,

$$
\begin{aligned}
p(\theta|z_{1:N}) &\propto p(z_{1:N}|\theta)p(\theta) \\
&= \prod_{i=1}^{N} p(z_i|\theta) \prod_{k=1}^{K} Dirichlet(\theta_k|\alpha) \\
&\propto \prod_{k=1}^{K} \theta_k^{d_k} \prod_{k=1}^{K} \theta_k^{\alpha-1} \\
&= \prod_{k=1}^{K} \theta_k^{\alpha+d_k-1} \\
p(\theta|z_{1:N}) &= \prod_{k=1}^{K} Dirichlet(\theta_k|\alpha_1 + d_1, \ldots, \alpha_K + d_K)
\end{aligned}
$$

where d_k equals the number of documents assigned to topic k.

Making these moves in succession over and over is an instance of *Gibbs sampling* and eventually the topic assignments and the specification of each topic multinomial distribution will converge to a representation reflecting what, given our prior beliefs and the articles' word frequencies, our updated beliefs ought to be.[2] And even if we had never encountered Gibbs sampling before, it is clear that making these inferential moves in succession leads to more plausible topic assignments (i.e., those documents containing "literary" end up in their own category). Table A.2 shows the topic assignments, on average, for each document after 500 iterations (ignoring the first 100).[3] Table A.3 shows the most probable under each topic distribution, on average.

The model indicates that the articles come from two different groups. This is indeed the case. The first six articles are from the early years of the journal *New*

2. For an introduction to Gibbs sampling see chapter six of Hoff (2009). Other introductions include Resnik and Hardisty (2010) and Casella and George (1992).

3. The problem of identifiability is sidestepped here by imposing the constraint that at ever iteration the group with the smallest number of articles is the first group.

Table A.2: Group assignments for the twenty-article corpus.

Docs	Topic 1	Topic 2
Article 1	0.04	0.96
Article 2	0.04	0.96
Article 3	0.04	0.96
Article 4	0.04	0.96
Article 5	0.04	0.96
Article 6	0.04	0.96
Article 7	0.96	0.04
Article 8	0.96	0.04
Article 9	0.96	0.04
Article 10	0.96	0.04
Article 11	0.96	0.04
Article 12	0.96	0.04
Article 13	0.96	0.04
Article 14	0.96	0.04
Article 15	0.96	0.04
Article 16	0.96	0.04
Article 17	0.96	0.04
Article 18	0.96	0.04
Article 19	0.96	0.04
Article 20	0.96	0.04

Table A.3: Characteristic words for each group in the twenty-article corpus.

Topic 1	literary	literature	writers	authors	century
Topic 2	critique	economic	century	literature	texts

German Critique and the remaining fourteen articles focus on German literature. A list of articles follows.

1. Karl Korsch. "The Crisis of Marxism." *New German Critique.* Autumn, 1974

2. Rainer Paris. "Class Structure and Legitimatory Public Sphere: A Hypothesis on the Continued Existence of Class Relationships and the Problem of Legitimation in Transitional Societies." *New German Critique.* Spring, 1975

3. Herbert Marcuse. "The Failure of the New Left?." *New German Critique.* Autumn, 1979

4. Paul Piccone. "Karl Korsch o el Nacimiento de una Nueva Epoca." *New German Critique.* Autumn, 1975

5. Paul Piccone. "From Tragedy to Farce: The Return of Critical Theory." *New German Critique.* Winter, 1976

6. Peter Laska. "A Note on Habermas and the Labor Theory of Value." *New German Critique.* Autumn, 1974

7. Leland R. Phelps. "The Emergence of German as a Literary Language." *Monatshefte.* Apr. - May, 1960

8. Andreas Kiryakakis. "Dictionary of Literary Biography: Volume 66: German Fiction Writers, 1885-1913 Part I: A-L." *German Studies Review.* May, 1990

9. Marianne Henn. "Benedikte Naubert (1756-1819) and Her Relations to English Culture." *The German Quarterly.* Fall, 2006

10. Stephen Brockmann. "German Literature of the 1990s and Beyond: Normalization and the Berlin Republic." *Monatshefte.* Summer, 2006

11. Willa Schmidt. "German Fiction Writers, 1885-1913." *Monatshefte.* Spring, 1993

12. Dieter Cunz. "Pennsylvania German Literature (Changing Trends from 1683 to 1942)." *The German Quarterly.* Mar., 1945

13. Helga Schreckenberger. "Major Figures of Contemporary Austrian Literature." *The German Quarterly.* Spring, 1990

14. Wulf Koepke. "After the Fires: Recent Writing in the Germanies, Austria and Switzerland." *German Studies Review.* May, 1988

15. Carl Steiner. "Bitter Healing: German Women Writers from 1700 to 1830." *German Studies Review.* May, 1991

16. Henry J. Schmidt. "Dictionary of Literary Biography. Vol. 56: German Fiction Writers, 1914-1945." *The German Quarterly.* Winter, 1989

17. James Hardin. "Der Weg in die Gegenwart: Geschichte des deutschen Romans." *The German Quarterly.* Mar., 1980

18. Lynn M. Kutch. "The Modern Restoration: Re-thinking German Literary History 1930-1960." *German Studies Review.* Oct., 2006

19. Thomas W. Kniesche. "A Companion to Twentieth-Century German Literature." *German Studies Review.* Oct., 1993

20. Ingeborg M. Goessl. "Austrian Fiction Writers: 1875-1913." *Monatshefte.* Spring, 1991

Starting only with the assumption that there were two latent topics in the corpus and the two assumptions stated in the opening paragraph, this topic model recovers two distinct groups of documents and their characteristic words.

Appendix B

Latent Dirichlet Allocation

Chapter 2 makes use of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). The following describes briefly the model and a common method for fitting the model to a corpus of texts.

"Dirichlet" in the name of LDA refers to the Dirichlet distribution, which figures prominently in the model. The distribution is the multivariate extension of the Beta distribution and describes a distribution over the $K-1$ simplex. The density of a $Dirichlet(\alpha)$ distribution is written

$$p(\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \beta^{\alpha_k - 1} = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \beta^{\alpha_k - 1}$$

If the Dirichlet distribution is parameterized by a single scalar value ($\alpha_1 = \ldots = \alpha_K$) it is referred to as a symmetric Dirichlet distribution.

In many models of text collections the Dirichlet distribution occurs as prior on the parameters of a multinomial distribution. Integrating out the Dirichlet distribution

yields the Dirichlet compound multinomial distribution or Pólya distribution. The combination of the multinomial distribution with a Dirichlet prior on its parameters is given as

$$\theta \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$$
$$x_i \sim Multinomial(1, \theta),\ i \in \{1, \ldots, N\}$$

The marginal probability of $\mathbf{x}$ is found by integrating over θ,

$$\begin{aligned} p(\mathbf{x}|\alpha_1, \ldots, \alpha_K) &= \int Dirichlet(\theta|\alpha_1, \ldots, \alpha_K) \prod_{i=1}^{N} Multinomial(x_i|\theta) d\theta \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \theta^{\alpha_k + n_k - 1} d\theta \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_k + \alpha_k)}{\Gamma(N + \sum_{k=1}^{K} \alpha_k)} \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(N + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \end{aligned}$$

where n_k is the number of times k appears in the values taken by $\{x_i\}_{i=1}^{N}$.

The follow description of LDA makes use of the following notation: D is the number of documents in the corpus, w_{di} is the ith word of document d, and $n_k^{(d)}$ is the number of words in document m associated with topic k. $m_k^{(v)}$ is the number of times words corresponding to the index v are associated with topic k. A dot in the sub- or superscript in the $n_k^{(d)}$ term expresses summation. For example, the number of words in document d is $n_\cdot^{(d)} = \sum_k n_k^{(d)}$.

