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

Allen Beye Riddell

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 introduces 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.[1] 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.

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 do you do with a million books?"[2] 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.[3] 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 rereading the nearly 10,000 diary entries of Martha Ballard, a midwife in Maine around 1800.[4] 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 ca. 2,500 issues of the weekly magazine *Die Gartenlaube* (The Garden Bower) printed between 1853 and 1900.[5] 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.[6] 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.[7] 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 interrater reliability.

Another example of collaborative reading is Larry Isaac's study of the "labor problem novel" in nineteenth- and early twentieth-century American fiction.[8] 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, from 1870 to 1918. Since thousands of novels were published in the United States during this period, reading through all 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 five hundred 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, then 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 as not only 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.

Machine Reading: Latent Dirichlet Allocation and Topic Models

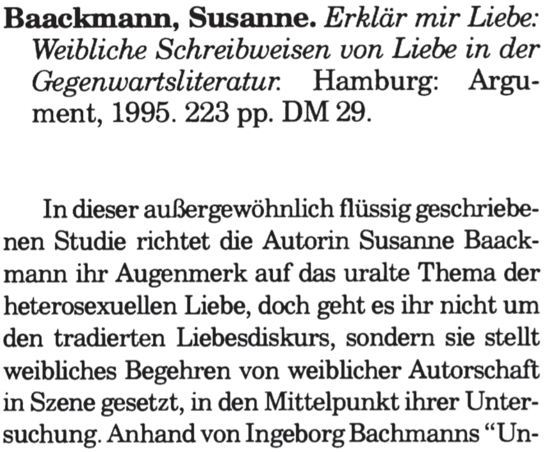
Other ways of reading a very large collection of texts exist. A range of alternative approaches might be labeled, following N. K. Hayles, "machine reading."[9] In this section, I will introduce one of these alternatives, known informally as a *topic model.*

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. *German Studies Review*, which first appeared in 1978[10]

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 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 3.1 shows a page scan of a



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 3.1. 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)

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 the image in figure 3.1 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 word 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 all numbers (fig. 3.2).[11] 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

```
<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 3.2. JSTOR XML for Karin U. Herrmann's review *Erklär mir Liebe*

Lines have been reordered to enable comparison with figure 3.1.

idiosyncrasies like the mangling of umlauts. That this is not possible—that US and international law blocks the noncommercial 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.[12]

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.

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.[13] 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 3.1 (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. 3.2).

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 Theodor Fontane's novel *Effi Briest* (1894). Each chapter is considered as a separate text document. If our vocabulary were limited to two solitary words: "Effi" and "Innstetten"—the names of the two main characters—the resulting table of word counts would be Table 3.2. This table provides a compact, if impoverished, representation of each chapter. Each row of counts (each chapter) may also be considered alone as a pair of numbers—for example, (21, 7). These pairs may be interpreted as *vectors*—specifically, vectors in two-dimensional space (fig. 3.3). This is where the name *vector*

	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	···
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 3.1. Word frequencies for book reviews in *The German Quarterly* (Summer 1997)

	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

Table 3.2. Word frequencies for selected chapters of *Effi Briest*

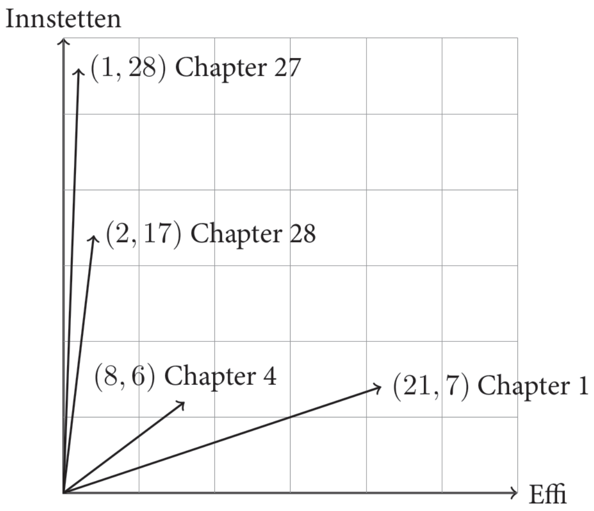


Figure 3.3. Chapters of *Effi Briest* represented as vectors in a two-dimensional plane

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.

