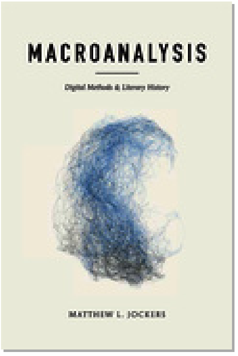
Macroanalysis

Matthew L. Jockers



Published by University of Illinois Press

Matthew L. Jockers.
Macroanalysis: Digital Methods and Literary History.
Champaign: University of Illinois Press, 2013.
Project MUSE. Web. 7 Jul. 2015. http://muse.jhu.edu/.

8 THEME

All ideas are second hand, consciously and unconsciously drawn from a million outside sources, and daily used by the garnerer with a pride and satisfaction born of the superstition that he originated them.

—Mark Twain, letter to Anne Macy (1903)

A typical complaint about computational stylistics is that such studies fail to investigate the aspects of writing that readers care most deeply about, namely, plot, character, and theme.* In the previous chapter, we saw how stylistic information can be usefully extracted from texts in a corpus and how the derivative data can be used to chart linguistic macro patterns and macro trends present in a century's worth of novels. I also began to address the trickier business of theme through a discussion of a particularly "British" word cluster, a cluster that I suggested as a possible surrogate for an expression or thread of "confidence" that runs through much British prose and much less through Irish prose. My analysis of this word cluster, selected from among other frequently occurring word tokens, represented a small and imperfect step in the direction of thematic discovery. If we are to capture the great richness of thematic diversity in the corpus, however, then from the small step, a giant leap is now required.

Summarizing the arguments of the Russian "preformalist" Alexander Veselovsky, Victor Erlich writes that "the main concern of the literary historian is not with assessing the unique contributions of individual writers, but with spotting the migratory poetic formulae; accounting for their appearance in various ethnic milieus . . . and tracing them through all vicissitudes back to the starting point" (1980, 29). Of general types and themes in literature, Alexander Veselovsky explains that a "gifted poet . . . may by chance hit upon this or that motif, produce

* See, for example, Withshire's response (1988) to Burrow's study of Jane Austen, which was noted in chapter 4.

imitators, create a school of writers . . . [but] these minor details . . . are hardly discernible in the broad alternation of the socio-poetic demand and supply" (as cited in ibid., 29). Veselovsky sought to define a science of literary poetics that would allow him to argue that literature evolves partially—or even completely—independent of individual creativity.* Literary history in Veselovsky's conception should be viewed as a series of recurring narrative plots, motifs, and devices that overshadow and dwarf the minor contributions of individual authors. These recurring elements exist in a larger literary system that is external to, or at least "outside" of, the immediate consciousness of the authors.

In the 1890s, Veselovsky and his brother Aleksey theorized along these lines about the origins of poetry. They attempted to trace the genesis of current poetic themes by adopting the methods of mythographers and linguists who conceived of recurrent themes as the products of external influence (see Polonsky 1998, esp. 16–17). Fascinating work to be sure, but, more generally, the Veselovsky brothers were interested in comparative literature and specifically in understanding and defining the influence of Western culture on Russian literature. Aleksey Veselovsky's study on the subject, *The Western Influence in New Russian Literature,* has not been translated into English, but Rachael Polonsky's rendering of the opening paragraph provides a usable jumping-off point for understanding just how ambitious the goal was:

> "The exchange of ideas, images, fables, artistic forms between tribes and peoples of the civilized world is one of the most important things studied by the still-young science of literary history." This process of exchange is "one of the laws of development of artistic creativity." On its way through history, a people assimilates the tales and myths, ideas and dreams, fables and folk motifs of others; "all this merges [*slivaet*] with its own birthright." "Borrowing [*zaimstvovanie*] can go from people to people . . . moving through time and space so that it becomes indirect . . . peoples can be influenced by peoples they have never touched." The exchange of ideas is an "eternal principle" that will be encountered whether a scholar studies literature by genre or by school. (1998, 18–19; brackets in the original)

These conclusions—hypotheses, really—set forth in the opening paragraph are ambitious, and, regrettably, Veselovsky never manages to take them beyond the anecdotal type of analysis to which we are still accustomed, which is to say a close reading. Hoping to show the broad interinfluences of literature, Aleksey

* The brothers Veselovsky, Alexander and Aleksey, may be seen as precursors, or "forerunners," as Erlich writes, of the main formalist movement of the 1920s and 1930s. Although their "preoccupation with genealogy . . . was largely abandoned by the Formalist theoreticians," their conceptions of plot and theme are echoed in works by Shklovsky, Propp, and others. See Erlich 1980, chap. 13.

Veselovsky wrestles with Pushkin and Tolstoy in the context of Rabelais, Goethe, and Cervantes, but in the end, *The Western Influence in New Russian Literature* fails to get beyond the limits of human-scale synthesis. Instead of generating big theories, the work becomes sidetracked by a more political battle that has Veselovsky wrestling his conservative colleagues who wished to reject the entire premise: Russian literature was most certainly too "original" and "pure" to be open to external linguistic and thematic pollution! However unrealized their objectives may have been, the Veselovskys provide nothing short of a call to arms for the modern, digitally equipped, scholar. With big data and computation, we possess the ability to identify and track the "migratory formulae" of literary history that the Veselovskys imagined.*

In terms of giving scholars a means of tracking thematic trends over time, the release of the Google Ngram Viewer in 2011 lowered the bar considerably.† The Ngram Viewer offers users the opportunity to track an n-gram's relative frequency behavior over the course of time and in a corpus of several million books. The Ngram Viewer, however, is not without problems, and users must exercise caution in terms of what can be said about theme based on the relative frequency of individual words. Among the more obvious problems is the issue of metadata, or more precisely the lack of metadata available to users of the Ngram tool. The only metadata provided are publication dates, and even these are frequently incorrect. Different printings, different editions, and the unaccounted-for presence of duplicate works in the corpus complicate matters even further. Even if these issues are resolved, something that the authors of the tool and the accompanying paper in *Science* (Michel et al. 2011) promise to do, other problems remain for scholars wishing to make the interpretive move from word to theme.‡ When we examine a word, or an n-gram, out of the context in which it appears, we inevitably lose information about how that word is being employed. If we are merely interested in stylistic habits, such out-of-context counting can be usefully employed, but when it comes to drawing semantic meaning from a word, we require more than a count of that word's occurrence in the corpus. A word's meaning is derived through context; as the English linguist John Rupert Firth has famously noted, "You shall know a word by the company it keeps" (1957, 11).

* I am grateful to my colleague Glen Worthey, who read and translated for me relevant sections of Veselovsky's *Western Influence in New Russian Literature.*

† http://books.google.com/ngrams.

‡ My blog post on December 12, 2010, offers a more specific breakdown of the most obvious problems. See http://www.matthewjockers.net/2010/12/22/unigrams-and-bigrams-and-trigrams-oh-my/.

Douglas Biber's *Corpus Linguistics: Investigating Language Structure and Use* is an excellent primer on many factors complicating word-focused text analysis and the subsequent conclusions one might draw regarding word meanings. As Biber points out, using a concordance program that produces keyword-in-context lists is a good start because such lists offer context, but these lists are hardly practical when it comes to an analysis of anything beyond even a few thousand occurrences of a key word in a given corpus (Biber et al. 2006). For investigations of this scale, the next option is to examine keywords and their collocates: the words that tend to co-occur with a keyword being investigated. Biber notes the "strong tendency for each collocate of a word to be associated with a single sense or meaning" of the word. He adds that "identifying the most common collocates of a word provides an efficient and effective means to begin analyzing senses" (ibid., 36). Biber also describes the problems associated with words that have varying parts of speech and thus "mean" differently depending on how they are being used. The Google Ngram Viewer provides no way of distinguishing between the word *deal* when used as a noun and *deal* when used as a verb (to use Biber's example). Given that the Google Ngram Viewer does allow for bigram, trigram, and so on searching, we are able to take some rough stabs at separating the two meanings by including context words. We may, for example, search for *deal the,* as in "deal the cards," versus *deal of,* as in "a great deal of monkey business." Here again, however, the quality of our results is limited by our ability to anticipate, in advance, the number and variety of potential collocates: for example, *deal to, deal in, deal with,* and so forth. Complicating matters still further, the researcher exploring the Google corpus must also consider the lemmas of the words under investigation. What, for example, of *dealt* and *dealing*?*

This is not to say that the Google Ngram Viewer and its corpus is of no, or even of little, value. On the contrary, there is much to be learned from the way that words "dance through history together."† Nevertheless, the researcher and layman alike must be cognizant of these limitations, especially when making the leap from word frequency to word meaning, or from word frequency to theme. The latter is precisely what many in the blogosphere seemed to be doing in the days following the release of the Ngram tool. I wager, in fact, that to date the vast majority of words entered into the Ngram Viewer have been nouns.

* I must acknowledge that the Ngram Viewer was built, at least in part, out of a desire to study these kinds of variants. I do not wish to disparage the larger research project and would point out that the Ngram tool is only the most public manifestation of the research.

† This lovely phrase is from Ryan Heuser, a former student and current coordinator of the Stanford Literary Lab.

Moreover, I suspect that the great majority of users entering these nouns have entered them with the assumption that they have some inherent or tangible meaning in connection with "culture," as if to say that the usage of a word is evidence of its purchase and relevance to the community represented in the corpus, as if a word could fully stand in for a particular concept, a theme. In fact, the Ngram Viewer offers little in terms of interpretive power: it cannot tell us *why* a particular word was popular or not; it cannot address the historical *meaning* of the word at the time it was used (something at which the *Oxford English Dictionary* is particularly good), and it cannot offer very much at all in terms of *how* readers might have perceived the use of the word. Forgetting even these issues, we cannot ignore the life that words have outside of written discourse. When we talk about the Ngram Viewer as a window into culture, or "culturomics," we speak only of written culture; even less so, we speak of written culture as it is curated by librarians at major research universities who have partnered with Google in scanning the world's books. If we believe the Ngram data, usage of the word *cool* peaked in 1940s and then began a precipitous drop that continued at least until the mid-1980s. Forgetting entirely the problem of disambiguating the meanings of *cool* (that is, "cold" versus "interesting"), surely the word's actual frequency in "culture" is not reflected in the Ngram data. *Cool,* as in "That's a cool car, man," remains ubiquitous in spoken dialogue, if not in published work.

The meanings of words are found in their contexts, and the Ngram Viewer provides only a small peephole into context. The size of the digital library, and even of the much smaller Stanford Literary Lab corpus of some thirty-five hundred nineteenth-century novels, makes n-gram viewing, KWIC lists, and collocate studies untenable for all but the most infrequent of words. If our goal is to understand the narrative subjects and the recurrent themes and motifs that operate in the literary ecosystem, then we must go beyond the study of individual n-grams, beyond the words, beyond the KWIC lists, and beyond even the collocates in order to capture what is at once more general and also more specific. Cultural memes and literary themes are not expressed in single words or even in single bigrams or trigrams. Themes are formed of bigger units and operate on a higher plane. If we may know the sense of a word by the company of words that surround it in a sentence, we may know a theme by the sentences, paragraphs, chapters, and even full books that express it. In short, simple word-to-word collocations and KWIC lists do not provide enough information to rise to the level of theme. What is needed in order to capture theme are collocations of collocations on a much larger scale.