An advantage of LDA over previous models, such as Probabilistic Latent Semantic Analysis (PLSA), is that it provides a generative description of a corpus, permitting predictions to be made about unseen documents (Hofmann 1999). LDA assumes the following generative model for the corpus:

1. For $k = 1, \ldots, K$
   (a) draw topic distribution over words $\beta_k \sim Dirichlet(\eta)$.
2. For $d = 1, \ldots, D$
   (a) draw document-specific mixture weights $\theta_d \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$
   (b) for $i = 1, \ldots, n_{\cdot}^{(d)}$
      i. draw topic index $z_{di} \sim Discrete(\theta_d)$
      ii. draw word $w_{di} \sim Discrete(\beta_{z_{di}})$

As the notation indicates, the prior distribution for each β_k is a symmetric Dirichlet distribution with scalar parameter η. Using an asymmetric prior distribution for the document-specific topic proportions θ_d has been discussed in Wallach, Mimno, and McCallum (2009).

The joint probability of the LDA model has the following factorization

$$p(\mathbf{w}, \mathbf{z}, \theta_{1:D}, \beta_{1:K} | \boldsymbol{\alpha}, \eta) = \prod_{k=1}^{K} Dir(\beta_k | \eta) \times \prod_{d=1}^{D} Dir(\theta_d | \boldsymbol{\alpha}) \times \prod_{d=1}^{D} \prod_{i=1}^{n_{\cdot}^{(d)}} Discrete(x_{di} | \beta_{z_{di}})$$

Taking advantage of the conjugacy between the Dirichlet prior distribution and the multinomial distribution discussed previously, we may integrate over $\theta_{1:D}$ and $\beta_{1:K}$

$$p(\mathbf{w}, \mathbf{z}|\boldsymbol{\alpha}, \eta) = \prod_{d=1}^{D} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(n_{\cdot}^{(d)} + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k^{(d)} + \alpha_k)}{\Gamma(\alpha_k)} \right) \times$$
$$\prod_{k=1}^{K} \left(\frac{\Gamma(V\eta)}{\Gamma(m_k^{(\cdot)} + V\eta)} \prod_{v=1}^{V} \frac{\Gamma(m_k^{(v)} + \eta)}{\Gamma(\eta)} \right)$$

Inference may be performed via Gibbs Sampling (Griffiths and Steyvers 2004; Casella and George 1992). Conditional on knowing the values for all but one of the topic assignments, $\mathbf{z}_{-i}$, a new topic assignment for z_i may be sampled using the following equation

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}, \boldsymbol{\alpha}, \eta) \propto \frac{m_{-i,k}^{(w_i)} + \eta}{n_{-i,k}^{(\cdot)} + V\eta} \frac{n_{-i,k}^{(d)} + \alpha_k}{n_{-i,\cdot}^{(d)} + \sum \alpha_j}$$

For a detailed derivation of this result, see Carpenter (2010).

Appendix C

Novels Used

The following list of novels maintains the formatting found in Garside and Schöwerling (2000) and supplements (Garside et al. 2006).

1786	gothic	[BECKFORD, William]	[VATHEK]. AN ARABIAN TALE, FROM AN UNPUBLISHED MANUSCRIPT: WITH NOTES CRITICAL AND EXPLANATORY.
1788	gothic	SMITH, Charlotte	EMMELINE, THE ORPHAN OF THE CASTLE. BY CHARLOTTE SMITH. IN FOUR VOLUMES
1790	gothic	[RADCLIFFE, Ann]	A SICILIAN ROMANCE. BY THE AUTHORESS OF THE CASTLES OF ATHLIN AND DUNBAYNE. IN TWO VOLUMES.
1791	gothic	[RADCLIFFE, Ann]	THE ROMANCE OF THE FOREST: INTERSPERSED WITH SOME PIECES OF POETRY. BY THE AUTHORESS OF "A SICILIAN ROMANCE," &C. IN THREE VOLUMES.
1793	gothic	SMITH, Charlotte	THE OLD MANOR HOUSE. A NOVEL, IN FOUR VOLUMES. BY CHARLOTTE SMITH.
1794	gothic	RADCLIFFE, Ann	THE MYSTERIES OF UDOLPHO, A ROMANCE; INTERSPERSED WITH SOME PIECES OF POETRY. BY ANN RADCLIFFE, AUTHOR OF THE ROMANCE OF THE FOREST, ETC. IN FOUR VOLUMES
1796	gothic	LEWIS, M[atthew] G[regory]	THE MONK: A ROMANCE. IN THREE VOLUMES. BY M. G. LEWIS, ESQ. M.P.
1797	gothic	RADCLIFFE, Ann	THE ITALIAN, OR THE CONFESSIONAL OF THE BLACK PENTITENTS. A ROMANCE. BY ANN RADCLIFFE, AUTHOR OF THE MYSTERIES OF UDOLPHO, &C. &C. IN THREE VOLUMES.
1799	gothic	GODWIN, William	ST. LEON: A TALE OF THE SIXTEENTH CENTURY. BY WILLIAM GODWIN. IN FOUR VOLUMES
1800	nationaltale	Maria EDGEWORTH	CASTLE RACKRENT, AN HIBERNIAN TALE. TAKEN FROM FACTS, AND FROM THE MANNERS OF THE IRISH SQUIRES, BEFORE THE YEAR 1782
1801	random	ANON	MYSTERIOUS FRIENDSHIP: A TALE. IN TWO VOLUMES
1801	random	Mary CHARLTON	THE PIRATE OF NAPLES. A NOVEL. IN THREE VOLUMES. BY MARY CHARLTON, AUTHOR OF ROSELLA, ANDRONICA, PHEDORA, &C