The advantages of using the vector space model are best understood in the following context: mathematicians have spent nearly two hundred years developing machinery for manipulating, comparing, and creating vectors.[14] 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. 3.4). This notion of "pointing" in the same direction can be made precise by referring to the angle between vectors. This angle is easy to calculate when it is used to compare two vectors: it goes by the name *cosine distance*.[15]

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 3.3 lists these articles.

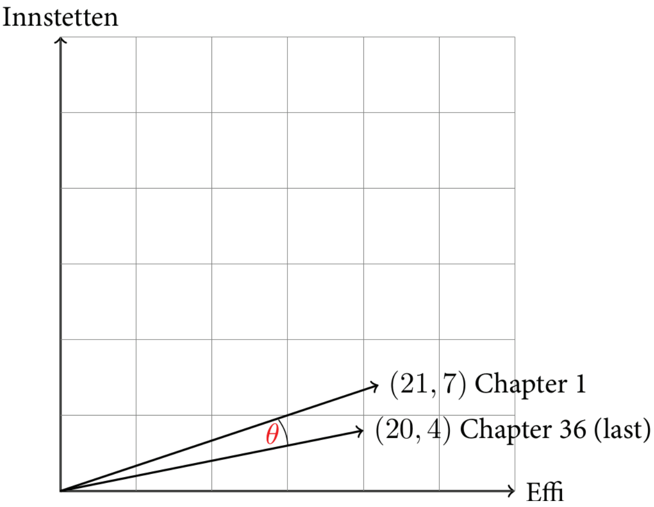


Figure 3.4. Cosine distance between chapters 1 and 36

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.

Table 3.3. Articles similar and dissimilar to Karin U. Herrmann's review of *Erklär mir Liebe*

Like any abstraction, the vector space model obscures important aspects of texts, word order chief among them—for example, "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. And as a practical matter, in cases where one is dealing with roughly comparable texts, experiments have shown that cosine distance and related measures are only loosely correlated with human judgments of similarity.[16]

Another objection to 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 Innstetten, but she is called "gnädige Frau" by others. We would also be interested in the possessive form "Effis" along 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 in 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 which 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.

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.[17] 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 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).[18] 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.*[19] 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 journal article corpus.[20]

To understand how LDA works it is easiest to start with the end result.[21] 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 percent topic 25, 17 percent topic 19, and 9 percent topic 20 (with 27 percent distributed with smaller shares over the remaining 27 topics; fig. 3.5). The plurality of the words is 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.”

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 in an online appendix to this chapter.[22] 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. The addition of probability to the model permits the association of the word “weimar” with two different topics.

Should we trust that the description of documents in terms of topics corresponds at all with what our judgments would have been, had we read the 22,198 articles? The titles of journal articles provide a 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. Verifying that what the topic

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

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.

	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

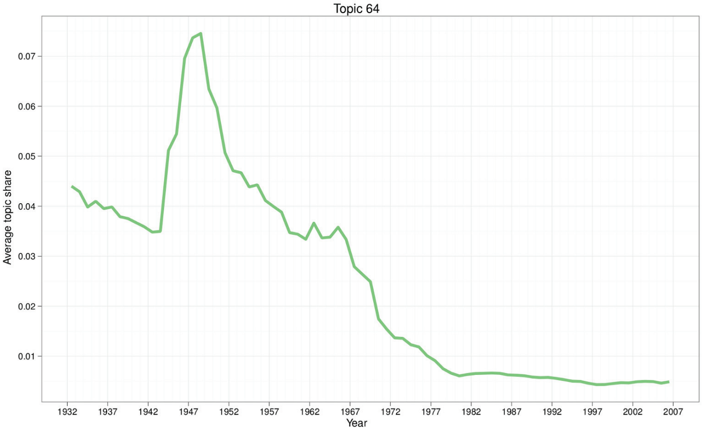
Figure 3.5. Catherine Dollard’s *German Studies Review* article viewed in terms of prominent topics. Shares and words are based on a topic model (LDA) with thirty topics. Considered separately, each of the remaining topics contributes less than 0.05.