Probabilistic latent semantic indexing (Hofmann 1999, 2001) and, more specifically, probabilistic topic modeling employing latent Dirichlet allocation (LDA) take us a very long way toward fulfilling that need (see Blei, Ng, and

Jordon 2003; Blei et al. 2004; Griffiths and Steyvers 2002, 2003, 2004; Steyvers and Griffiths 2007). Topic models are, to use a familiar idiom, the mother of all collocation tools. This algorithm, LDA, derives word clusters using a generative statistical process that begins by assuming that each document in a collection of documents is constructed from a mix of some set of possible topics. The model then assigns high probabilities to words and sets of words that tend to co-occur in multiple contexts across the corpus.*

Aside from the researcher's somewhat arbitrary setting of the number of topics to be "discovered," the entire process is done in an unsupervised fashion. "Unsupervised" means that the machine does not know in advance what themes to look for. With no human input into what constitutes a theme, a motif, a topic, the model collects distributions of co-occurring words and then returns them in a manner that allows us to examine, assess, interpret, and intuit what they all have in common, that is, their shared "theme."† The reality, of course, is that the process is far more complicated, and readers with a high tolerance for equations and plate notation will find satisfying reading in Blei, in Steyvers, and in Newman. Those less concerned with the statistics and more interested in humanistic knowledge, I direct to historian Sharon Block, whose work topic modeling an eighteenth-century newspaper is, to my knowledge, the earliest example of topic modeling in the humanities. Working with David Newman, Block performed an analysis of articles taken from the *Pennsylvania Gazette* in the 1700s. In the paper, Block describes topic modeling as follows:

* A layman's explanation of the LDA process can be found online at http://www.matthewjockers.net/macroanalysisbook/lda/.

† In this chapter, I use the terms *theme, topic,* and *motif* as synonyms for the same general concept: namely, a type of literary material, that is, "subject matter," that recurs with some degree of frequency throughout and across a corpus. This material functions as a central and unifying unit of a text or texts. Despite a long history of studying theme and motif in literature and even more extensively in folklore, these terms do remain ambiguous, terms of convenience. I believe that the word clusters discussed here are self-evidently thematic and that even while the matter of what constitutes a theme or motif is a broad area in which some things are black, some white, and some gray, most readers will recognize in the word distributions the larger thematic category to which these words belong. Although handbooks such as *Themes and Motifs in Western Literature* (Daemmrich and Daemmrich 1987) may help us to understand theme, even these scholarly compilations are open to the charge of being arbitrary—they define by example, not by concise definition. Daemmrich offers the theme of "Death," for example, and I am comfortable accepting "Death" as a theme. But Daemmrich also records a theme called "Eye." To my mind, "Eye" would be more appropriately chronicled in a dictionary of symbols than in a handbook of themes.

> Topic modeling is based on the idea that individual documents are made up of one or more topics. It uses emerging technologies in computer science to automatically cluster topically similar documents by determining the groups of words that tend to co-occur in them. Most importantly, topic modeling creates topical categories without a priori subject definitions. This may be the hardest concept to understand about topic modeling: unlike traditional classification systems where texts are fit into preexisting schema (such as Library of Congress subject headings), topic modeling determines the comprehensive list of subjects through its analysis of the word occurrences throughout a corpus of texts. The content of the documents—not a human indexer—determines the topics collectively found in those documents. (2006, n.p.)

Readers who find Block's work on a corpus of newspaper articles useful may find Cameron Blevin's blog post (2010) about topic modeling Martha Ballard's diary equally appealing.* For our purposes here, suffice to say that the model identifies words that tend to co-occur together in multiple places in multiple documents. If the statistics are rather too complex to summarize here, I think it is fair to skip the mathematics and focus on the end results. We needn't know how long and hard Joyce sweated over *Ulysses* to appreciate his genius, and a clear understanding of the LDA machine is not required in order to see the beauty of the result.

The results that I shall describe and explore here were derived from David Mimno's implementation of LDA in the open-source MAchine Learning for LanguagE Toolkit, or "MALLET," software packaged developed by Andrew McCallum (2002) and other contributors at the University of Massachusetts–Amherst. In my experience, the MALLET implementation is the most robust and best-tested topic-modeling software package. The documentation accompanying Mimno's implementation of LDA describes topic modeling as follows: "Topic models provide a simple way to analyze large volumes of unlabeled text. A 'topic' consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings" (McCallum 2009, n.p.). In short, topic modeling provides a scalable method for word-sense disambiguation that specifically addresses the limitations of traditional collocation and KWIC lists that were discussed above. Unlike KWIC and collocate lists, which require careful human interrogation in order to parse out one word sense from another, topic modeling works in an unsupervised way, inferring information about individual word senses based on their repeated appearance in similar contextual situations. The resulting "topics" in a given collection of texts are not provided to the re-

* See also Robert K. Nelson's online commentary "Of Monsters, Men—and Topic Modeling" that appeared in the *New York Times* online opinion pages, May 29, 2011, http://opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/.

searcher in the form of labeled themes (for example, the theme of "Seafaring"), but rather as sets of words, which are "ranked," or "weighted," according to their probabilities of appearing together in a given topic. The ease with which these resulting word clusters can be readily identified as topics or themes is referred to as topic "coherence" or topic "interpretability." Figures 8.1 and 8.2 use word clouds to visualize two topics that were harvested from the Stanford Literary Lab's collection of 3,346 books. The individual words are weighted, and the weights represent the probability of the word in the given topic.*

It is not difficult for a human interpreter to make sense of the word clouds and to assign labels to the themes represented by these algorithmically derived word lists. Without getting too specific, the first may be labeled "Native Americans" and the second "Ireland."† Within each topic are words that, when taken in isolation, would provide very little in terms of thematic information. Consider, for example, the word *stream* in topic 271, the "Native Americans" topic.

* That is, *Indians* is the word most likely to appear in topic 271 followed by *chief* and *Indians,* and so forth. Another topic may also have the word *Indian* with a much smaller emphasis. In this topic, for example, we find the word *party.* In this context, we understand that *party* is being used in the sense of *war party.* Another topic, about celebrations, for example, might have the same word type, *party,* with a different weight and a different sense.

† Obviously, these are generalities: each of the word clusters contains potential subclusters or subtopics, and, given different parameters, the LDA process might be tuned to detect different levels of granularity. Also found in these collections of words are markers indicative of attitudes toward the overarching theme: *savages, scalp,* and perhaps even *death* in the context of the other words found in topic 271 give us an inkling of how the Indians of this topic are being represented in the corpus. Were we interested in pursuing a finer level of granularity—if we were interested, for example, in separating the depiction of Indians as peaceful frontiersmen from the depictions of Indians as scalping savages—we might rerun the topic model and set the parameters so as to define a larger number of topics, or we might use text segmentation to create smaller units of text for processing. These techniques will be described in more detail below.

Likewise, it appears that topic 25 contains the seeds of what might be two separate themes under the larger heading of "Ireland." Consider how a sublist of words including *heart, cabin, bog, county, mountains, Derry, house, air, evening, turf, friends,* and *heaven* presents a picture of the native Irish with perhaps some idealization of the Irish countryside. Contrast that with a sublist of words containing only *night, land, cabin, bog, sorrow, divil, peasantry, whiskey, manner, officer, barracks, agent,* and *officers,* and we can conjecture the presence of a topic with greater negative associations, one that leads us to consider how British rule in Ireland impacted the depiction of Ireland and provided a contrast to the more idealized Ireland. Taken together, the two sublists offer a broad view and capture the idea of Ireland as a bifurcated country: *One Island, Two Irelands,* as the title of a documentary film portraying the "Troubles" puts it, or *The Two Irelands,* as in David Fitzpatrick's study of Irish partition from 1912 to 1939.

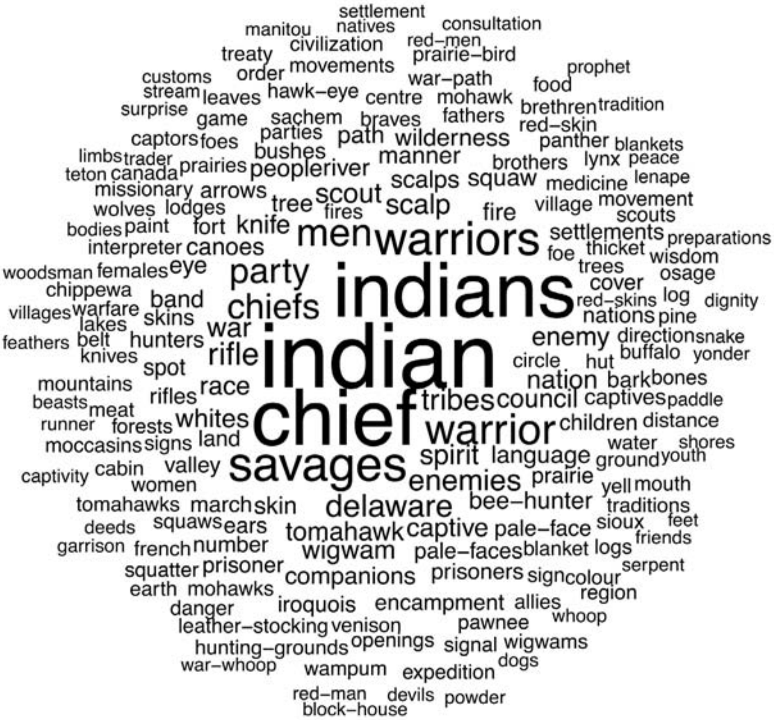


Figure 8.1. "Native Americans" theme

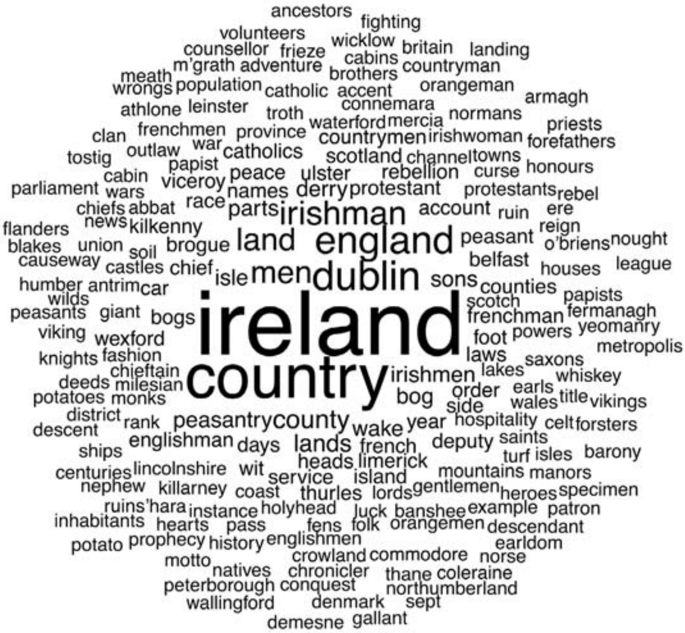


Figure 8.2. "Ireland" theme

In conjunction with the much larger company of other words that the model returns, it is easy to see that this particular use of *stream* is not related to the "jet stream" or to the "stream of immigrants" entering the United States in the 1850s. Nor is it a word with any affiliations to contemporary "media streaming." This *stream* refers to a body of flowing water. Its contextual relationship to other words such as *wilderness* and *prairie*—not to mention the more obvious *Indians* and *chiefs*—leaves us knowing with no uncertainty that the larger "theme" of which *stream* is one component is Native Americans. The words returned from the model paint both setting and subject; the result is a useful and quantifiable representation of a very particular theme. Importantly, though, note further that *stream* is not a word likely to arise in one's consciousness upon hearing the words *Native American. Stream* is, however, entirely appropriate in the context of these other words because it fits with and belongs in the theme. Were we conducting a traditional scholarly analysis of the "theme" of Native Americans in nineteenth-century fiction, would we have even thought to include streams? Consider the following passage from Mayne Reid's 1851 novel *The Scalp Hunters.* The first occurrence of the word *stream(s)* is located 225 words away from the occurrence of the word *Indians,* approximately the distance between the top and bottom of a standard book page:

> The scenery was altogether new to me, and imbued me with impressions of a peculiar character. The *streams* were fringed with tall groves of cottonwood trees, whose column-like stems supported a thick frondage of silvery leaves.