1805	random	Isaac D'ISRAELI	FLIM-FLAMS! OR, THE LIFE AND ERRORS OF MY UNCLE, AND THE AMOURS OF MY AUNT! WITH ILLUSTRATIONS AND OBSCURITIES, BY MESSIEURS TAG, RAG, AND BOBTAIL. WITH AN ILLUMINATING INDEX! IN THREE VOLUMES, WITH NINE PLATES
1806	random	ANON	FORRESTI; OR, THE ITALIAN COUSINS. A NOVEL. IN THREE VOLUMES. BY THE AUTHOR OF VALAMBROSA [sic]
1806	gothic	Charlotte DACRE	ZOFLOYA; OR, THE MOOR: A ROMANCE OF THE FIFTEENTH CENTURY. IN THREE VOLUMES. BY CHARLOTTE DACRE, BETTER KNOWN AS ROSA MATILDA, AUTHOR OF THE NUN OF ST. OMERS, HOURS OF SOLITUDE, &C
1806	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE WILD IRISH GIRL; A NATIONAL TALE. BY MISS OWENSON, AUTHOR OF ST. CLAIR, THE NOVICE OF ST. DOMINICK, &C. &C. &C. IN THREE VOLUMES
1807	nationaltale	Anne Louise Germaine de STAËL-HOLSTEIN	CORINNA; OR, ITALY. BY MAD. DE STAËL HOLSTEIN. IN THREE VOLUMES
1808	nationaltale	Charles Robert MATURIN	THE WILD IRISH BOY. IN THREE VOLUMES. BY THE AUTHOR OF MONTORIO
1808	random	Ellen Rebecca WARNER	HERBERT-LODGE; A NEW-FOREST STORY. IN THREE VOLUMES. BY MISS WARNER, OF BATH
1809	nationaltale	Maria EDGEWORTH	TALES OF FASHIONABLE LIFE, BY MISS EDGEWORTH, AUTHOR OF PRACTICAL EDUCATION, BELINDA, CASTLE RACKRENT, ESSAY ON IRISH BULLS, &C. IN THREE VOLUMES
1809	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	WOMAN: OR, IDA OF ATHENS. BY MISS OWENSON, AUTHOR OF THE "WILD IRISH GIRL," THE "NOVICE OF ST. DOMINICK," &C. IN FOUR VOLUMES
1810	gothic	Percy Bysshe SHELLEY	ZASTROZZI, A ROMANCE. BY P. B. S
1811	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE MISSIONARY: AN INDIAN TALE. BY MISS OWENSON. WITH A PORTRAIT OF THE AUTHOR. IN THREE VOLUMES
1815	gothic	Anne Julia Kemble HATTON	SECRET AVENGERS; OR, THE ROCK OF GLOTZDEN. A ROMANCE. IN FOUR VOLUMES. BY ANNE OF SWANSEA, AUTHOR OF CAMBRIAN PICTURES; SICILIAN MYSTERIES; CONVICTION, &C. &C
1815	gothic	ANON	THERESA; OR, THE WIZARD'S FATE. A ROMANCE. IN FOUR VOLUMES. BY A MEMBER OF THE INNER TEMPLE
1815	gothic	ANON	DANGEROUS SECRETS. A NOVEL. IN TWO VOLUMES
1815	gothic	Catherine SMITH	BAROZZI; OR THE VENETIAN SORCERESS. A ROMANCE OF THE SIXTEENTH CENTURY. IN TWO VOLUMES. BY MRS. SMITH, AUTHOR OF THE CALEDONIAN BANDIT, &C. &C
1815	nationaltale	Christian Isobel JOHNSTONE	CLAN-ALBIN: A NATIONAL TALE. IN FOUR VOLUMES
1816	nationaltale	Elizabeth APPLETON	EDGAR: A NATIONAL TALE. BY MISS APPLETON, AUTHOR OF PRIVATE EDUCATION, &C. IN THREE VOLUMES
1816	gothic	Henrietta Rouviere MOSSE	CRAIGH-MELROSE PRIORY; OR, MEMOIRS OF THE MOUNT LINTON FAMILY. A NOVEL. IN FOUR VOLUMES
1816	silverfork	Lady Caroline LAMB	GLENARVON. IN THREE VOLUMES
1816	gothic	Mary Ann SULLIVAN	OWEN CASTLE, OR, WHICH IS THE HEROINE? A NOVEL. IN FOUR VOLUMES. DEDICATED BY PERMISSION TO THE RIGHT HONOURABLE LADY COMBERMERE, BY MARY ANN SULLIVAN, LATE OF THE THEATRES ROYAL, LIVERPOOL, MANCHESTER, NEWCASTLE, BIRMINGHAM, AND NORWICH
1816	gothic	Sophia F. ZIEGENHIRT	THE ORPHAN OF TINTERN ABBEY. A NOVEL. IN THREE VOLUMES. BY SOPHIA F. ZIEGENHIRT, AUTHOR OF SEABROOK VILLAGE, AND SEVERAL HISTORICAL ABRIDGEMENTS
1817	gothic	Anne Julia Kemble HATTON	GONZALO DE BALDIVIA; OR, A WIDOW'S VOW. A ROMANTIC LEGEND. IN FOUR VOLUMES. INSCRIBED, BY PERMISSION, TO WILLIAM WILBERFORCE, ESQ. BY THE AUTHOR OF CAMBRIAN PICTURES, SICILIAN MYSTERIES, CONVICTION, SECRET AVENGERS, CHRONICLES OF AN ILLUSTRIOUS HOUSE, &C. &C

1817	gothic	Anne KER	EDRIC, THE FORESTER: OR, THE MYSTERIES OF THE HAUNTED CHAMBER. AN HISTORICAL ROMANCE, IN THREE VOLUMES. BY MRS. ANNE KER, OF HIS GRACE THE DUKE OF ROXBURGH'S FAMILY, AUTHOR OF THE HEIRESS DI MONTALDE—ADELINE ST. JULIAN—EMMELINE, OR THE HAPPY DISCOVERY—MYSTERIOUS COUNT—AND MODERN FAULTS
1817	gothic	ANON	HOWARD CASTLE; OR A ROMANCE FROM THE MOUNTAINS. IN FIVE VOLUMES. BY A NORTH BRITON
1817	gothic	Edward MOORE	THE MYSTERIES OF HUNGARY. A ROMANTIC HISTORY, OF THE FIFTEENTH CENTURY. IN THREE VOLUMES. BY EDWARD MOORE, ESQ. AUTHOR OF SIR RALPH DE BIGOD, &C. &C
1817	nationaltale	Maria EDGEWORTH	HARRINGTON, A TALE; AND ORMOND, A TALE. IN THREE VOLUMES. BY MARIA EDGEWORTH, AUTHOR OF COMIC DRAMAS, TALES OF FASHIONABLE LIFE, &C. &C
1817	gothic	Nugent BELL	ALEXENA; OR, THE CASTLE OF SANTA MARCO, A ROMANCE, IN THREE VOLUMES. EMBELLISHED WITH ENGRAVINGS
1818	gothic	ANON	THE BANDIT CHIEF; OR, LORDS OF URVINO. A ROMANCE. IN FOUR VOLUMES
1818	nationaltale	Charles Robert MATURIN	WOMEN; OR, POUR ET CONTRE. A TALE. BY THE AUTHOR OF "BERTRAM," &C. IN THREE VOLUMES
1818	gothic	Mary Wollstonecraft SHELLEY	FRANKENSTEIN; OR, THE MODERN PROMETHEUS. IN THREE VOLUMES
1818	nationaltale	Susan Edmonstone FERRIER	MARRIAGE, A NOVEL. IN THREE VOLUMES
1819	random	Adelaide O'KEEFFE	DUDLEY. BY MISS O'KEEFFE, AUTHOR OF PATRIARCHAL TIMES, OR THE LAND OF CANAAN; ZENOBIA, QUEEN OF PALMYRA; &C. IN THREE VOLUMES
1819	gothic	Anne Julia Kemble HATTON	CESARIO ROSALBA; OR, THE OATH OF VENGEANCE. A ROMANCE. IN FIVE VOLUMES. BY ANN OF SWANSEA, AUTHOR OF SICILIAN MYSTERIES, CONVICTION, GONZALO DE BALDIVIA, SECRET AVENGERS, SECRETS IN EVERY MANSION, CAMBRIAN PICTURES, CHRONICLES OF AN ILLUSTRIOUS HOUSE, &C
1819	gothic	ANON	THE CASTLE OF VILLA-FLORA. A PORTUGUESE TALE, FROM A MANUSCRIPT LATELY FOUND BY A BRITISH OFFICER OF RANK IN AN OLD MANSION IN PORTUGAL. IN THREE VOLUMES
1819	random	Elizabeth BENNETT	EMILY, OR, THE WIFE'S FIRST ERROR; AND BEAUTY & UGLINESS, OR, THE FATHER'S PRAYER AND THE MOTHER'S PROPHECY. TWO TALES. IN FOUR VOLUMES. BY ELIZABETH BENNET, AUTHOR OF FAITH AND FICTION, &C. &C
1819	gothic	Zara WENTWORTH	THE RECLUSE OF ALBYN HALL. A NOVEL. IN THREE VOLUMES. BY ZARA WENTWORTH
1820	gothic	Charles Robert MATURIN	MELMOTH THE WANDERER: A TALE. BY THE AUTHOR OF "BERTRAM," &C. IN FOUR VOLUMES
1820	gothic	Francis LATHOM	ITALIAN MYSTERIES; OR, MORE SECRETS THAN ONE. A ROMANCE. IN THREE VOLUMES. BY FRANCIS LATHOM, AUTHOR OF THE MYSTERIOUS FREEBOOTER; LONDON; THE UNKNOWN; MEN AND MANNERS; ROMANCE OF THE HEBRIDES; HUMAN BEINGS; FATAL VOW; MIDNIGHT BELL; IMPENETRABLE SECRET; MYSTERY; &C. &C
1820	gothic	Mrs ISAACS	EARL OSRIC; OR, THE LEGEND OF ROSAMOND. A ROMANCE. BY MRS. ISAACS, AUTHOR OF "TALES OF TO-DAY,"—"WANDERINGS OF FANCY," &C. &C. &C. IN THREE VOLUMES
1820	gothic	Sarah Scudgell WILKINSON	THE SPECTRE OF LANMERE ABBEY, OR THE MYSTERY OF THE BLUE AND SILVER BAG; A ROMANCE. BY SARAH WILKINSON: AUTHORESS OF THE BANDIT OF FLORENCE, FUGITIVE COUNTESS, WHEEL OF FORTUNE, &C. IN TWO VOLUMES
1821	gothic	J. M. H. HALES	DE WILLENBERG; OR, THE TALISMAN. A TALE OF MYSTERY. IN FOUR VOLUMES. BY I. M. H. HALES, ESQ. AUTHOR OF THE ASTROLOGER