shares imply is also what the article title implies is a convenient way to check that a topic model has succeeded in capturing important themes in a collection of texts.[23]

Four German Studies Journals (1928–2006)

To explore the corpus of journal articles using LDA, I fixed the number of topics at a hundred.[24] As described previously, LDA infers the distribution of the hundred topics across all the 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. 3.6). While some of the decline in articles on language instruction is surely an artifact of the corpus (in 1968 *The German Quarterly* split off

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.

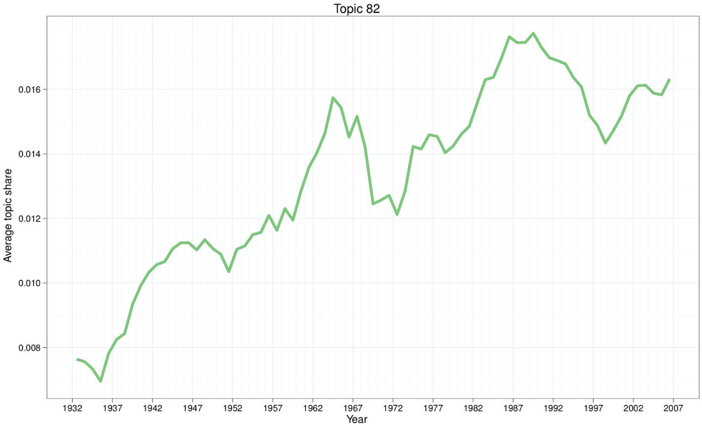
Figure 3.6. Topic 64: Characteristic words, five-year moving average, and representative articles

a separate journal for language instruction, *Die Unterrichspraxis*, which is not included in the corpus), the decline in the share of these articles is visible well before 1968.

The second trend is the gradual rise in articles concerned with literature and literary criticism (fig. 3.7). This trend is connected with a topic characterized by words such as “literature,” “literary,” “writers,” and “authors.”

The recent history of US universities offers a 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.[25]

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

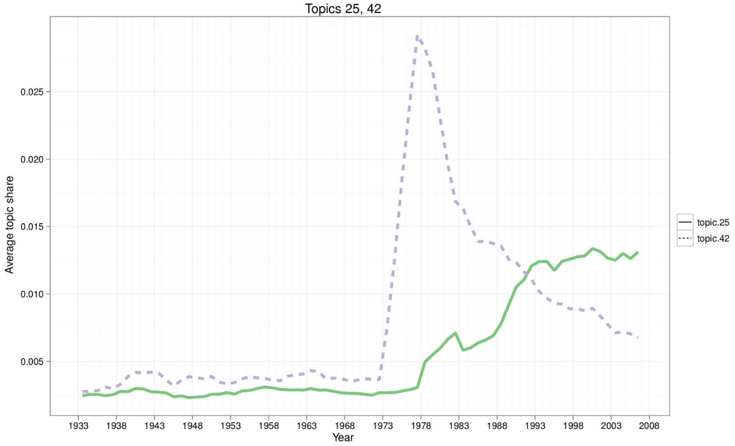


- Leland R. Phelps, review of *The Emergence of German as a Literary Language* by Eric A. Blackall, *Monatshefte* 52 (April-May 1960): 213-14.