. . . [199 words removed here] . . .

> As we approached the Arkansas, we saw mounted *Indians* disappearing over the swells. They were Pawnees; and for several days clouds of these dusky warriors hung upon the skirts of the caravan. But they knew our strength, and kept at a wary distance from our long rifles.

The word *stream* does not occur "close" enough to the word *Indian* to have been picked up in a KWIC list, and even a collocation concordance would be unlikely to identify *stream* as an important word in the context of Indians, since *stream* is not going to be a very frequent collocate. Yet the topic-modeling methodology, which treats documents as "bags" of words, is not restricted from recognizing that there is something important and contextually homogeneous about *stream* and *Indians.*

The two topics described above were produced during a run of the MALLET topic modeler that sought to identify five hundred topics in a corpus of 3,346 nineteenth-century books. When running an LDA model, the researcher sets a variety of parameters to determine how the model is constructed. The primary parameter, and the one that causes no small amount of discussion and confusion, is the parameter determining the number of topics to be harvested from these data. There is neither consensus nor conventional wisdom regarding a perfect number of topics to extract, but it can be said that the "sweet spot" is largely dependent upon and determined by the scope of the corpus, the diversity of the corpus, and the level of granularity one is seeking in terms of topic interpretability and coherence. As noted previously, the "interpretability and coherence" of a topic mean specifically the ease with which a human being can look at and identify (that is, "name" or "characterize") a topic from the word list that the model produces. Setting the number of topics too high may result in topics lacking enough contextual markers to provide a clear sense of how the topic is being expressed in the text; setting the number too low may result in topics of such a general nature that they tend to occur throughout the entire corpus. Chang et al. (2009) have conducted one of the most compelling studies of topic interpretability. The authors test statistical and unsupervised approaches designed to evaluate topic coherence against qualitative human evaluations. Their conclusion is that automated methods for assessing topic interpretability are negatively correlated with human evaluations of interpretability. In other words, though the machine does a very good job in identifying the topics latent in a corpus, the machine does a comparatively poor job when it comes to auto-identifying which of the harvested topics are the most interpretable by human beings.

Interpretability is of key concern to us since we are most interested in investigating and exploring themes that we understand and recognize. This is an important point to dwell on because one criticism of unsupervised models is that they are black boxes. In one sense, it is true that with no, or very little, input from the human operator, the machine churns through a corpus and harvests out what it "believes" to be the "n" most important or prevalent topics in the corpus. From an information-retrieval perspective, this may be all that is required. If our goal is only to find documents with similar content, then the results of the topic model can be further processed so that texts with high proportions of "topic A" are grouped in one pile and those with strong proportions of "topic B" in another. This method has proven to work extremely well when it comes to document clustering and information retrieval (see, for example, Cohn and Hofmann 2001; Hofmann 2001; Wei and Croft 2006). For literary scholars, however, it is the interpretability and coherence of the topics that ultimately count. For illustrative purposes, figure 8.3 shows the word cloud generated for topic 123. Although it is true that an uninhibited reader might be able to make an argument for the interpretability of this topic, this set of words is clearly something other than what was seen in the topics of figures 8.1 and 8.2. Out of five hundred topics derived in this research, I identified fourteen as being "uncertain" or "unclear."* What, then, does this mean—that

* Word clouds for all five hundred topics can be found online at http://www.matthewjockers.net/macroanalysisbook/macro-themes/.

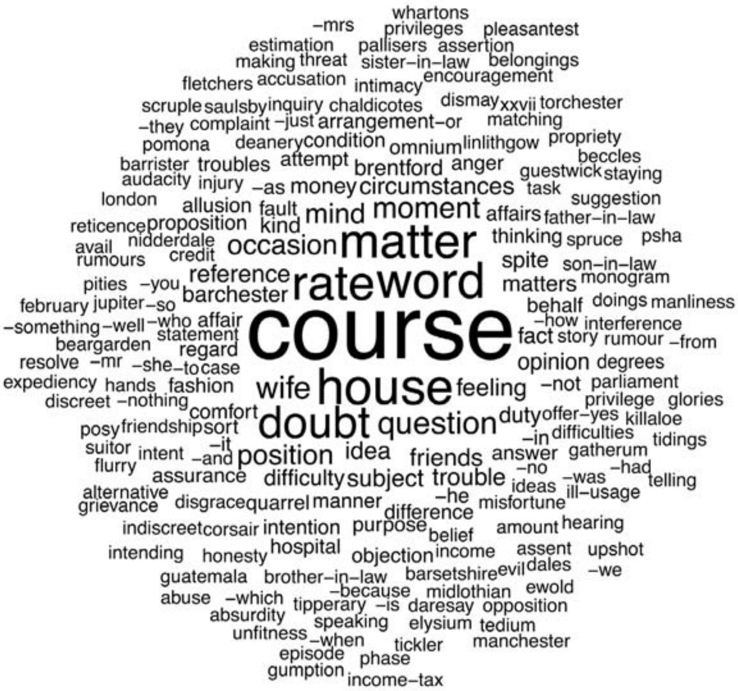


Figure 8.3. An incoherent theme

some topics lack clarity—and how are we supposed to trust the algorithms if alongside two homogenous, pristine, interpretable topics we also find these impenetrable juggernauts?

For computer scientists and linguists working in the field of topic modeling, this is an active area of research, but it is not a problem for us to dwell upon here. Ideally, all of the topics the machine produces would be perfectly interpretable; the presence of uninterruptable topics, however, does not undermine the usefulness of the topics that are interpretable.* If we choose to ignore some topics, say the ambiguous one, and focus our attention on others, say only the most interpretable, we do no disservice to the overall model, and we in no way compromise our analysis. Choosing to examine only some of the machine's

* For information-retrieval researchers and scholars working in machine learning, the problem of uninterruptable topics is a nut to be cracked, and a variety of approaches are being explored: for example, approaches that attempt to automatically score topic coherence. It is fair to say that the more interpretable the resulting topics, the better the model's use, and so there is legitimate motivation to improve the modeling algorithms to a point where they turn out only pristine topics. See, for example, Lau, Newman, et al. 2010; Lau, Grieser, et al. 2011; Newman, Lau, et al. 2010; Newman, Noh, et al. 2010.

topics and not others is a legitimate use of these data and should not be viewed with suspicion by those who may be wary of the "black box." Indeed, as literary scholars, we are almost constantly focusing our gaze on the specific elements of a text that we consider to be the interpretable elements. That we may do so at the expense of some other valuable insight we might gain by focusing our attention elsewhere is understood. Put simply, not everything in a book can be studied at once: *Moby Dick* explores both whaling and religion, but a scholar writing of religion in *Moby Dick* cannot be faulted for failing to discuss cetology. The place where error, or more correctly "debate," can creep into the topic-modeling process is in the assignment of thematic labels, that is, in the human interpretation of the machine-extracted topics. Some interpreters will disagree over the appropriate label to assign to a given word-topic distribution, but these disagreements will almost always be over the precision, or imprecision, of the human-generated topic labels and not over the underlying theme represented by the words in the topic. For this research, I have found it useful to visualize the topic-word distributions as word clouds. This has the effect of accentuating those words that are most central to the topic while pushing the related but less central words to the periphery of the visualization. In general, the labels that I have assigned to the five hundred topics I explore in this research are derived from the primary, or "key," words revealed by these topic clouds. In some cases, however, my labels are a synthesis of the words, that is, they are labels that use words not necessarily found in the topic, labels designed to offer a more general summary: the label "Native Americans," for example, is what I use to describe the cluster of words in figure 8.1.* As noted above, the labeling of topics is subjective and can be contentious; for this reason, I have placed all five hundred topic clouds online. Readers wishing to verify my labeling choices are encouraged to study the topic-cloud visualizations.

With that said, let us unpack the topic-model results a bit further. In addition to outputting the word clusters seen in the figure, the model also provides output indicating the percentage or proportion of each topic assigned to each document. Remember that each document contains a mixture of topics in different proportions, like a pie in which some pieces are larger than others. The model provides similar information for the corpus as a whole. Figure 8.4 shows the distribution of the top-ten topics found in *Moby Dick* alongside the same topics in the entire corpus. The presence of topic 347 ("Seas and Whaling"), a theme dealing with islands, whaling, and sea voyages, is prominent in *Moby Dick:* according to the model, about 20 percent of the novel. This seafaring theme is followed by theme 85, which deals with ships. Next is theme 492 about boats

* This is a particularly useful label, given that another topic deals with India and another set of "Indians."

and water, and then 455, a theme that deals with captains and crews. In *Moby Dick,* these are big themes—in the entire corpus, however, they are barely noticeable, as seen in the darker gray bars representing the corpus mean for these themes. Figure 8.5 shows the top-ten topics from the overall corpus alongside those same topics in *Moby Dick.* With the exception of the themes associated with morning and night, the themes that dominate the overall corpus are comparatively foreign to *Moby Dick.* Indeed, *Moby Dick* is very much an outlier in terms of the themes that dominate this nineteenth-century corpus.

With a basic sense of how topic modeling works and what it produces, we can now delve into my specific application of LDA to this corpus of 3,346 nineteenth-century books. When running a topic model, a researcher has the option of excluding "stop words" from the analysis. Most commonly, stop words are defined as high-frequency "function" words such as articles and prepositions. The MALLET software provides a setting that will ignore a standard list of very common English adverbs, conjunctions, pronouns, and prepositions. Unlike the stylistic analysis of the previous chapter, where we sought to exclude context-sensitive words, in topic modeling we want to do the exact opposite: we want to remove words that carry no thematic weight and concentrate on those words that best convey meaning. For this analysis of theme, I determined that even the standard stop list in MALLET was unsatisfactory. In addition to the usual set of common words, many other word types can "pollute" an otherwise pristine thematic signal. The repeated use of a common name in multiple books, "John," for example, can come to dominate a topic that would otherwise be purely topical in nature ("John" from one book being not the same as "John" from another). Prior to topic modeling the texts for this research, therefore, I employed the Stanford Named Entity Recognition software package (Finkel, Grenager, and Manning 2005) to identify named entities in the corpus. These character and personal names, identified by the NER software package, were then added to the list of common stop words. Ultimately, the stop-word list totaled 5,631 distinct word types.* Marks of punctuation and numbers were also specifically excluded.

Through experimentation it was observed that even finer-grained and more interpretable topics could be derived through a further reduction of the vocabulary being modeled. For modeling theme, I discovered that highly interpretable and thematically coherent topics could be derived through a model built entirely from nouns.† Constructing a noun-based model required preprocessing all of

* See http://www.matthewjockers.net/macroanalysisbook/expanded-stopwords-list/.