1821	gothic	Miss C. D. HAYNES [afterwards GOLLAND, Mrs C. D.]	ELEANOR; OR, THE SPECTRE OF ST. MICHAEL'S. A ROMANTIC TALE. IN FIVE VOLUMES. BY MISS C. D. HAYNES, AUTHOR OF CASTLE LE BLANC; FOUNDLING OF DEVONSHIRE; AUGUSTUS AND ADELINA, &C. &C
1821	gothic	Thomas Henry MARSHAL	THE IRISH NECROMANCER; OR, DEER PARK. A NOVEL. IN THREE VOLUMES. BY THOMAS HENRY MARSHAL
1822	random	Isabel HILL	CONSTANCE, A TALE. BY ISABEL HILL, AUTHOR OF 'THE POET'S CHILD,' A TRAGEDY
1822	random	Jean Charles Léonard SIMONDE DE SISMONDI	JULIA SEVERA; OR THE YEAR FOUR HUNDRED AND NINETY-TWO; TRANSLATED FROM THE FRENCH OF J. C. L. SIMONDE DE SISMONDI, AUTHOR OF NEW PRINCIPLES OF POLITICAL ECONOMY; THE HISTORY OF FRANCE; THE ITALIAN REPUBLICS OF THE MIDDLE AGE; THE LITERATURE OF THE SOUTH OF EUROPE, &C. IN TWO VOLUMES
1823	nationaltale	Alicia LEFANU	TALES OF A TOURIST. CONTAINING THE OUTLAW, AND FASHIONABLE CONNEXIONS. IN FOUR VOLUMES. BY MISS LEFANU, AUTHOR OF STRATHALLAN, LEOLIN ABBEY, HELEN MONTEAGLE, &C
1823	random	George JONES	TEMPTATION. A NOVEL. BY LEIGH CLIFFE, AUTHOR OF "THE KNIGHTS OF RITZBERG,"—"PARGA," "SUPREME BON TON," &C. IN THREE VOLUMES
1823	silverfork	Lady Caroline LAMB	ADA REIS, A TALE. IN THREE VOLUMES
1824	random	Hannah Maria JONES	THE FORGED NOTE: OR, JULIAN AND MARIANNE. A MORAL TALE, FOUNDED ON RECENT FACTS. BY MRS. H. M. JONES, AUTHORESS OF GRETNA GREEN,—WEDDING RING,—BRITISH OFFICER, &C
1824	nationaltale	Susan Edmonstone FERRIER	THE INHERITANCE. BY THE AUTHOR OF MARRIAGE. IN THREE VOLUMES
1825	silverfork	Constantine Henry, Marquis of Normanby PHIPPS	MATILDA; A TALE OF THE DAY
1825	silverfork	Eyre Evans CROWE	THE ENGLISH IN ITALY. IN THREE VOLUMES
1826	silverfork	Benjamin, Earl of Beaconsfield DISRAELI	VIVIAN GREY
1826	random	Sir Walter SCOTT	WOODSTOCK; OR, THE CAVALIER. A TALE OF THE YEAR SIXTEEN HUNDRED AND FIFTY-ONE. BY THE AUTHOR OF "WAVERLEY, TALES OF THE CRUSADERS," &C. IN THREE VOLUMES
1826	silverfork	Thomas Henry LISTER	GRANBY. A NOVEL. IN THREE VOLUMES
1827	random	Sarah Wilmot WELLS	TALES; MOURNFUL, MIRTHFUL, AND MARVELLOUS. BY MRS. WILMOT WELLS, OF MARGATE. IN THREE VOLUMES
1827	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE O'BRIENS AND THE O'FLAHERTYS; A NATIONAL TALE. BY LADY MORGAN. IN FOUR VOLUMES
1828	silverfork	Edward George BULWER LYTTON	PELHAM; OR, THE ADVENTURES OF A GENTLEMAN. IN THREE VOLUMES
1828	nationaltale	John BANIM	THE ANGLO-IRISH OF THE NINETEENTH CENTURY. A NOVEL. IN THREE VOLUMES
1828	silverfork	Lady Caroline Lucy SCOTT	A MARRIAGE IN HIGH LIFE. EDITED BY THE AUTHORESS OF 'FLIRTATION.' IN TWO VOLUMES
1828	silverfork	Thomas Henry LISTER	HERBERT LACY. BY THE AUTHOR OF GRANBY. IN THREE VOLUMES
1831	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	THE YOUNG DUKE. BY THE AUTHOR OF "VIVIAN GREY." IN THREE VOLUMES.
1831	nationaltale	Ferrier	DESTINY; OR, THE CHIEF'S DAUGHTER. BY THE AUTHOR OF "MARRIAGE," AND "THE INHERITANCE."
1831	random	REYNOLDS, Frederick	A PLAYWRIGHT'S ADVENTURES
1832	silverfork	[GORE, Catharine Grace Frances]	THE OPERA: A NOVEL. BY THE AUTHOR OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1833	silverfork	[BULWER LYTTON, Edward George]	GODOLPHIN. A NOVEL. IN THREE VOLUMES.
1833	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	THE WONDROUS TALE OF ALROY. THE RISE OF ISKANDER. BY THE AUTHOR OF "VIVIAN GREY," "CONTARINI FLEMING," &C.; IN THREE VOLUMES.
1833	silverfork	[GARDINER, Marguerite], Countess of Blessington	THE REPEALERS. A NOVEL. BY THE COUNTESS OF BLESSINGTON. IN THREE VOLUMES

1833	silverfork	[GORE, Catherine Grace Frances]	THE SKETCH BOOK OF FASHION. BY THE AUTHOR OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1833	silverfork	[HOOK, Theodore Edward]	LOVE AND PRIDE. BY THE AUTHOR OF "SAYINGS AND DOINGS," ETC. IN THREE VOLUMES.
1833	random	[TONNA], Charlotte Elizabeth	DERRY, A TALE OF THE REVOLUTION. BY CHARLOTTE ELIZABETH, AUTHORESS OF OSRIC, THE ROCKITE, THE SYSTEM, &C.; &C.;
1835	random	CAUNTER, J[ohn] Hobart	POSTHUMOUS RECORDS OF A LONDON CLERGYMAN. EDITED BY THE REV. HOBART CAUNTER, B.D., AUTHOR OF THE ORIENTAL ANNUAL.
1835	random	[DEACON, William Frederick]	THE EXILE OF ERIN; OR, THE SORROWS OF A BASHFUL IRISHMAN. IN TWO VOLUMES.
1835	nationaltale	MORGAN, Lady [Sydney] [née OWENSON, Sydney]	THE PRINCESS; OR, THE BEGUINE. BY LADY MORGAN, AUTHOR OF "O'DONNEL," &C.; IN THREE VOLUMES.
1835	random	[SULLIVAN, Arabella Jane]; DACRE, Lady [Barbarina] (editor)	TALES OF THE PEERAGE AND PEASANTRY. EDITED BY LADY DACRE. IN THREE VOLUMES.
1836	silverfork	[GARDINER, Marguerite], Countess of Blessington	THE CONFESSIONS OF AN ELDERLY GENTLEMAN. ILLUSTRATED BY SIX FEMALE PORTRAITS, FROM HIGHLY FINISHED DRAWINGS BY E. T. PARRIS. BY THE COUNTESS OF BLESSINGTON.
1836	silverfork	[GORE, Catherine Grace Frances]	MRS. ARMYTAGE; OR, FEMALE DOMINATION. BY THE AUTHORESS OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1836	silverfork	[HOOK, Theodore Edward]	GILBERT GURNEY. BY THE AUTHOR OF ""SAYINGS AND DOINGS,"" ""LOVE AND PRIDE,"" ETC. IN THREE VOLUMES.
1837	silverfork	[BURY, Lady Charlotte Susan Maria]	THE DIVORCED. BY LADY CHARLOTTE BURY, AUTHORESS OF FLIRTATION, &c. &c IN TWO VOLUMES.
1837	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	VENETIA. BY THE AUTHOR OF "VIVIAN GREY" AND "HENRIETTA TEMPLE." IN THREE VOLUMES.
1847	silverfork	[GORE, Catherine Grace Frances]	CASTLES IN THE AIR. A NOVEL. BY MRS. GORE. IN THREE VOLUMES.