- Andreas Kiryakakis, review of *Dictionary of Literary Biography: Volume 66: German Fiction Writers, 1885-1913 Part I: A-L* by James Hardin, *German Studies 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 3.7. Topic 82: Characteristic words, five-year moving average, and representative articles

In addition to the decline of articles on teaching and rise of articles on research, two other topics exhibit distinctive trends (fig. 3.8). 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," "political," "class,"

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 3.8. Topics 25 and 42: Characteristic words, five-year moving averages, and representative articles

and "society." Herbert Marcuse's "The Failure of the New Left" numbers among the articles most strongly associated with this 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."[26]

All the topics mentioned so far appear in different proportions in the corpus. Figure 3.9 shows the frequency of several topics over time on the same scale. Recall 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 emphasizing that the LDA model makes use of relative rather than absolute word frequencies. That is, a 500-word review that is 20 percent topic 64 is treated the same, in certain important respects, as a 9,000-word article that is 20 percent 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. With topics associated with only a few articles a year, such as the "folktales" topic discussed later, selection bias becomes a concern. It is possible that some trends are not

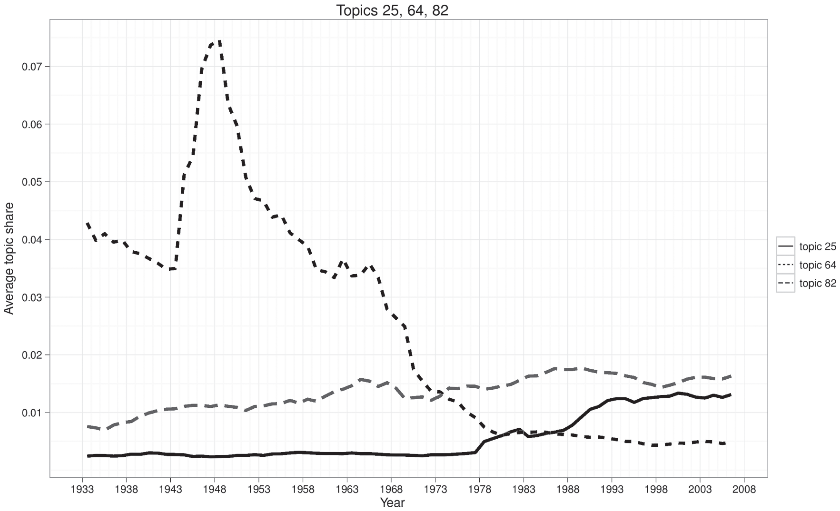


Figure 3.9. Comparison of topics 25 ("women . . ."), 64 ("students . . ."), and 82 ("literature . . .")

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.

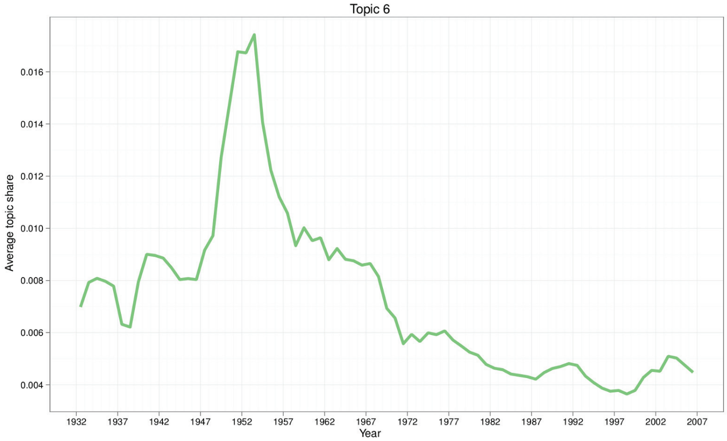
Long Nineteenth-Century Topics

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. 3.10). 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. 3.11). With peaks around 1955 and 1990, there is a temptation to think that

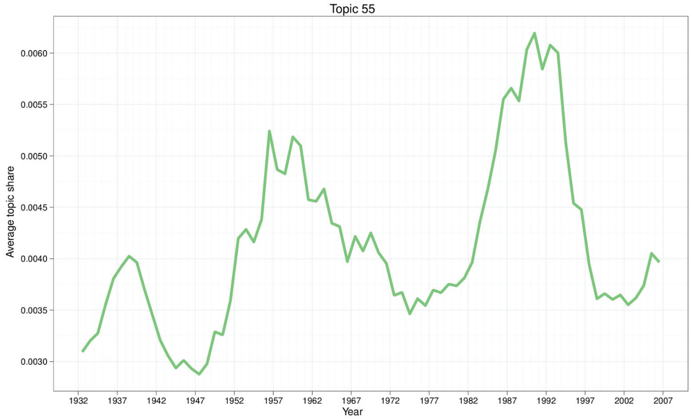
goethe faust goethes wilhelm werther weimar iphigenie ottilie gretchen charlotte meisters mephisto meister dichtung wahlverwandtschaften



- 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 3.10. Topic 6: Characteristic words, five-year moving average, and representative articles

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 3.11. Topic 55: Characteristic words, five-year moving average, and representative articles

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.