† Depending on what one wishes to analyze in the topics, the exclusion of certain word classes could be viewed as a controversial step. By eliminating adjectives, for example, the resulting topics may lack information expressive of attitudes or sentiments. Likewise, the

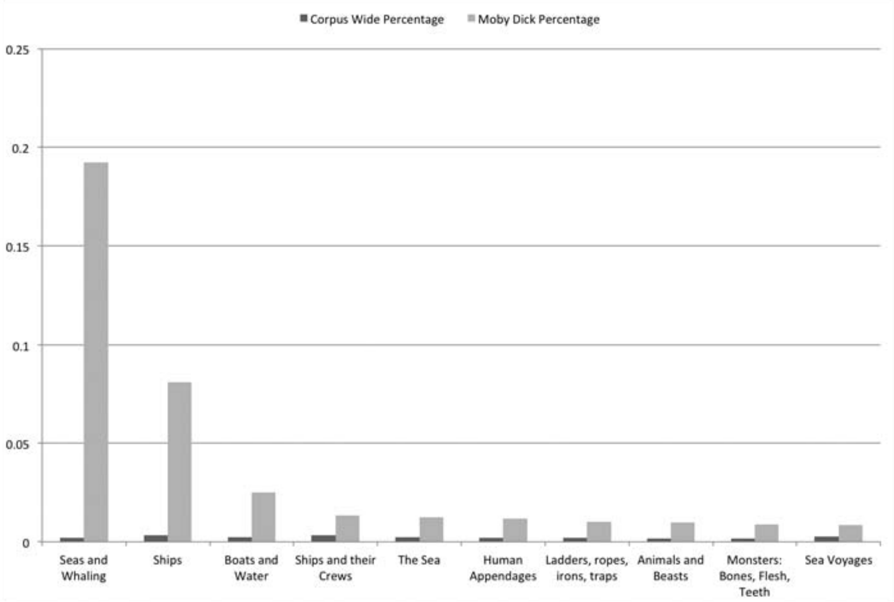


Figure 8.4. Top-ten topics in *Moby Dick*

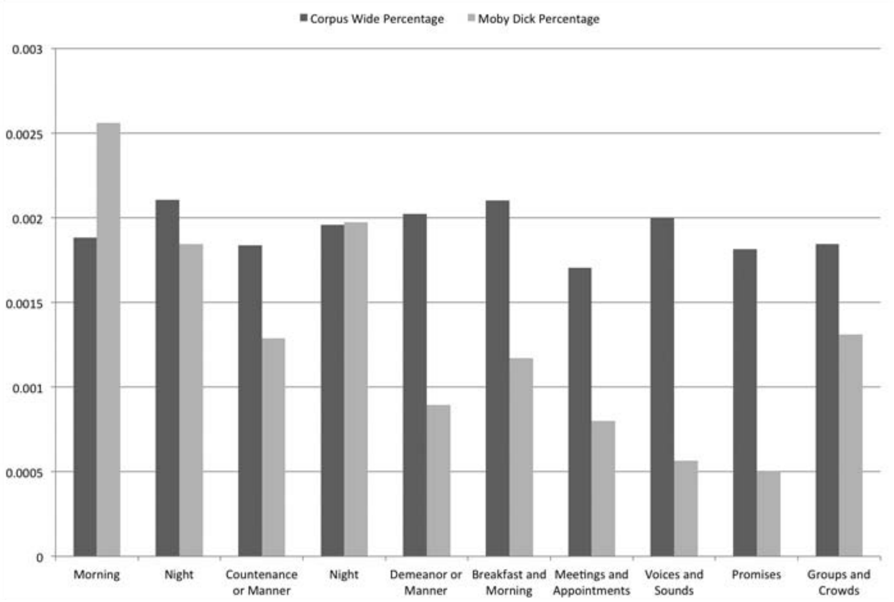


Figure 8.5. Top-ten topics in corpus

the texts in the corpus with a part-of-speech (POS) tagger.* In this manner, Austen's "It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters" is transformed into "truth man possession fortune want wife feelings views man neighbourhood truth minds families property other daughters." Once the 3,346 texts in the corpus were tagged for part of speech, a custom script chunked each text into 1,000 word segments and then extracted the words identified as nouns from each segment and created a new corpus of text segments composed entirely of

elimination of verbs may result in topics that fail to capture the actions associated with the novels. I must offer the caveat, therefore, that the noun-based approach used here is specific to the type of thematic results I wish to derive; I do not suggest this as a blanket approach. In my experiments, I have often found that a combination of specific parts of speech can reveal aspects of narrative and narrative style that are not captured by nouns alone. Nevertheless, the noun-based approach that I describe here proves to be extremely effective in generating coherent topics that can be easily identified and interpreted. The effectiveness of nouns for this purpose was determined through an iterative process of trial and experimentation. Various combinations of different parts of speech were tried and examined. Each combination had certain advantages and disadvantages. For the purpose of modeling theme, nouns produced the best results. This will be made clear in the analysis that follows. Verbs were discovered to have only limited value, and though adjectives were found to be highly effective in capturing sentiment, they were not effective in capturing or further elucidating theme. The development and use of an adjective-based model for detecting sentiment alongside theme is a current area of experimentation in a project of the Stanford Literary Lab that is funded by the Mellon Foundation.

* There are a variety of POS taggers available for this kind of work, and all of them have advantages and disadvantages. The Stanford POS tagger, for example, that was used in this research is known to be highly accurate, but since it is trained on hand-tagged samples of journalistic prose from the *Wall Street Journal,* it may not be as sensitive to literary prose as another tagger trained on a corpus of nineteenth-century fiction. Modern part-of-speech taggers rely on supervised machine learning and machine classification. In simple terms, the tagger is given a training corpus of documents that have been marked up by a human coder, and once "trained" on this corpus, the tagger is then given new, unmarked, text for which it assigns part-of-speech tags based on what it has learned from the training data. I have not done extensive comparisons to determine whether one or the other tagger is "better." Matthew Wilkins provides a useful review of several taggers on his blog at http://workproduct.wordpress.com/2009/01/27/evaluating-pos-taggers-conclusions/. See also http://nora.hd.uib.no/corpora/1997–3/0161.html and Boggess et al. 1999.

nouns. MALLET was then employed to model the corpus and extract 500 latent topics from this noun-based and segmented corpus of 631,577 novel chunks.*

Text segmentation was done in order to improve topic quality and interpretability. Topic modeling treats each document as a "bag of words" in which word order is disregarded. Without segmentation, a single novel would be processed as a single "bag." Since the topic model works by identifying words that tend to co-occur, the bigger the bag, the more words that will tend to be found together in the same bag. If novels tended to be constrained to only a very small number of topics or themes, then treating an entire novel as one bag might be fruitful. In reality, though, novels tend to have some themes that run throughout and others that appear at specific points and then disappear. In order to capture these transient themes, it was useful to divide the novels into "chunks" and run the model over those chunks instead of over the entire text. There appears to be no conventional wisdom regarding ideal text-segmentation parameters. David Mimno—author of the MALLET topic-modeling software—reports in email correspondence that he frequently chunks texts down to the level of individual paragraphs. For the books in this corpus, I found through experimentation that 1,000-word chunking was effective.† Knowledge of the scope and variety of the corpus, along with a sense of the research questions to be investigated, proved to be the most useful guides in determining an ideal number of topics and ideal degree of text segmentation. Until new research provides an algorithmic

* In work sponsored by the Mellon Foundation, this process was later formalized by Loretta Auvil and Boris Capitanu as a SEASR workflow.

† Other experiments included building models based on dividing each novel into ten equal-size chunks, page-based chunks, paragraph chunks, and 250-consecutive-noun chunks. None of these produced topics of greater quality (interpretability) than what was observed in the 1,000-word chunking. In fact, some topics that represent major themes in specific novels were "lost" when segmentation became too fine grained. It was observed, for example, that the theme "Whaling Ships" disappears from the topic results when the chunk size becomes too small. This is because the theme of whaling in *Moby Dick,* for example, tends to get expressed across the breadth of the entire text and to a lesser degree in specific sections. If the chunks get too small, the concurrence of whaling terms within a chunk do not rise to a level worthy of "topic status"; instead, the whaling terms gets absorbed into larger themes associated with "seafaring." For this corpus, segmentation of each text into 1,000-word chunks produced high-quality results. Undoubtedly, these results could be improved still further, but ultimately it is the interpretability of the topics that is important. In this research, experiments that made small adjustments to text chunk size or number of topics, or both, resulted in trivial or unperceivable change. Experimentation with segmentation revealed very little difference between 500 topics generated using texts that are segmented into 1,000-word chunks versus using 1,500-word chunks.

alternative, trial and experimentation augmented by domain expertise appear to be the best guides in setting segmentation and topic-number parameters.*

MALLET output includes two derivative files: a file containing topic "keys" and a file containing the proportions of each topic found in each text, or each text segment in this case. The keys file is a simple matrix in which the first column of data contains a unique topic identifier, and a second column contains the top words assigned to the topic. A researcher examines these topic-word distributions and assigns them a label. In this work, I have found it useful to visualize the topics as word clouds. In some cases, an appropriate label can be difficult to determine. In this research, I labeled ambiguous topics as "uncertain," and then as a check to the legitimacy of my other labels, I consulted with members of the Stanford Literary Lab to confirm the appropriateness of my choices. There was general consensus about the interpretability and labeling choices I had made. Of the 500 topics, I labeled four as "Bad Data." These were topics that resulted from the poor quality of optical character recognition of some texts in the corpus. Another four topics were labeled "Book Metadata." These topics were obviously derived from the words found in the title pages and metadata of the electronic files. I identified 18 topics as "uncertain." The remaining 474 were found to be interpretable and were given thematic labels.† The second derivative file that MALLET produces provides data regarding the amount (proportion) of each topic found in each text segment. The modeling process assumes that every document is a "mixture" of all of the 500 possible topics in the corpus. Thus, each document is composed of some proportion of each of the 500 topics.‡

Motivated by the work of the Veselovsky brothers and their interest in studying literary evolution in terms of recurring motifs and national literatures, I began my analysis by plotting the mean proportions of every topic, in every year, separated first by nation, then by gender, and finally by nation and gender combined. Linking all of the thematic and topical data to the metadata facilitated

* David Mimno and others are presently working on exactly this sort of research. Tuning the algorithm to automatically identify the best parameters requires training data for which the topics have been manually vetted for quality. In the course of my research, I have produced, studied, and labeled several dozen topic models, and I have given many of these results to Mimno for analysis. Using my labeled models, Mimno is able to study the differences between topics that I have identified as interpretable versus ambiguous and look for patterns in the word distributions of the two. In time, Mimno hopes to develop an algorithmic solution for generating the most coherent results.

† www.matthewjockers.net/macroanalysisbook/macro-themes/

‡ This proportions file is a table in which each row is a text segment and each column a topic. The proportions of each topic in each text segment are the values held in the individual cells.

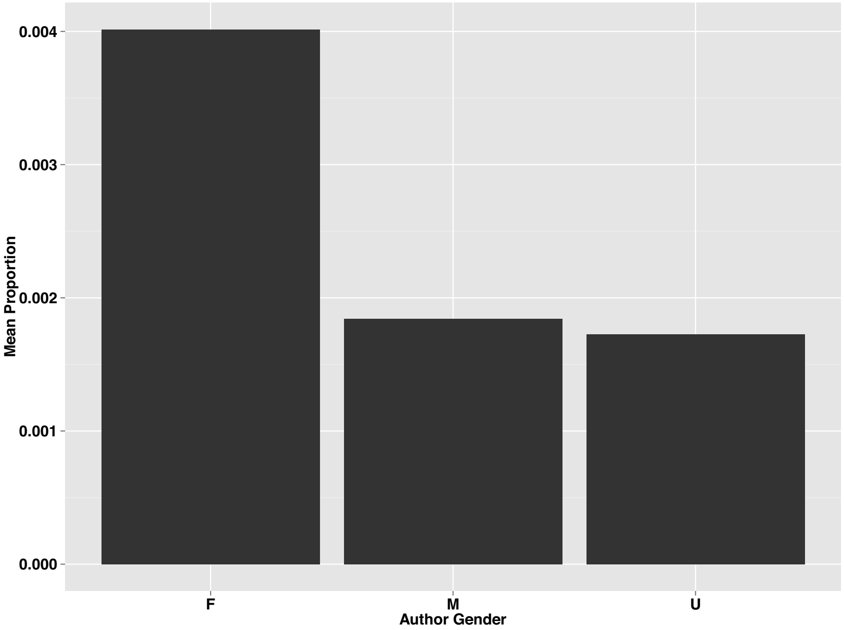


Figure 8.6. "Affection and Happiness" theme by gender

the identification of thematic and topical patterns at the level of the corpus, at the level of the individual book, and across facets of time, gender, and nationality. All 474 interpretable themes were plotted across each metadata facet; the resulting charts can be found on the companion website.*

Examining the charts, a number of things were obvious, not the least of which was that nations and genders have clear thematic "preferences" or tendencies. Some of these preferences correspond rather closely to our expectations and even our stereotypes. For example, a theme associated with strong emotions and feelings that I labeled "Affection and Happiness" appears more than twice as often in female authors as in male authors (figure 8.6). Female authors also tend to write more about women's apparel, as evidenced by their greater usage of topic 29 (figure 8.7), which is a theme I have labeled "Female Fashion." And women are at least twice as likely to write about the care of infants and children as their male counterparts. Figure 8.8, labeled "Infants," shows this topic distribution by author gender.