Appendix D

Inferring Relations with a Vector Autoregressive Model

D.1 Introduction and Notation

An observation in a VAR(p) model is an m-dimensional column vector that typically has the form

$$\mathbf{y}'_t = \mathbf{c} + \mathbf{y}'_{t-1}\,\mathbf{\Phi}_1 + \mathbf{y}'_{t-2}\,\mathbf{\Phi}_2 + \cdots + \mathbf{y}'_{t-p}\,\mathbf{\Phi}_p + \boldsymbol{\epsilon}_t \tag{D.1}$$

where $\mathbf{c}$ is a $1 \times m$ unknown vector, each $\mathbf{\Phi}_i$ is an $m \times m$ matrix, and $\boldsymbol{\epsilon}_t$ is an m-dimensional column vector of evolution errors. In the case of normally-distributed errors, $\boldsymbol{\epsilon}_t \sim N_m(0, \Sigma)$. $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_j$ are independent for any $i \neq j$.

If we assume that there are a total of T observations that the first p are observed without error, we may write the model more conveniently by defining $\mathbf{x}'_t = (1, \mathbf{y}'_{t-1}, \cdots, \mathbf{y}'_{t-p})$ and writing

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}'_{p+1} \\ \vdots \\ \mathbf{y}'_T \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_T \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{c} \\ \boldsymbol{\Phi_{p+1}} \\ \vdots \\ \boldsymbol{\Phi_p} \end{pmatrix}, \mathbf{E} = \begin{pmatrix} \boldsymbol{\epsilon}'_{p+1} \\ \vdots \\ \boldsymbol{\epsilon}'_{p+1} \end{pmatrix}$$

yielding the concise expression and an entrée for multivariate linear regression

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{D.2}$$

where $\mathbf{Y}$ and $\mathbf{E}$ are $(T-p) \times m$ matrices, $\mathbf{X}$ is a $(T-p) \times (1+pm)$, and $\mathbf{B}$ is a $(1+pm) \times m$ matrix.

D.2 Conditional Likelihood

If we condition on the first p values of $\mathbf{y}_{1:T}$ and let $k = 1 + pm$, the likelihood may be written as

$$\begin{aligned} f(\mathbf{y}_{(p+1):T} \,|\, \mathbf{y}_{1:p}) &\propto |\Sigma|^{-(T-p)/2} \exp\{-\frac{1}{2} \sum_{t=p+1}^{T} (y_t - \mathbf{B}'\, x_t)' \Sigma^{-1} (y_t - \mathbf{B}'\, x_t)\} \\ &= |\Sigma|^{-(T-p)/2} \operatorname{etr}\{-\frac{1}{2} (\mathbf{Y} - \mathbf{X}\,\mathbf{B})' (\mathbf{Y} - \mathbf{X}\,\mathbf{B}) \Sigma^{-1}\} \\ &= |\Sigma|^{-(T-p-k)/2} \operatorname{etr}\{-\frac{1}{2} \Sigma^{-1} S\} \times |\Sigma|^{-k/2} \operatorname{etr}\{-\frac{1}{2} (\mathbf{B} - \widehat{\mathbf{B}})' X' X (\mathbf{B} - \widehat{\mathbf{B}}) \Sigma^{-1}\} \end{aligned}$$

where $\operatorname{etr}(\cdot)$ is the exponential of the trace, $S = (Y - X\,\widehat{\mathbf{B}})'(Y - X\,\widehat{\mathbf{B}})$, and $\hat{B} = (X'X)^{-1}X'Y$, the maximum likelihood estimate of B. In the final line, we recognize that the conditional likelihood is proportional to the product of an inverse Wishart distribution, $\mathrm{IW}(\Sigma | T-p-k-m-1, S^{-1})$, and a matrix normal distribution, $N(B|\hat{B}, (X'X)^{-1}, \Sigma)$.

Recall that an $m \times m$ positive definite, symmetric matrix Σ has an inverse Wishart distribution, written $IW(d, A)$, when its probability distribution function (pdf) is

$$p(\Sigma) = c|\Sigma|^{-(d+m+1)/2} \operatorname{etr}(-\Sigma^{-1}A^{-1}/2)$$

with normalizing constant

$$c^{-1} = |A|^{d/2}2^{dm/2}\pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma((d+1-i)/2).$$

A $k \times m$ random vector B has a matrix normal distribution, written $B \sim N(M, U, V)$, when its pdf is written

$$p(B) = (2\pi)^{-km/2}|U|^{-m/2}|V|^{-k/2} \times \operatorname{etr}\{-(B-M)'U^{-1}(B-M)V^{-1}/2\}.$$

D.3 Prior

The matrix normal, inverse Wishart distribution is a conjugate prior (Rossi, Allenby, and McCulloch 2005). It has the form

$$B|\Sigma \sim N(B_0, V_0, \Sigma)$$
$$\Sigma \sim IW(d_0, S_0^{-1})$$

D.4 Posterior

The prior and conditional likelihood above lead to the posterior distribution for B given y,

$$B|\Sigma, y \sim N(\tilde{B}, \tilde{V}, \Sigma)$$
$$\Sigma|y \sim IW(\tilde{d}, \tilde{S}^{-1})$$

where

$$\tilde{V} = [V_0^{-1} + X'X]^{-1}$$

$$\tilde{B} = \tilde{V}\left[V_0^{-1}B_0 + X'X\hat{B}\right]$$

$$\tilde{d} = d_0 + T - p$$

$$\tilde{S} = S + S_0 + \hat{B}'X'X\hat{B} + B_0'V_0^{-1}B_0 - \tilde{A}'(V_0^{-1} + X'X)\tilde{A}$$

D.5 Ancestral VAR

The "ancestral VAR" described in 4 is just another multivariate linear regression problem.

The ancestral graph pictured as figure 4.3 has an adjacency matrix:

$$\Lambda = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0
\end{pmatrix}$$

The graph shown as figure 4.3 conditions on the first 5 observations. If all the observations are stacked into the rows of a matrix Z then our model,

$$y_t = \Phi\left(\frac{\sum_{i \in pa(t)} y_i}{|pa(t)|}\right) + \epsilon_t,$$

for $t \in 6, \ldots, 14$ may be written as

$$Y = \begin{pmatrix} \mathbf{1}_n & \hat{\Lambda}^T Z \end{pmatrix} B + E \tag{D.3}$$

where $\hat{\Lambda}$ is simply Λ where the columns have divided by their sums to capture the $\frac{1}{|pa(t)|}$ term. Y, B, and E are the familiar counterparts from equation D.2.

For the pedagogical example described in chapter 4, the following rudimentary prior parameters were used, $B_0 = \mathbf{0}_{k \times q}, V_0 = 10 \cdot \mathbf{I}_k, d_0 = q + 1, S_0 = \mathbf{I}_q$.

Bibliography

Abrams, M. H., and Stephen J. Greenblatt, eds. 2000. *The Norton Anthology of English Literature.* 7th ed. Vol. 2. New York: Norton.

Adamic, Lada, and Natalie Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election." In *Proceedings of the 3rd International Workshop on Link Discovery,* 36–43. LinkKDD 2005. New York: ACM.