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., held-out likelihood) do not always align with human readers' assessments of the coherence of inferred topics, suggesting a mismatch at some level between topic models and topics familiar to human readers.[27] 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.

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. 3.5) and a hundred-topic fit to characterize the same corpus of journal articles. While many of the topics of the thirty-topic fit resemble those of the hundred-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 the particular model in use. Blei and his coauthors are admirably clear on this point.[28]

LDA and other topic models also make assumptions known to be incorrect.[29] For example, LDA assumes that the 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.

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 1924.

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.[30] For example, David Hall takes up a hypothesis inspired by Thomas Kuhn's account of the historical trajectory of science as one punctuated by periodic "revolutions" in dominant methods.[31] Hall observes that there have been widely acknowledged shifts in the prominence of certain methods within computational linguistics over the past twenty years. If these methodological shifts represented a revolutionary change of "paradigm" in Kuhn's sense, then Hall anticipated that the researchers associated with "insurgent" methods would not be participants in a field—that is, authors of articles—with long standing. In other words, these researchers would be new arrivals, not established scholars abandoning existing methodologies in favor of new ones. 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 a 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 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.

Notes

[1] Sharon Block and David Newman, "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts," *Journal of Women's History* 23, no. 1 (2011): 81–109; Justin Grimmer, "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis* 18, no. 1 (2010): 1–35; David Hall, "Tracking the

Evolution of Science" (bachelor's thesis, Stanford University, 2008); David Hall, Daniel Jurafsky, and Christopher D. Manning, "Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI: Association for Computational Linguistics, 2008), 363–71; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *ACM Journal of Computing in Cultural Heritage* 5, no. 1 (2012), doi:10.1145/2160165.2160168.

[2] Gregory Crane, "What Do You Do with a Million Books?" *D-Lib Magazine* 12, no. 3 (March 2006), doi:10.1045/march2006-crane.

[3] David Mimno and David Blei, "Bayesian Checking for Topic Models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Somerset, NJ: Association for Computational Linguistics, 2011), 227–37; Robert K. Nelson, "Mining the Dispatch," Digital Scholarship Lab, University of Richmond, accessed March 28, 2012, http://dsl.richmond.edu/dispatch.

[4] Laurel Ulrich, *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary, 1785–1812* (New York, NY: Knopf, 1990).

[5] Kirsten Belgum, *Popularizing the Nation: Audience, Representation, and the Production of Identity in Die Gartenlaube, 1853–1900* (Lincoln: University of Nebraska Press, 1998); Fritz K. Ringer, *The Decline of the German Mandarins: The German Academic Community, 1890–1933* (Cambridge, MA: Harvard University Press, 1969).

[6] Larry Isaac, "Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel," *American Sociological Review* 74, no. 6 (2009): 938–65, doi:10.1177/000312240907400605; Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005); Carl P. Simon and Eric S. Rabkin, "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*, ed. Gastone Castellani et al. (Bologna: Bononia University Press, 2008), 279–94; John Unsworth, "20th-Century American Bestsellers," accessed October 29, 2013, http://people.lis.illinois.edu/~unsworth/courses/bestsellers.

[7] Eric S. Rabkin, "Science Fiction and the Future of Criticism," *PMLA* 119, no. 3 (2004): 457–73; Simon and Rabkin, "Culture, Science Fiction, and Complex Adaptive Systems."

[8] Isaac, "Movements, Aesthetics, and Markets in Literary Change."

[9] N. Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis* (Chicago: University of Chicago Press, 2012), 55–80.

[10] *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*. 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 two hundred 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 size of the remaining lexicon was 74,158 unique terms. The total number of words in all articles was 15,680,621.

[11] This final step—removing all numbers—creates a special problem with this corpus. Since the Eszett (ß) is mangled by JSTOR OCR into "13," all words containing ß are removed as they contain a numeric character ("3"). 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.