Not surprisingly, women also appear to have the market cornered when it comes to even more specific expressions of strong emotion. Whether these are expressions of "Happiness" (topic 109), "Passion" (topic 316), or "Grief and

* www.matthewjockers.net/macroanalysisbook/macro-themes

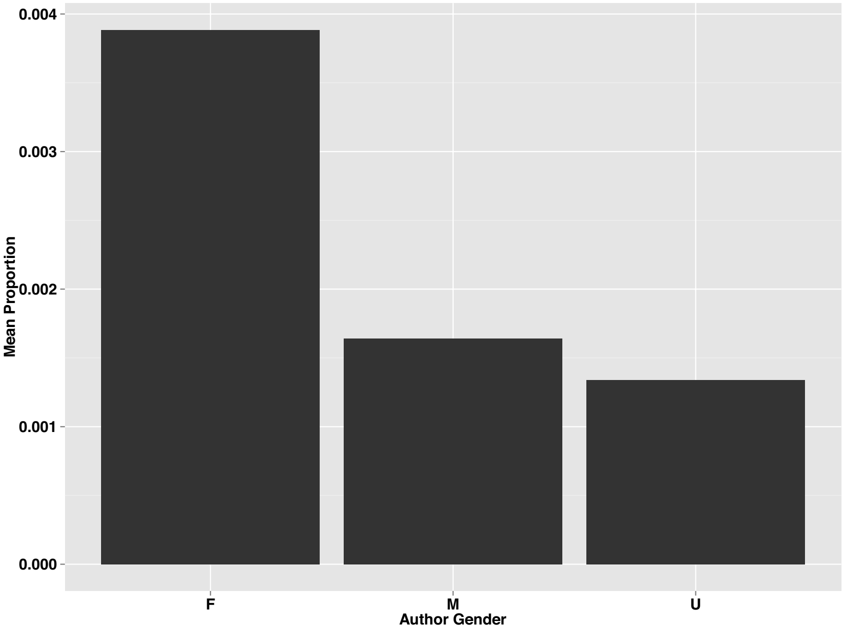


Figure 8.7. “Female Fashion” theme by gender

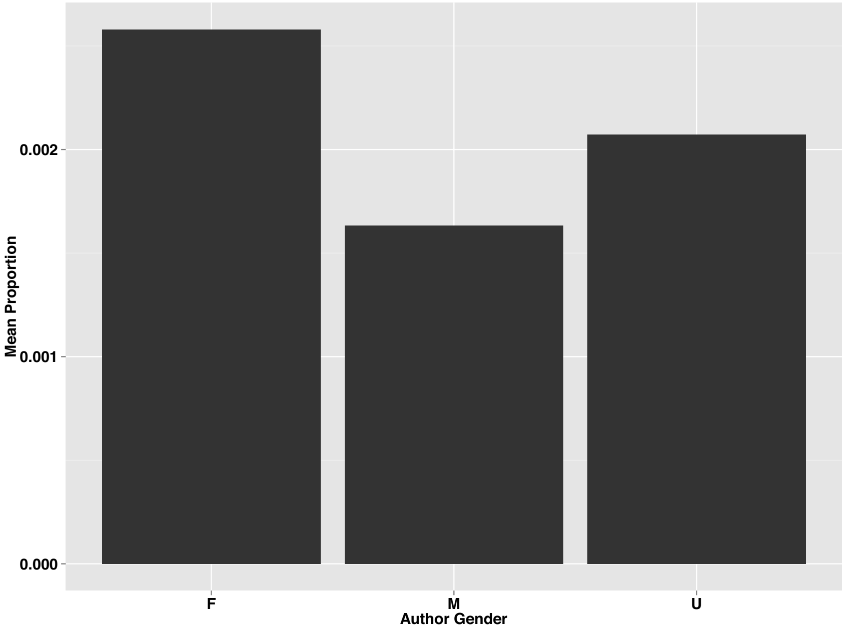


Figure 8.8. “Infants” theme by gender

Sorrow" (topic 43), female authors tend to deal with expressions of feeling and emotion far more often than males. Male authors, on the other hand, are about twice as likely as women to write about "Villains and Traitors" (figure 8.9), about "Quarrels and Dueling" (figure 8.10), and about "Enemies" (figure 8.11).

Other themes in the corpus are not specifically gendered. The use of a rural theme associated with scenes of natural beauty (topic 471) is equally distributed among male and female authors. So too is a related theme of "The Land" (topic 190). And three themes associated with wealth, business affairs, and social rank (topics 338, 359, 375) are all equally distributed across genders. Male and female authors in this corpus are also equally likely to write of humor, jesting, laughter, and joking (topic 381).

The works in this corpus include 1,363 by female authors, 1,753 by males, and 230 of unknown authorship (the "unidentified" gender class labeled as *U* in the figures). It turned out that these anonymous works are quite interesting. Indeed, several themes in the corpus are overrepresented in works of anonymous authorship, and these themes often relate to sociopolitical institutions such as the monarchy, as seen in topics 97, 108, and 239; or to religious institutions, as in the "Convents and Abbeys" theme found in topic 31; or the theme of "Religion"

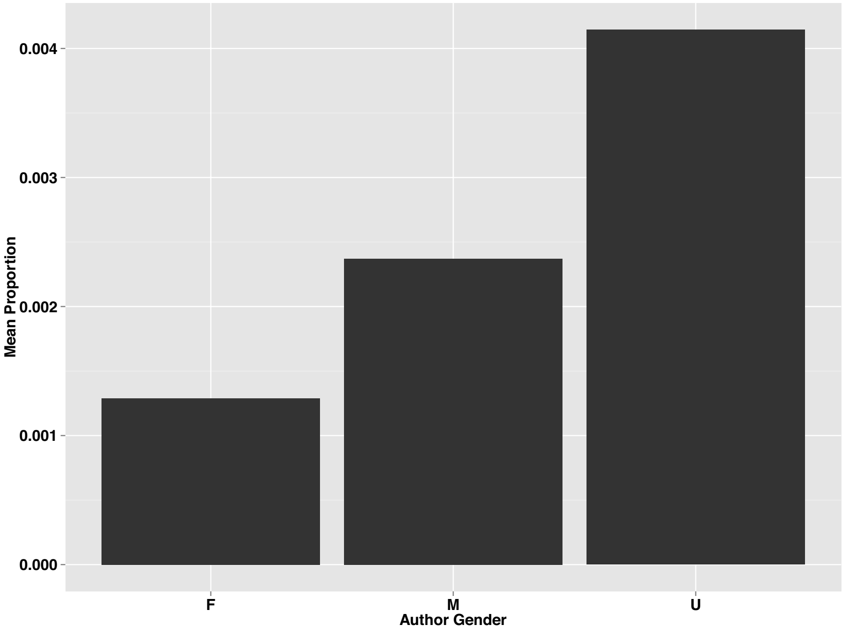


Figure 8.9. "Villains and Traitors" theme by gender

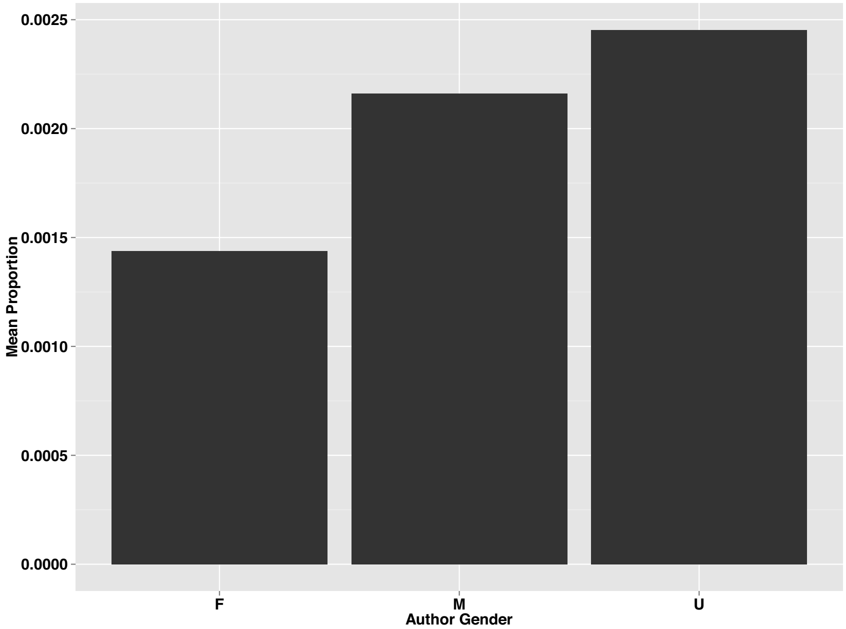


Figure 8.10. “Quarrels and Dueling” theme by gender

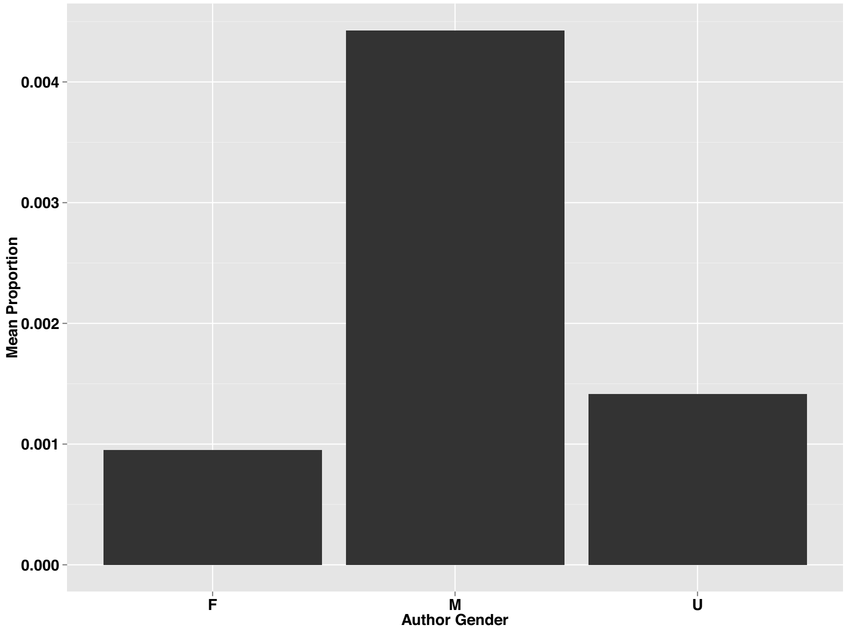


Figure 8.11. “Enemies” theme by gender

found more generally in topic 448. At the same time, these anonymous works are also very high on a theme dealing with the expression of opinions, topic 28. All of this makes perfect sense if we believe that these authors felt the need to conceal their identities in order to present a more candid portrait of politics or religion or both. This is an undoubtedly rich area for further research, and a closer probing of this data led me to a further observation that these anonymous writers have a higher usage of topics 25 and 324, which are "Ireland" and "Scotland," respectively. Taken together, this evidence suggests a class of writers generating thinly veiled narratives that express opinions about religious and nationalistic matters that would have likely been awkward or impossible to express without the use of a pseudonym.