Adburgham, Alison. 1983. *Silver Fork Society.* London: Constable.

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment.* Pamphlet 1. Stanford Literary Lab: Stanford University.

Barbrook, Adrian C., Christopher J. Howe, Norman Blake, and Peter Robinson. 1998. "The Phylogeny of The Canterbury Tales." *Nature* 394 (6696): 839.

Barthes, Roland. 1967. "The Death of the Author." *Aspen* (5-6).

Belgum, Kirsten. 1998. *Popularizing the Nation: Audience, Representation, and the Production of Identity in Die Gartenlaube, 1853-1900.* Lincoln, NE: University of Nebraska Press.

Bennett, Tony. 2009. "Counting and Seeing the Social Action of Literary Form: Franco Moretti and the Sociology of Literature." *Cultural Sociology* 3 (2): 277–297.

Bérubé, Michael. 2011. "The Science Wars Redux." *Democracy* (19).

Blei, David. 2012. "Introduction to Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Blei, David M., and John D. Lafferty. 2006. "Dynamic Topic Models." In *Proceedings of the 23rd International Conference on Machine Learning,* 113–120. Pittsburgh, PA: ACM.

———. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1): 17–35.

Blei, David M., and John D. Lafferty. 2009. "Topic Models." In *Text Mining: Classification, Clustering, and Applications,* edited by Ashok Srivastava and Mehran Sahami, 71–89. Boca Raton, FL: CRC Press.

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

Block, Andrew. 1961. *The English Novel, 1740-1850.* 2nd ed. London: Dawsons.

Block, Sharon, and David Newman. 2011. "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts." *Journal of Women's History* 23 (1): 81–109.

Bourdieu, Pierre. 1988. "Flaubert's Point of View." Translated by Priscilla Parkhurst Ferguson. *Critical Inquiry* 14 (3): 539–562.

———. 1996. *The Rules of Art: Genesis and Structure of the Literary Field.* Stanford: Stanford University Press.

Boyle, James. 2008. *The Public Domain: Enclosing the Commons of the Mind.* New Haven: Yale University Press.

Brin, Sergey, and Lawrence Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30 (1-7): 107–117.

Burrows, J. F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burt, Ronald S. 2004. "Structural Holes and Good Ideas." *American Journal of Sociology* 110 (2): 349–399.

Busa, Roberto. 1974. *Index Thomisticus.* 56 vols. Stuttgart: Frommann-Holzboog.

Carpenter, Bob. 2010. "Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling." Accessed March 12, 2012. `http://lingpipe.files.wordpress.com/2010/07/lda3.pdf`.

Casanova, Pascale. 1999. *Le republique mondiale des lettres.* Paris: Editions du Seuil.

Casanova, Pascale. 2004. *The World Republic of Letters.* Translated by M. B. DeBevoise. Cambridge, MA: Harvard University Press.

Casella, George, and Edward I. George. 1992. "Explaining the Gibbs Sampler." *The American Statistician* 46 (3): 167–174.

Casey, Ellen Miller. 1996. "Edging Women out?: Reviews of Women Novelists in the "Athenaeum," 1860-1900." *Victorian Studies* 39 (2): 151–171.

Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, July 5.

Cazamian, Louis François. 1973. *The Social Novel in England, 1830-1850: Dickens, Disraeli, Mrs. Gaskell, Kingsley.* London: Routledge / Kegan Paul.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems 22,* edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 288–296.

Cisne, John L. 2005. "How Science Survived: Medieval Manuscripts' "Demography" and Classic Texts' Extinction." *Science* 307 (5713): 1305–1307.

Cohen, Margaret. 2002. *The Sentimental Education of the Novel.* Princeton: Princeton University Press.

Colletti, Lucio. 1974. "A Political and Philosophical Interview." *New Left Review* I (86).

Collini, Stephan. 1993. "Introduction." In *The Two Cultures,* vii–lxxi. London: Cambridge University Press.

"Corvey Introduction." Sheffield Hallam Corvey Project. 2013. Accessed January 23. `http://extra.shu.ac.uk/corvey/intro/index.html`.

Crane, Gregory. 2006. "What Do You Do with a Million Books?" *D-Lib Magazine* 12 (3).

Crowe, Michael J. 1967. *A History of Vector Analysis: The Evolution of the Idea of a Vectorial System.* Notre Dame, IN: University of Notre Dame Press.

Dick, Philip K. 1968. *Do Androids Dream of Electric Sheep?* New York: New American Library.

Easley, David, and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* New York: Cambridge University Press.

Eastwood, Jonathan. 2007. "Bourdieu, Flaubert, and the Sociology of Literature." *Sociological Theory* 25 (2): 149–169.

Eley, Geoff. 2005. *A Crooked Line: From Cultural History to the History of Society.* University of Michigan Press, October 27.

Elson, David, Nicholas Dames, and Kathleen McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting*

of the Association for Computational Linguistics, 138–147. Uppsala, Sweden: Association for Computational Linguistics.

English, James F. 2008. "Literary Studies." In *The SAGE Handbook of Cultural Analysis,* edited by Tony Bennett and John Frow. London: SAGE.

———. 2010. "Everywhere and Nowhere: The Sociology of Literature After "the Sociology of Literature"." *New Literary History* 41 (2): v–xxiii.

Feinerer, Ingo, Kurt Hornik, and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5): 1–54.

Fouetillou, Guilhem. 2006. "Ecologie de la Blogopole." Observatoire Présidentielle 2007. Accessed January 2, 2013. `http://www.observatoire-presidentielle.fr/?pageid=31`.

Frow, John. 2006. *Genre.* London: Routledge.

———. 2008. "Thinking the Novel." *New Left Review* II (49): 137–145.

Gallagher, Catherine. 1985. *The Industrial Reformation of English Fiction: Social Discourse and Narrative Form, 1832-1867.* Chicago: University of Chicago Press.

Gallop, Jane. 2007. "The Historicization of Literary Studies and the Fate of Close Reading." *Profession* 2007 (1): 181–186.

Garside, Peter. 1991. "Popular Fiction and National Tale: Hidden Origins of Scott's Waverley." *Nineteenth-Century Literature* 46 (1): 30–53.

———. 2000. "The English Novel in the Romantic Era." In *The English novel, 1770-1829,* edited by Peter Garside, James Raven, and Rainer Schöwerling. Vol. 2. 2 vols. Oxford: Oxford University Press.

Garside, Peter, Jacqueline Belanger, and Anthony Mandal. 2001. *'The English Novel, 1800-1829': Update 1 (Apr 2000-May 2001).* 6. Centre for Editorial and Intertextual Research.

Garside, Peter, Anthony Mandal, Verena Ebbes, Angela Koch, and Rainer Schöwerling. 2006. "The English Novel, 1830-36: A Bibliographic Survey of Fiction Published in the British Isles." The English Novel, 1830–36. January 26. Accessed August 12, 2011. `http://www.cardiff.ac.uk/encap/journals/corvey/1830s/index.html`.

Garside, Peter, and Rainer Schöwerling. 2000. *The English Novel, 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles.* Edited by Peter Garside, James Raven, and Rainer Schöwerling. Vol. 2. 2 vols. Oxford: Oxford University Press.

Gerrish, Sean M, and David M. Blei. 2010. "A Language-based Approach to Measuring Scholarly Impact." In *Proceedings of the 27th International Conference on Machine Learning,* 375–382. Omnipress.

Godfrey-Smith, Peter. 2003. *Theory and Reality: An Introduction to the Philosophy of Science.* Chicago: University of Chicago Press.

———. 2007. *Popper's Philosophy of Science: Looking Ahead.* Draft, forthcoming in The Cambridge Companion to Popper, Shearmur and G. Stokes (eds.) Accessed March 8, 2013.

———. 2009. *Darwinian Populations and Natural Selection.* Oxford: Oxford University Press.

Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101:5228–5235.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35.