[12] James Boyle, *The Public Domain: Enclosing the Commons of the Mind* (New Haven, CT: Yale University Press, 2008); Lawrence Lessig, *Free Culture: The Nature and Future of Creativity* (New York: Penguin Press, 2005).

[13] 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} ≠ {5,4,4}.

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

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

[16] Michael Lee, Brandon Pincombe, and Matthew Welsh, "An Empirical Evaluation of Models of Text Document Similarity," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Erlbaum, 2005), 1254–59.

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

[18] 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. See I. M. James, *Remarkable Mathematicians: From Euler to von Neumann* (Washington, DC: Mathematical Association of America, 2002).

[19] David Blei, "Introduction to Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (2012): 77–84, doi:10.1145/2133806.2133826. 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" (79).

[20] For subsequent developments, see David M. Blei and John D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, ed. William Cohen and Andrew Moore (Pittsburgh, PA:

Association for Computing Machinery, 2006), 113–20; Yee Whye Teh et al., "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association* 101, no. 476 (2006): 1566–81; Hannah Wallach, David Mimno, and Andrew McCallum, "Rethinking LDA: Why Priors Matter," in *Advances in Neural Information Processing Systems 22*, ed. Y. Bengio et al. (La Jolla, CA: Neural Information Processing Systems: 2009), 1973–81; Sinead Williamson, Chong Wang, Katherine A. Heller, and David M. Blei. "The IBP Compound Dirichlet Process and Its Application to Focused Topic Modeling," in *Proceedings of the 27th International Conference on Machine Learning*, ed. Thorsten Joachims and Johannes Fürnkranz (Madison, WI: International Machine Learning Society, 2010), 1151–58.

[21] Other introductions to LDA include Blei, "Introduction to Probabilistic Topic Models," and David M. Blei and John D. Lafferty, "Topic Models," in *Text Mining: Classification, Clustering, and Applications*, ed. Ashok Srivastava and Mehran Sahami (Boca Raton, FL: CRC Press, 2009), 71–94.

[22] Allen B. Riddell, "A Simple Topic Model," accessed October 29, 2013, http://purl.org/NET/how-to-read-n-articles-appendix.

[23] The validation of topic models is an area of research in its own right. For a discussion of the issue, see Jonathan Chang et al., "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems 22*, ed. Y. Bengio (La Jolla, CA: Neural Information Processing Systems, 2009), 288–96.

[24] 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, "Rethinking LDA,1973–81." The R software environment was used to model the data in conjunction with the tm and topicmodels packages; visualizations were made using ggplot2. See R Development Core Team, *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing, 2011); Ingo Feinerer, Kurt Hornik, and David Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software* 25, no. 5 (March 2008): 1–54; Bettina Grün and Kurt Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software* 40, no. 13 (2011): 1–30.

[25] Louis Menand, *The Marketplace of Ideas: Reform and Resistance in the American University* (New York: W. W. Norton, 2010), 64–66, 74–77.

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

[27] Chang et al., "Reading Tea Leaves," 288–96.

[28] Blei, Ng, and Jordan, "Latent Dirichlet Allocation," 996n1.

[29] Wallach, Mimno, and McCallum, "Rethinking LDA"; Williamson et al., "The IBP Compound Dirichlet Process and Its Application to Focused Topic Modeling"; David M. Blei and John D. Lafferty, "A Correlated Topic Model of Science," *The Annals of Applied Statistics* 1, no. 1 (2007): 17–35, doi:10.1214/07-AOAS114; Blei and Lafferty, "Dynamic Topic Models."

[30] Hall, "Tracking the Evolution of Science"; Hall, Jurafsky, and Manning, "Studying the History of Ideas Using Topic Models"; Yanchuan Sim, Noah Smith, and David Smith, "Discovering Factions in the Computational Linguistics Community," in *Proceedings of the ACL Workshop on Rediscovering Fifty Years of Discoveries* (Jeju, Korea: Association for Computational Linguistics, 2012): 22–23.

[31] Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).