Such an observation naturally leads us to inquire about other correlations between nationality and theme. Sadly, for these unidentified authors we know neither their genders nor their nationalities, and thus it is impossible to know if these writers were themselves Irish or Scottish. For the majority of authors in this corpus, however, we do have information about nationality, and this data can be used much in the same way that gender was used to plot the thematic preferences of male and female authors. When it comes to nationality, some themes were discovered to be predominantly American, others British, and some distinctly Irish.* Prominent in American writing, for example, is a theme associated with slavery† (figure 8.12) and another associated with American Indians (figure 8.13).

An especially Irish theme, which I have labeled "Tenants and Landlords," is found most often in Irish novels and is, in fact, the most prominent theme (that is, the topic given the greatest proportion) in 34 percent of the Irish novels in this corpus (figure 8.14). This theme will be familiar to scholars of Irish literature as the "Big House" theme that was made popular by writers exploring the relationship between ascendancy landlords and their typically poor, Catholic tenants. Topping the list of Irish authors who deploy this theme are Maria Edgeworth (whose *Castle Rackrent* registers 5.2 percent) and William Carleton (whose *Poor Scholar* scores 5 percent). In other words, according to the model, fully 5 percent of each of these books is concerned with the relationships between tenants and landlords. And should we wish to interrogate the usage of this theme at the smaller scale of the individual novel, it is possible to plot the progression, or "incidence," of the theme across the 1,000-word segments of each text. Figures 8.15 and 8.16 plot the usage of the "Tenants and Landlords" theme across "novel time" in the two books.

* Unless specifically noted, Scottish and Welsh authors are generally treated as part of the "British" corpus.

† Two themes in this corpus deal with slavery: one with an American context and the other Persian.

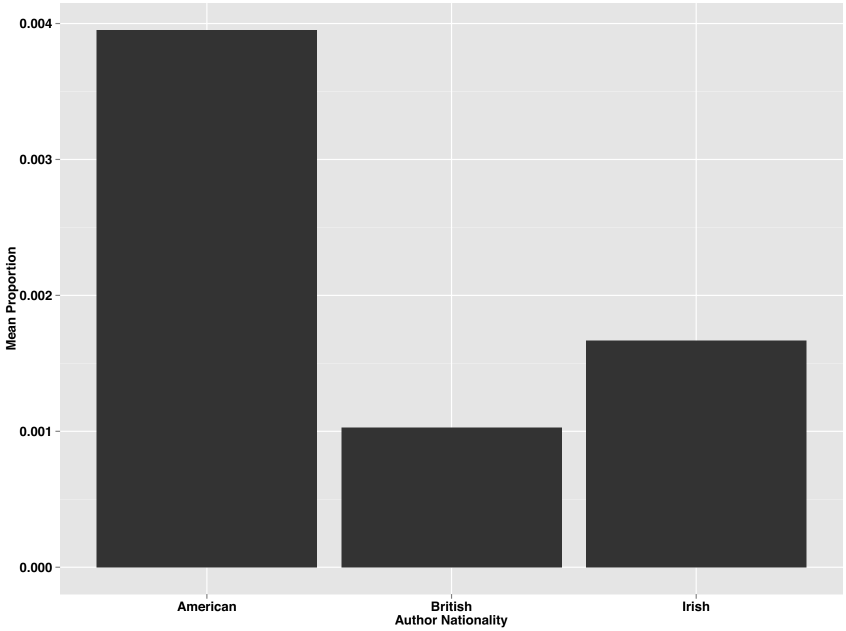


Figure 8.12. “American Slavery” theme by nation

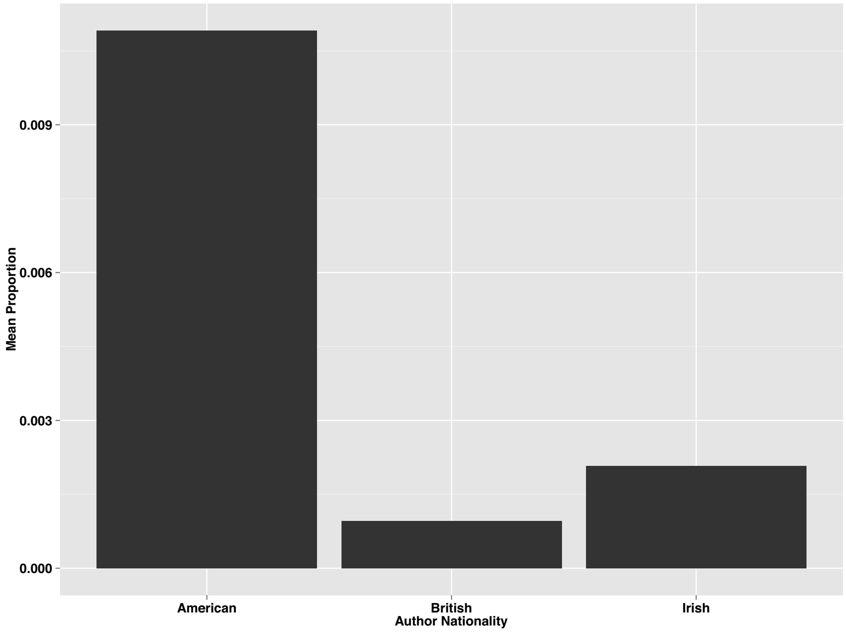


Figure 8.13. “Native Americans” theme by nation

By comparison, this Big House theme, which appears as most important, or most dominant, in 34 percent of Irish-authored novels, appears as dominant in just 0.007 percent of British novels and just 0.003 percent of American novels. Although it is a defining theme of the nineteenth-century Irish novel, it is not omnipresent throughout the entire century. Figure 8.17 shows the "Tenants and Landlords" theme separated by nation and plotted across time. To reduce noise, a five-year moving average was applied to the data. Among the Irish-authored texts in the corpus (the light-gray line), this theme experiences four distinct spikes. The first spike occurs around 1810; this is the time frame within which Maria Edgeworth's *Castle Rackrent* (1800) and Lady Morgan's *Wild Irish Girl* (1806) were published: we would expect to find the theme well represented in these texts. A second, smaller, spike appears in 1835. Then the topic peaks in the 1840s during the years of the Great Famine and a time when writers including Carleton, Lever, and Lytton were active. Whether the famine was also the catalyst for the simultaneous spike seen in American writers of the same period is a question for further investigation. Certainly, the influx of famine immigrants to America would have been very hard to ignore. That this catastrophe may have served as specific fuel for the American literary imagination is another

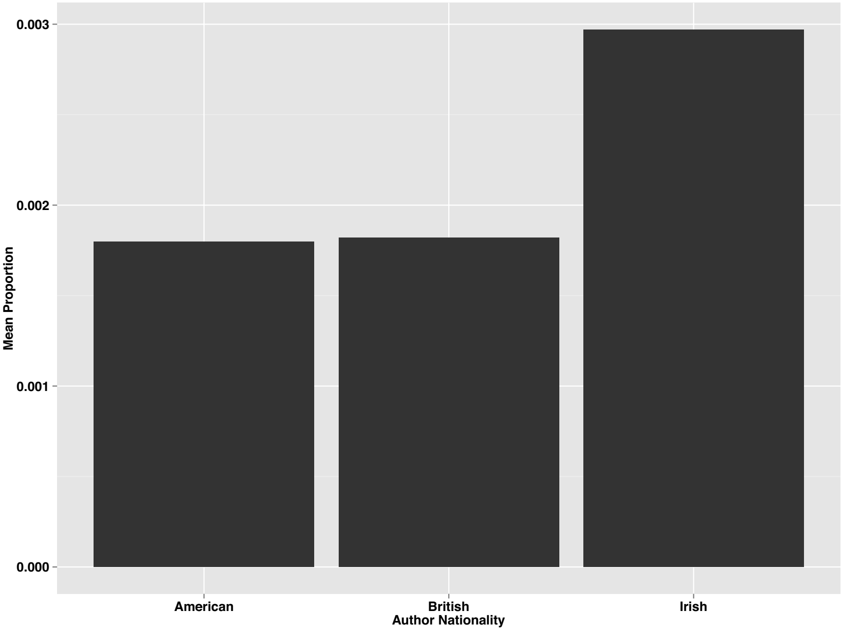


Figure 8.14. "Tenants and Landlords" theme by nation

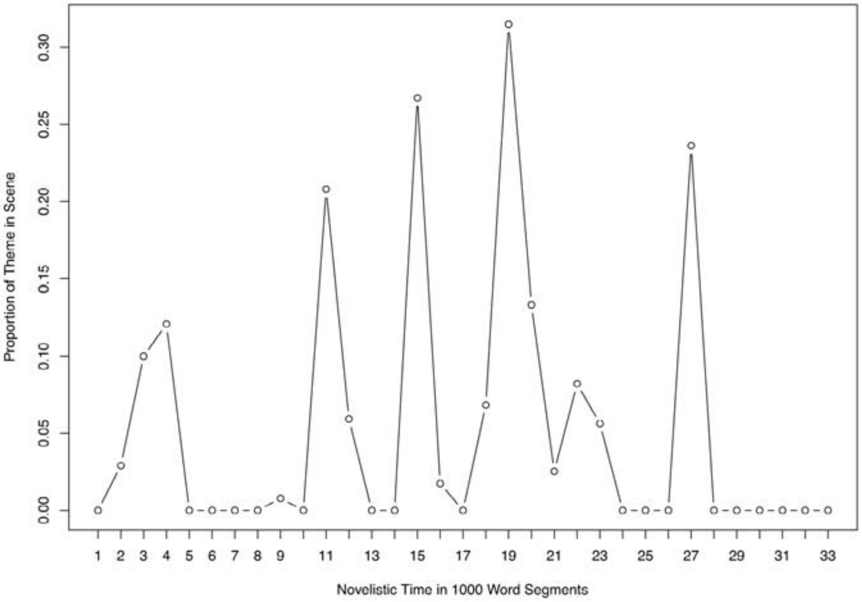


Figure 8.15. "Tenants and Landlords" in *Castle Rackrent*

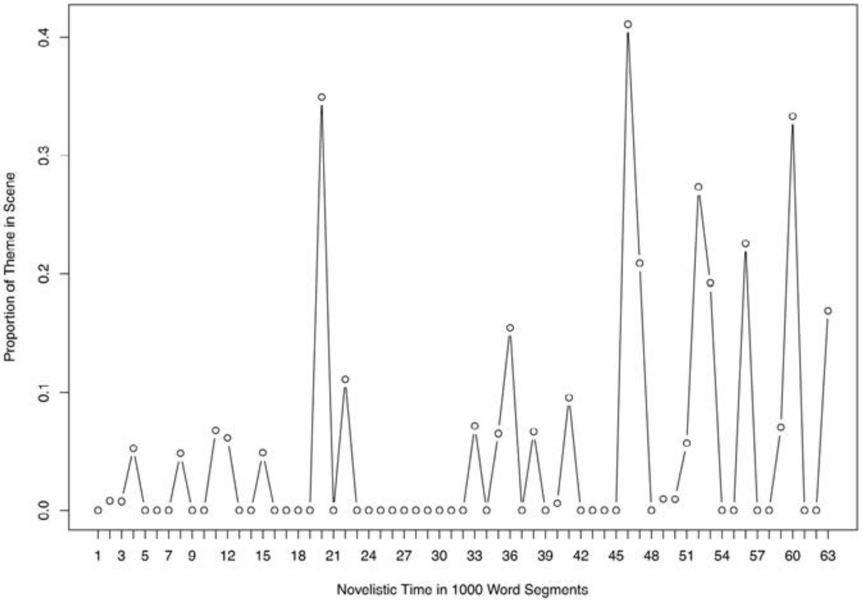


Figure 8.16. "Tenants and Landlords" in *The Poor Scholar*

matter. With James Fenimore Cooper, at least, we know of a looming dissatisfaction with the American aristocracy and a desire to avoid in America the kinds of disparity he witnessed in Europe in the '30s. One Cooper novel from this period, *The Redskins,* goes to the heart of "tenant-landlord" systems, and readers cognizant of the situation in Ireland at the time cannot help but note that Cooper's fictional footman, Barney, is just thirteen weeks out of Ireland.*

This spike of the 1840s is followed by a steep decline into a trough that runs from the 1850s until the late 1860s. Such a decline may very well be attributable to the aftermath of the famine. This was, after all, a national catastrophe that would have made writing fiction about strained tenant-landlord relationships seem almost cruel and unusual punishment for a devastated population. A final spike of increased writing about tenants and landlords is seen in the 1860s, where again it seems to have parallel currency in the American context.