Griswold, Wendy. 2008. *Regionalism and the Reading Class.* University Of Chicago Press.

Grün, Bettina, and Kurt Hornik. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13): 1–30.

Gulli, Antonio, and A. Signorini. 2005. "The Indexable Web is More Than 11.5 Billion Pages." In *Proceedings of the 14th International Conference on World Wide Web–Special Interest Tracks and Posters,* 902–903. ACM.

Hall, David. 2008. "Tracking the Evolution of Science." Bachelor's thesis.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. "Studying the History of Ideas Using Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 363–371. Association for Computational Linguistics.

Haraway, Donna. 1985. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." *Socialist Review* 15 (2).

Harvey, Paul H., and Mark D. Pagel. 1991. *The Comparative Method in Evolutionary Biology.* Oxford: Oxford University Press.

Haskell, Thomas L. 1975. "The True and Tragic History of 'Time on the Cross'." *New York Review of Books* 22 (15).

Hayles, N. Katherine. 2012. *How We Think: Digital Media and Contemporary Technogenesis.* University of Chicago Press.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods.* New York: Springer.

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 50–57. New York: ACM.

Hollingsworth, Keith. 1963. *The Newgate Novel, 1830-1847; Bulwer, Ainsworth, Dickens & Thackeray.* Detroit: Wayne State University Press.

Hope, Jonathan, and Michael Witmore. 2004. "The Very Large Textual Object: A Prosthetic Reading of Shakespeare." *Early Modern Literary Studies* 9 (3). Accessed September 20, 2011.

Howe, Christopher J., Adrian C. Barbrook, Matthew Spencer, Peter Robinson, Barbara Bordalejo, and Linne R. Mooney. 2001. "Manuscript Evolution." *Endeavour* 25 (3): 121–126.

Isaac, Larry. 2009. "Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel." *American Sociological Review* 74 (6): 938–965.

James, I. M. 2002. *Remarkable Mathematicians: From Euler to von Neumann.* Washington, DC: Mathematical Association of America.

Jameson, F. 1981. *The Political Unconscious: Narrative as a Socially Symbolic Act.* Ithaca: Cornell University Press, February 28.

Jauss, Hans Robert. 1989. "Literary History as a Challenge to Literary Theory." In *The Critical Tradition: Classic Texts and Contemporary Trends,* 1197–1218. New York: St. Martin's Press.

Jones, Beatrix, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. 2005. "Experiments in Stochastic Computation for High-Dimensional Graphical Models." *Statistical Science* 20 (4): 388–400.

Kadane, Joseph B. 2011. *Principles of Uncertainty.* Chapman & Hall/CRC.

Kelly, Gary. 1976. *The English Jacobin Novel 1780-1805.* Oxford: Clarendon Press.

Kerman, Judith B. 1997. *Retrofitting Blade Runner: Issues in Ridley Scott's Blade Runner and Phillip K. Dick's Do Androids Dream of Electric Sheep?* Bowling Green: Bowling Green State Univ. Popular Press.

Koch, Angela. 2002. "Gothic Bluebooks in the Princely Library of Corvey and Beyond." *Cardiff Corvey: Reading the Romantic Text* (9).

Kolaczyk, Eric D. 2009. *Statistical Analysis of Network Data: Methods and Models.* New York: Springer.

Kruschke, John K. 2010. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS.* Burlington, MA: Academic Press.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions.* University of Chicago Press.

Latour, Bruno. 2004. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30 (2): 225–248.

———. 2010. "Tarde's Idea of Quantification." In *The Social After Gabriel Tarde: Debates and Assessments,* edited by Matei Candea. London: Routledge.

Latour, Bruno, Pablo Jensen, Tommaso Venturini, Sébastian Grauwin, and Dominique Boullier. 2012. "'The Whole is Always Smaller Than its Parts' – a Digital Test of Gabriel Tardes' Monads." *The British Journal of Sociology* 63 (4): 590–615.

Lee, Michael, Brandon Pincombe, and Matthew Welsh. 2005. "An Empirical Evaluation of Models of Text Document Similarity." In *Proceedings of the 27th Annual Conference of the Cognitive Science Society,* edited by B. G. Bara, L. W. Barsalou, and M. Bucciarelli, 1254–1259. Mahwah, NJ: Erlbaum.

Lee, Peter M. 2004. *Bayesian Statistics: An Introduction.* 3rd ed. London: Wiley.

Lessig, Lawrence. 2005. *Free Culture: The Nature and Future of Creativity.* New York: Penguin Press.

Lévy, Maurice. 1968. *Le Roman gothique anglais, 1764-1824.* Toulouse: Association des publications de la Faculté des lettres et sciences humaines.

Lieberson, Stanley. 2000. *A Matter of Taste: How Names, Fashions, and Culture Change.* New Haven: Yale University Press.

MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms.* Cambridge: Cambridge University Press.

Madnani, N., J. Tetreault, and M. Chodorow. 2012. "Re-examining Machine Translation Metrics for Paraphrase Identification." In *Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2012),* 182–190. ACL.

Mann, Geoff. 2009. "Colletti on the Credit Crunch." *New Left Review* II (56): 119–127.

Manning, Christopher D., and Hinrich Schüzte. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Mayr, Ernst. 1976. "Typological versus Population Thinking." In *Evolution and the Diversity of Life,* 26–9. Cambridge, MA: Harvard University Press.

———. 2001. *What Evolution Is.* New York: Basic Books.

McGray, Douglas. 2002. "Japan's Gross National Cool." *Foreign Policy* 130 (May/June): 44–54.

Meilă, Marina. 2002. *Comparing Clusterings.* UW Statistics Technical Report 418. University of Washington.

———. 2007. "Comparing Clusterings—an Information Based Distance." *Journal of Multivariate Analysis* 98 (5): 873–895.

Menand, Louis. 2010. *The Marketplace of Ideas: Reform and Resistance in the American University.* New York: W. W. Norton.

Mimno, David. 2011. "Reconstructing Pompeian Households." In *Uncertainty in Artificial Intelligence.* Barcelona, Spain.

———. 2012a. "Computational Historiography: Data Mining in a Century of Classics Journals." *ACM Journal of Computing in Cultural Heritage* 5 (1): 3:1–3:19.

———. 2012b. "Topic Regression." Ph.D. thesis.

Mimno, David, and David Blei. 2011. "Bayesian Checking for Topic Models." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* 227–237. ACL.

Moretti, Franco. 1982. "L'Anima e l'arpia." *Quaderni Piacentini* (5): 43–83.

———. 1988. *Signs Taken for Wonders.* London: Verso.

———. 2000a. "Conjectures on World Literature." *New Left Review* (1).

———. 2000b. "The Slaughterhouse of Literature." Volume 61, Number 1, March 2000, *MLQ: Modern Language Quarterly* 61 (1): 207–227.

———. 2000c. *The Way of the World: The Bildungsroman in European Culture.* 2nd ed. London: Verso.

Moretti, Franco. 2003a. “Graphs, Maps, Trees: Abstract Models for Literary History-1.” *New Left Review* (24): 67–93.

———. 2003b. “More Conjectures.” *New Left Review* (20).

———. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History.* London: Verso.

———. 2006. “The End of the Beginning.” *New Left Review* (41): 71–87.

———. 2008. “The Novel: History and Theory.” *New Left Review* (52).

———. 2009. “Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850).” *Critical Inquiry* 36 (1): 134–158.

Mosteller, Frederick, and David L Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Reading, MA: Addison-Wesley.

Nelson, Robert K. 2011. “Mining the Dispatch.” Mining the Dispatch. Accessed March 28, 2012. `http://dsl.richmond.edu/dispatch/`.

Novembre, John, and Matthew Stephens. 2008. “Interpreting Principal Component Analyses of Spatial Population Genetic Variation.” *Nature Genetics* 40 (5): 646–649.

Nye, Joseph S. 1990. *Bound to Lead: The Changing Nature of American Power.* Basic Books.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics* 155 (2): 945–959.