In *The Irish Novel: A Critical History,* James Cahalan asserts that even in the work of Irish writer Laurence Sterne—who did not write novels specifically about Ireland—he, Cahalan, can identify the "features of a particular kind of Irish novel" (1988, 7). Cahalan argues that Irish novels are often unified and identifiable by distinctly "Irish" themes, even if those themes are not exclusively Irish. He argues, for example, that a theme of subjugation runs through Irish prose. Certainly, this "Tenant and Landlords" theme, which includes such terms as *rights, conditions, rents, landlord, servant, tenants, agents, grievance, complaints, taxes, neglect,* and *bailiff* could be seen as related to subjugation along the lines that Cahalan hypothesizes. And although Irish authors dominate this corpus when it comes to use of this theme, there are other themes of subjugation, which also find expression in the Irish corpus. The theme "American Slavery" is especially interesting. It is a theme most often found, as we would expect, in American prose. Nevertheless, the theme gets expressed rather profoundly in the Irish corner of the corpus, especially during the late 1850s and 1860s. This

* Michael J. Pikus has a revealing essay on this subject titled "*The Redskins;* or, *Indian and Injin* and James Fenimore Cooper's Continuing Historical Paradox." Pikus writes, "Despite his appreciation of the American aristocracy and landholding classes, the power of these classes in Europe dismays Cooper" (1997, n.p.). He cites Cooper as follows: "In America property is taxed as it should be . . . ; but in Europe as much is extorted as is possible from the pittance of the laborer, by means of excises. It is not unusual to term the political contests of the world, the struggle of the poor against the rich; but in Europe it is, in fact, the struggle of the rich against the poor. Governments, in this quarter of the world, are in fact degenerating into stock-jobbing companies, in which the mass are treated as so many producers to enable the few to get good securities for their money" (*The Letters and Journals of James Fenimore Cooper,* 2:345–46, cited in ibid., n.p.). Pikus goes on to write that the "sensitivity [Cooper] displays toward the European working classes reflects a benevolent republicanism" (ibid., n.p.).

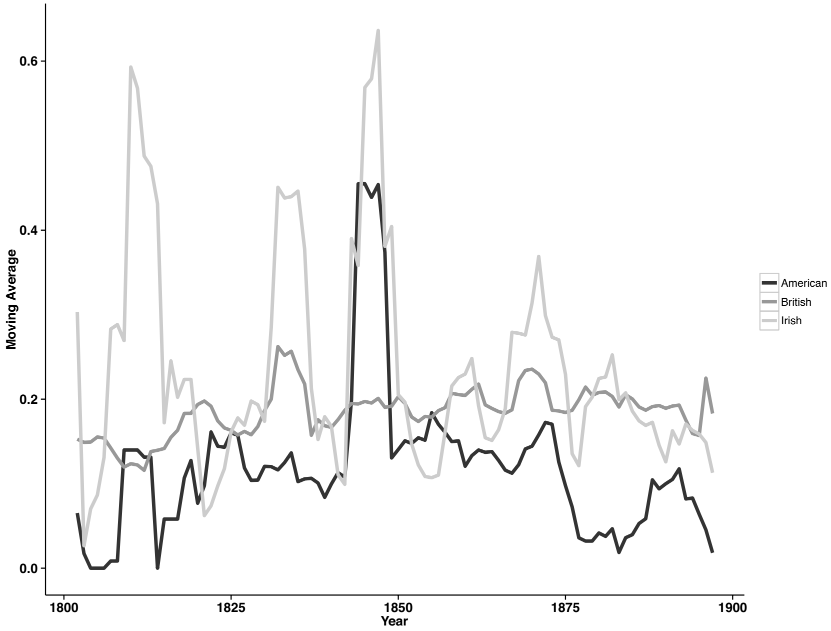


Figure 8.17. “Tenants and Landlords” across time and nation

surge is due almost entirely to the works of the very prolific Irish author Mayne Reid. Reid’s books from this period include the renowned antislavery novel *The Quadroon* (1856), which was somewhat shamelessly adapted for the stage by the more famous Irish author and playwright Dion Boucicault and retitled *The Octoroon.* Boucicault’s play, more so than Reid’s novel, was responsible for fueling debate in Ireland about abolition and the morality of slavery.* The American slavery theme is most prominent in the American part of the corpus, but it shows up twice as often in Irish books than it does in British books, as was seen in figure 8.12.

Thus far, my focus has been on describing thematic *presence,* but the thematic tendencies of nations and of genders may be seen not only in the themes that are overrepresented in their novels but also in those that are conspicuously absent. Despite the significant role played by ships and seafaring in Ireland, for example, the Irish novels in this corpus are almost completely lacking when it comes to themes associated with the sea. And Irish novels have much less than their British contemporaries when it comes to themes associated with the leisure activities of

* Irish writers, if not Irish American writers, were generally sympathetic to abolition and appeared to identify with the plight of black slaves in America.

the upper class. Somewhat ironically for a largely rural nation, the Irish authors in this corpus also have little to say about sheep, cattle, and livestock.

To be sure, there is much more that could be said about the general trends and tendencies of themes in this corpus of novels. The aforementioned themes are cited primarily as examples, as a way to give a fuller sense of the data, but they are also noted to serve as provocations for further thought. The thematic distributions are revealing; they incite all manner of questions about the relationships and the correlations between themes and historical events, between themes and author nationality, between theme and author gender, and more. The macro trends visualized here provide context, not on a keyword scale, but on a massive corpus-size scale.

• • •

What, then, can we learn of literary history from these trends? Surely, the things that writers choose to write about are in some sense tied to time and place. A good deal of literary criticism is invested in precisely this argument, that literary texts are historical artifacts, that literary texts are, whether overtly or more subtly, representative of the times in which they are produced. Where we may have been surprised to find that there are clear and measurable differences in word usage among British, Irish, and American writers, we are less surprised to learn that these authors write about different things, or that they write about similar things to differing degrees. Some themes are endemic to particular nations or to particular author genders; other themes cross both geographic boundaries and gender lines. Although this may not come as an incredible surprise to readers, and even less so to scholars of these three national literatures, an examination of exactly how connected the thematic signals are to factors such as gender and nationality turns out to be quite shocking.

The thematic data harvested here provide five hundred thematic data points, that is, five hundred measurements of thematic content for every novel in the corpus. Each of the five hundred variables is measured on a scale from 0 to 1, and each measurement represents a proportion of that thematic variable, or "topic," in a given text: in other words, the sum of all five hundred features for a given text will always equal one. In cases where the value of a topic is very low, or even zero, we can assume the theme is absent, or comparatively absent, from the text. In cases where the value is high, such as the 20 percent value assigned to the "Seas and Whaling" topic in *Moby Dick,* we can assume that the theme is prominent. Importantly, though, all the thematic values in between the maximum and the minimum values provide useful, even essential, information about the contextual makeup of the texts. A child may have only one freckle, but that freckle may carry valuable identifying information. Whether a genetic mutation, a paternal inheritance, or the result of too many hours spent at the beach, that freckle is a part of the person and tells some part of a larger story. Recalling the "LDA Buffet" fable from the website,* consider how a single thematic variable is like one item in a buffet of themes. Altering the quantity and type of one item changes the resulting meal: add one part of topic *X,* two parts of topic *Y,* and you get *Persuasion;* double the amount of topic *X,* and you get *Sense and Sensibility;* add several parts of topic *Z,* and you get *The Woman in White* by Wilkie Collins. This is a simplification, but conceptually it is a useful one.

The previous chapter's analysis of style has shown that there is more to novels than their thematic content; content is but one component in a complex, creative recipe that includes technique, language, theme, and other factors we have not yet analyzed (including, for example, plot, character, sentiment, and so on).† Like the cook who employs only locally grown ingredients, however, the elements of style and theme employed by nineteenth-century novelists turn out to be largely contingent upon and even determined by local conditions. This is not to say that writers cannot and do not import exotic elements into their fiction; they do. However, as we shall see shortly, the extent and degree (a teaspoon or a cup) to which an author employs a particular theme or writes in a particular style are closely correlated to the "raw materials" available in that author's local literary ecosystem.

Using the thematic data from the LDA model, an experiment was designed to assess the degree to which author nationality, author gender, and date of publication could be predicted by the thematic signals expressed in the books. In a series of tests, a classifier was repeatedly trained on data from a random sample of two-thirds of the novels in the corpus and then tested on the remaining one-third.‡ When it came to detecting, or classifying, nationality, the results were impressive. Using only this thematic data, the classifier was able to differentiate among British, Irish, and American texts with an average accuracy of 67 percent.§ In cases where the model erred, the error often involved a text by

* http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/.

† At the time of this writing, two projects of the Stanford Literary Lab are attacking the matter of novelistic "affect," or "sentiment," and the matter of character. The former project employs topic modeling and opinion-mining software to extract dominant shifts in emotional language. The second project is using Named Entity Recognition and Social Network Analysis to track character interactions. See http://litlab.stanford.edu.

‡ As in previous chapters, I have employed the NSC classifier.

§ The experiment was constructed such that in each run of the classifier, an equal number of samples were selected for training and testing. In this way, the probability of a chance assignment could be held constant at 33.3 percent. The observed F-measure, or "harmonic mean of precision and recall," of 67 was significantly better than what could be expected by mere chance.

a writer whom we may consider “cross-national.” The cases in point are James McHenry and Mayne Reid, two Irish authors who traveled in and wrote almost exclusively about the United States. When Scottish writers were added to the corpus, the overall model accuracy dropped to 52 percent. That said, this was still a big improvement beyond the expected accuracy of 25 percent. Given the close geographic proximity between Ireland, England, and Scotland as well as the long history of cross-cultural pollination, I was not surprised to find that English, Irish, and Scottish authors were hardest for the classifier to separate. When only English, Scottish, and Irish texts were included, a mean accuracy of 60 percent was observed across one hundred runs of the model—not a bad result given the three-class chance probability of 33 percent. Nevertheless, it was the Americans who proved to be the easiest to distinguish from the rest. For each possible pairing (that is, English and Irish, English and American, English and Scottish, and so on), I performed a series of one hundred independent two-class training and testing experiments. In each series, the mean performance was recorded. In these tests, we may assume that when the classifier's overall accuracy is higher, the two national literatures being tested are more distinct from each other. Put another way, when classification accuracy is high, the classes are most dissimilar; when classification accuracy is low, the machine is having a harder time distinguishing between the two literatures because they are more similar. Sequentially testing each pair provides an estimate of the thematic similarity of the three nations. The best accuracy, an 85 percent f-measure, was achieved when classifying American and Irish books. The next best result was seen in Scottish-versus-American authors, with an f-measure of 83. The American-versus-English test and an f-score of 81 followed this. The most difficult task was distinguishing English from Irish authors. Here the model achieved an f-score of only 62 (an f-score of 66 was observed when classing English and Scottish novels).

More interesting than these classification scores, however, are the data the classifier returns about which features, which themes in this case, were most useful in separating the national literatures. When all four nations are included, the most useful themes are the ones involving dialectical terms. Not too surprisingly, these linguistically potent themes are strong indicators of Irish, American, and Scottish authorship, a fact that returns us in some interesting ways to Yeats's notion of Irish accents and to Fanning's ideas about Irish authors engaging in a type of “linguistic subversion.” The data from the classifier provide us a way of exploring exactly which thematic ingredients, in which proportions, best define the three national literatures. They give us a way of backward engineering the thematic recipes favored by the authors of a given nation, and from this we may also readily identify the outliers, those writers and books that are most atypical of the general tendencies. Figures 8.18, 8.19, 8.20, and 8.21 show the twenty-five themes that the classifier identified as being the most positively associated with

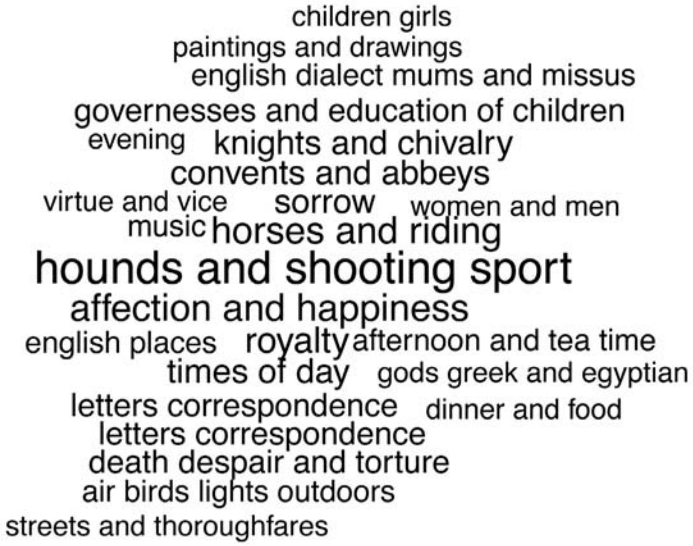


Figure 8.18. Dominant British themes

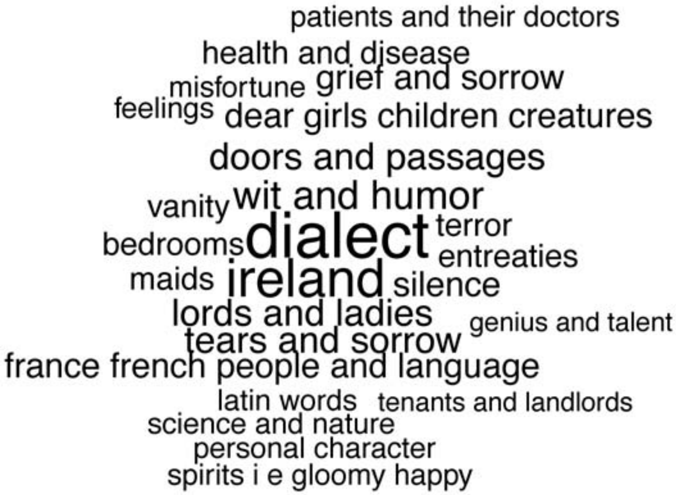


Figure 8.19. Dominant Irish themes

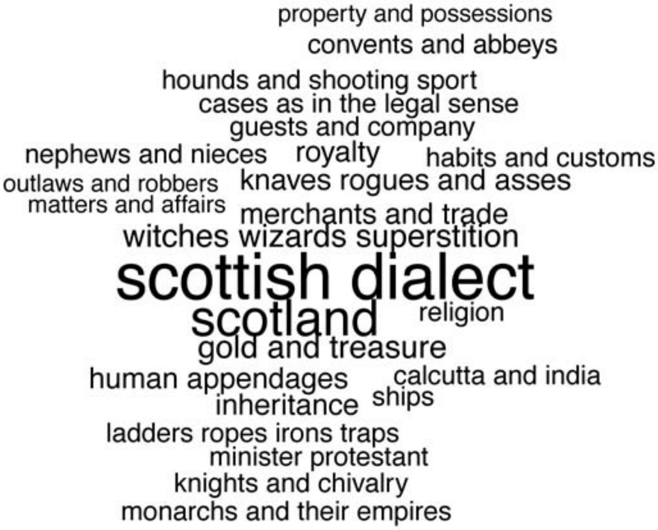


Figure 8.20. Dominant Scottish themes

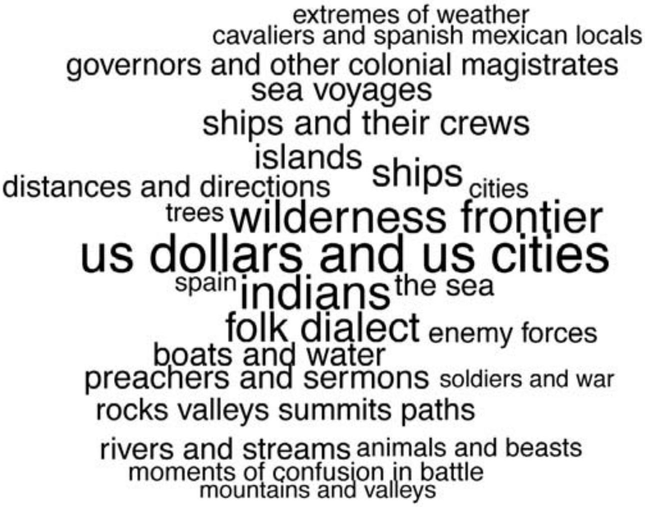


Figure 8.21. Dominant American themes

each of the four national classes. In other words, it was the comparatively high presence of these themes in each nation that the machine found most useful in distinguishing the classes. The importance of the themes is registered by their size in the word clouds.

The American corpus is typified by a set of themes largely about the natural world: "Wilderness," "The Frontier," "The Sea," "Native Americans," "Trees," and so forth. This cluster of dominant themes is very much in keeping with Longfellow's call for a uniquely national literature. In his novel *Kavanagh* (1849), he writes of a national literature that would be "commensurate with our mountains and rivers . . . a national epic that shall correspond to the size of the country. . . . We want a national drama in which scope shall be given to our gigantic ideas and to the unparalleled activity of our people. . . . In a word, we want a national literature altogether shaggy and unshorn, that shall shake the earth, like a herd of buffaloes thundering over the prairies." English authors, on the other hand, are unified by themes related to the aristocracy and upper classes: for example, "Royalty," "Music," "Private Education," "Running of Hounds," "Shooting Sports," and "Art." There are very few among the top English themes that we could consider negative: one theme of sorrow and another I have labeled "Death, Despair, and Torture." Contrast this, however, with the themes that characterize the Irish corpus. Along with the more positive "Wit and Humor," we find an abundance of darker themes such as "Doubt and Fear," "Misfortune," "Health and Disease," "Terror," "Silence," "Tears and Sorrow," and "Grief."

The Scottish corpus shares a lot with the English (as we might expect), such as "Monarchs," "Royalty," and "Running of Hounds," but also in contrast to the English, we see more emphasis on the theme "Religion" (especially "Protestantism") and the somewhat odd presence of "Witches, Wizards, and Superstition." As noted previously, we may also learn much from the absence of theme. English authors are silent on the matter of "Tenants and Landlords"; they avoid discussion of "Ireland." American authors are disinclined to discuss the themes of "Lords and Ladies," "Servants," "Housekeepers," and "Maids." Surprising to this researcher, Irish authors were found to be far less interested in natural beauty than either the English or the Americans. Compared to the other nationalities, Irish authors avoid prolonged engagement with the themes "Rivers and Streams" and "Mountains and Valleys"; there is no idyllic "Lake Isle of Innisfree" in this nineteenth-century Irish corpus.

The national data confirm some intuitions and challenge others. Irish authors write more about Native Americans and slavery than we might have imagined, and the American authors—who Hemingway once claimed were intent on imitating the English—proved to be far from "English" in terms of both their style and their theme. Much more dramatic, however, are the data related to gender. The gender data from this corpus are a ringing confirmation of virtually

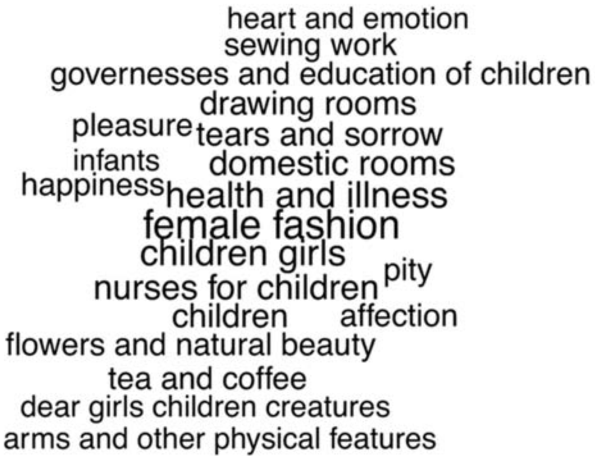


Figure 8.22. Female themes

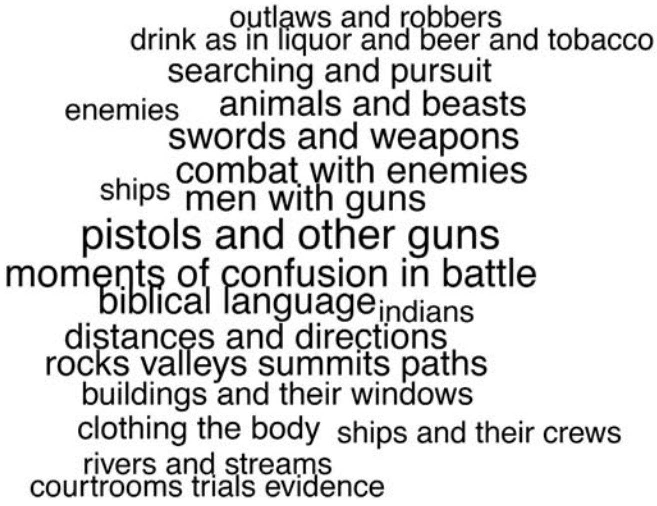


Figure 8.23. Male themes

all of our stereotypes about gender. Smack at the top of the list of themes most indicative of female authorship is “Female Fashion.” “Fashion” is followed by “Children,” “Flowers,” “Sewing,” and a series of themes associated with strong emotions (see figure 8.22). In contrast stand the male authors with their weapons and war. Topping the list of characteristic themes for men is “Pistols,” followed in turn by “Guns,” “Swords,” “Weapons,” “Combat,” and a series of themes related to the rugged masculine places where such implements of war are most likely to be employed: battlefields, mountains, and so on (see figure 8.23). To be sure, these are not the only things that men and women write about, but so striking are these differences that the classifier achieves 86 percent accuracy when guessing the gender of an author using these thematic data. This 86 percent is a full 6 percent better than what was observed with the stylistic data.

When considering these findings, it is important to remain mindful of the exceptions. These are macro trends we have been exploring, and they provide a generalized view of the whole. These are not necessarily the tendencies of individual authors or even of individual books. Indeed, 33 percent of the books in the nationality experiment and 14 percent of the books in this gender experiment were misclassified: some Irish authors were thought to be American, some male authors to be females, and so on. There are the outliers and the exceptions, and these are the subject of the next chapter.