Propp, Vladimir. 1968. *Morphology of the Folktale.* 2nd ed. Edited by Louis A. Wagner. Translated by Laurence Scott. Austin: University of Texas Press.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rabkin, Eric S. 2004. “Science Fiction and the Future of Criticism.” *PMLA* 119 (3): 457–473.

Radway, Janice A. 1991. *Reading the Romance: Women, Patriarchy, and Popular Literature.* 2nd ed. Chapel Hill: University of North Carolina Press.

———. 1999. *A Feeling for Books: The Book-of-the-Month Club, Literary Taste, and Middle-Class Desire.* Chapel Hill: University of North Carolina Press.

Redhead, Steve. 2010. "From Marx to Berlusconi: Lucio Colletti and the Struggle for Scientific Marxism." *Rethinking Marxism* 22 (1): 148–156.

Resnik, Philip, and Eric Hardisty. 2010. *Gibbs Sampling for the Uninitiated.* CS-TR-4956. University of Maryland.

Ringer, Fritz K. 1969. *The Decline of the German Mandarins: The German Academic Community, 1890-1933.* Cambridge, MA: Harvard University Press.

Rogers, Deborah S., Marcus W. Feldman, and Paul R. Ehrlich. 2009. "Inferring Population Histories Using Cultural Data." *Proceedings of the Royal Society B: Biological Sciences* 276 (1674): 3835–3843.

Rollin, Charles. 1804. *The Ancient History of the Egyptians, Carthaginians, Assyrians, Babylonians, Medes & Persians, Macedonians, and Grecians.* Vol. 6. 8 vols. London: Printed for W. J. / J. Richardson et al.

Rossi, Peter E., Greg M. Allenby, and Rob McCulloch. 2005. *Bayesian Statistics and Marketing.* 1st ed. Wiley, December 9.

Schuh, Randall T., and Andrew V. Z. Brower. 2009. *Biological Systematics: Principles and Applications.* 2nd ed. Ithaca: Cornell University Press.

Sewell Jr., William H. 2005. "The Political Unconscious of Social and Cultural History, or, Confessions of a Former Quantitative Historian." In *The Politics of Method in the Human Sciences: Positivism and Its Epistemological Others.* Duke University Press Books.

Shalizi, Cosma. 2006. "Graphs, Trees, Materialism, Fishing." The Valve - A Literary Organ. January 24. Accessed May 3, 2008. `http://www.thevalve.org/go/valve/article/graphs_trees_materialism_fishing/`.

———. 2011. "Graphs, Trees, Materialism, Fishing." In *Reading Graphs, Maps, and Trees: Responses to Franco Moretti,* edited by John Holbo and Jonathan Goodwin. Parlor Press.

Sim, Yanchuan, Noah A. Smith, and David A. Smith. 2012. "Discovering Factions in the Computational Linguistics Community." In *Proceedings of the ACL Workshop on Rediscovering Fifty Years of Discoveries,* 22–23.

Simon, Carl P., and Eric S. Rabkin. 2008. "Culture, Science Fiction, and Complex Adaptive Systems: The Work of the Genre Evolution Project." In *Biocomplexity at the Cutting Edge of Physics, Systems Biology and Humanities,* edited by

Castone Castellani, Elena Lamberti, Vita Fortunati, and Claudio Franceschi, 279–294. Bologna: Bononia University Press.

Smith, Barbara. 2006. *Scandalous Knowledge: Science, Truth and the Human.* Durham, NC: Duke University Press.

Snow, C. P. 1993. *The Two Cultures.* London: Cambridge University Press.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach.* Oxford: Blackwell.

Spivak, Gayatri Chakravorty, and Cathy Caruth. 2010. "Interview with Gayatri Chakravorty Spivak." *PMLA* 125 (4): 1020–1025.

Sutherland, John. 1988. "Publishing History: A Hole at the Centre of Literary Sociology." *Critical Inquiry* 14 (3): 574–589.

———. 1989a. "Review: Edging Women Out." *The American Journal of Sociology* 95 (3): 814–816.

———. 1989b. *The Stanford Companion to Victorian Fiction.* Stanford, CA: Stanford University Press.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–1581.

Tëmkin, I., and N. Eldredge. 2007. "Phylogenetics and Material Cultural Evolution." *Current Anthropology* 48 (1): 146.

Trumpener, Katie. 1997. *Bardic Nationalism: The Romantic Novel and the British Empire.* Princeton, NJ: Princeton University Press.

———. 1998. "National Tale." In *Encyclopedia of the Novel,* edited by Paul E. Schellinger, 910–11. Chicago: Fitzroy Dearborn.

Tuchman, Gaye. 1989. *Edging Women Out: Victorian Novelists, Publishers, and Social Change.* In collaboration with Nina E. Fortin. New Haven: Yale University Press.

Tuchman, Gaye, and Nina Fortin. 1980. "Edging Women Out: Some Suggestions about the Structure of Opportunities and the Victorian Novel." *Signs* 6 (2): 308–325.

Tuchman, Gaye, and Nina E. Fortin. 1984. "Fame and Misfortune: Edging Women Out of the Great Literary Tradition." *The American Journal of Sociology* 90 (1): 72–96.

Tynyanov, Yuri. 1927. "On Literary Evolution." In *The Critical Tradition,* edited by David H. Richter, translated by C. A. Luplow. New York: St. Martin's Press, 1989.

Ulrich, Laurel. 1990. *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary, 1785-1812.* New York, NY: Knopf.

Unsworth, John. 2006. "20th-Century American Bestsellers." Accessed April 2, 2012. `http://people.lis.illinois.edu/~unsworth/courses/bestsellers/`.

Vinh, N.X., J. Epps, and J. Bailey. 2010. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance." *Journal of Machine Learning Research* 11:2837–2854.

Wallach, Hanna, David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." In *Advances in Neural Information Processing Systems 22,* edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 1973–1981.

West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models.* 2nd. New York: Springer.

White, Hayden. 1999. *Figural Realism: Studies in the Mimesis Effect.* Baltimore: The Johns Hopkins University Press.

———. 2003a. "Anomalies of Genre: The Utility of Theory and History for the Study of Literary Genres." *New Literary History* 34 (3): 597–615.

———. 2003b. "Commentary: Good of Their Kind." *New Literary History* 34 (2): 367–376.

Wilhite, Allen W., and Eric A. Fong. 2012. "Coercive Citation in Academic Publishing." *Science* 335 (6068): 542–543.

Williamson, S., C. Wang, K. Heller, and D. Blei. 2010. "The IBP Compound Dirichlet process and its Application to Focused Topic Modeling." In *Proceedings of the 27th International Conference on Machine Learning,* edited by Thorsten Joachims and Johannes Fürnkranz, 1151–1158. Haifa, Israel: Omnipress, June.

Wimsatt, W. K., and Beardsley C. Monroe. 1954. "The Intentional Fallacy." In *The Verbal Icon: Studies in the Meaning of Poetry.* Lexington: University of Kentucky Press.

Winter, Sidney G. 1987. "Natural Selection and Evolution." In *The New Palgrave: a Dictionary of Economics,* edited by John Eatwell, Murray Milgate, and Peter Newman. New York: Stockton Press.

Winthrop-Young, Geoffrey. 1999. "How the Mule Got Its Tale: Moretti's Darwinian Bricolage." Volume 29, Number 2, Summer 1999, *Diacritics* 29 (2): 18–40. Accessed May 2, 2008.

Ferrier, Susan Edmonstone (1782–1854). 2004. In *Oxford Dictionary of National Biography,* online edn, Oct 2006. Oxford: Oxford University Press, by Elspeth Yeo. Accessed January 29, 2013.

Biography

Allen Beye Riddell was born in San Diego, California in 1980. He received his B.A. in Comparative Literature from Stanford University in 2004, his M.S. in Statistics from Duke University in 2013, and his Ph.D. from the Graduate Program in Literature at Duke University in 2013.