CHAPTER 2

A DESCRIPTION OF THE STUDIED TEXT CORPORA AND A DISCUSSION OF OUR MODELING STRATEGY

2.1 Introduction to the Corpora: Selecting the Texts

For our study, we selected two particularly large corpora that document significant legal and political discursive processes in American history. These corpora were selected because they hold the promise to yield multiple avenues for research. These corpora had never before been fully digitized. We expected to find treasures within these troves merely by digitizing them for search possibilities. The extremely laborious process of digitization and then cleaning, described in more detail in the next chapter, would pay off if the effort could be expected to yield several more avenues to explore in future research projects. Yet these texts have the potential to yield even greater insights when subjected to systematic text analysis.

In matters of legal and constitutional research, the historical records are important resources. Neither of these texts had been digitized before, so there was much that could be learned from them. In constitutional matters in particular, issues debated while the terms are being drafted and enacted have particular significance. We chose two corpora that followed 1789, the constitution's birth, but influenced the path by which state governments were created, and when the Civil War required reconstruction of the Constitution, we turned to look at the Congressional debates during Reconstruction. Given the form in which these debates had been printed, little had been done on the immense number of speeches and debates taking place in Congress. These were indexed, snippets had been collected, but we chose to look at the entire corpora of debates for the decade of Reconstruction.

There is an additional jurisprudential reason for focusing attention on the Reconstruction debates due to a concept called "original intent." Courts charged with interpreting the language of constitutions, or of statutes, routinely consult the legislature's discussions taking place at the time that the new law was enacted. There are entire schools of thought about how these legislative statements should be interpreted and weighed by courts, but lawyers and judges agree on one premise: They are relevant. Thus, by digitizing the speeches of the members of Congress for the critical years that they discussed enacting three major constitutional amendments, we have assembled relevant information that had previously been hidden in the thousands of pages of a dozen volumes of the *Congressional Globe*.

2.2 Debates of the 39th U.S. Congress, as recorded in the *Congressional Globe*

Although we have digitized the proceedings of the U.S. Congress from 1864 to 1872, for this book's focus on method, we apply text mining techniques only to the speeches given during one 2-year Congress, the 39th U.S. Congress that met from 1865 to 1867. Everything spoken on the floor of either the Senate or the House of Representatives, while those entities were in session, was ultimately printed in volumes called the *Congressional Globe. The Globe* was the precursor to the modern *Congressional Record*.

These books are unique because they are both so detailed and so comprehensive. These volumes are the official account of exactly what was said and done on the floor of Congress, published for the reading of the American people and for posterity. Not only were 2-hour long speeches transcribed and published in their entirety, so too were brief outbursts and interruptions from other congressmen on the floor, outburst and applause from the galleries, and even laughter. Thus, this corpus of the 39th Congress contains about 100,000 spoken statements. A statement can be an entire speech, an uninterrupted 2-hour oratory given by a particularly long-winded congressman, or it can be the attempt of one congressman to interrupt another or to take the floor. We refer to all of these verbal statements as speeches, and the speeches in this collection have extremely variable lengths. What characterized each one is that it is the sustained expression of a single congressman.

The 39th Congress is particularly interesting because of the monumental events occurring at that time, events that were sometimes transpiring and sometimes discussed on the floor of the two houses of Congress. The 39th Congress met from March 4, 1865, to March 4, 1867. It convened very soon after the 38th Congress had finally concluded after passing the Thirteenth Amendment to the Constitution abolishing slavery, after the South had surrendered, and after President Lincoln had been assassinated and the unpopular vice president Andrew Johnson was sworn in as President. During the 39th Congress, the Thirteenth Amendment was sent to the states for ratification, and Congress continued to turn its attention to further civil rights reforms, such as birthright citizenship and an expanded Federal role in ensuring due process and equal protection for all persons in the United States. The Congress confronted both the problems of securing true freedom for the freedmen and finding the terms on which to reunify the defeated South into the union from which it had attempted to secede.

We started from the printed record of the speeches. Each page was subjected to optical character recognition (OCR) and converted into an

electronic record and stored in Microsoft Word format. We combined pages into a giant master file, stripped off all book page breaks, headers, footers, and page numbers, and entered separation symbols to indicate the start of each speech. We replaced as many typos and OCR errors as possible. Many of these changes had to be made manually using text replacement code. The initial two words of each speech allowed us to infer the speaker. However, inconsistent spellings of the names of the speakers and inconsistencies in the placement of the speaker's name within the speech made it necessary to hand edit the speaker information, which also took considerable time. Such text cleaning is certainly time-consuming. We spent months performing these cleaning operations. Whereas more cleaning could certainly have been done, we tried to do our best. Working on such large projects, one quickly realizes that the "perfect text document" with zero errors is an elusive ideal.

Digitization has given us access to these speeches in a way that has never been available before. From our digitized corpora, we can retrieve speeches that may not have been indexed in one of the few subject matter indices to the volumes. From our analysis, we seek insights into the lawmakers' discussions about important issues of public concern: race, slavery, equality, and justice. We are also interested in the dynamics of the discourse. How did the discussions of slaves, freedmen, and African American citizens change over time? Who were the movers and shakers among this mix of congressmen? Who dominated the Reconstruction debates, with which issues, and which speakers taking the lead, and which others echoing the words of the leader? In this sometimes heated and excited exchange, who addressed whom directly? All of these questions can be addressed by sustained and systematic methods of text analysis.

From analyzing the occurrence of words such as "white", "black", "negro", "labor", "slaves", and so on, we can learn about the predominant views of congressmen on these subjects. We can identify the connections they made linguistically and theoretically between reforms on civil rights and other reforms such as bankruptcy and taxes, for example. During the 39th Congress, the Thirteenth Amendment was in the process of ratification by the states. The Congress was predominately Republican and in increasing tensions with the Democratic Andrew Johnson who had just ascended to the presidency. The 39th Congress passed much more legislation including the Civil Rights Act of 1866, a statute creating the Freedmen's Bureau, which would attempt to police the conditions of freed slaves in the South. We can also learn more about the political and rhetorical dynamics of different congressmen. We can also learn more about how specific words and phrases change over the course of discussion. And we can identify certain words used during discussion that become dedicated—that is, exclusively applied to specific topics.

2.3 *The Territorial Papers of the United States*

The second source of texts that we selected to demonstrate text mining techniques were *The Territorial Papers of the United States*. These several volumes, commissioned by Congress contain the most comprehensive collection of letters between the antebellum western frontier and the U.S. government that has ever been assembled and published. The letters, written between 1789 and 1848, detail the communications between the national capital and residents of the territories. During these years, residents of distant territories under American sovereignty communicated with their government by longhand correspondence. During the 20th century, Congress authorized that these documents, which pertained to frontier territories becoming states, be assembled, curated, transcribed, and ultimately published. The effort took more than 20 years and eventually produced a series of 28 volumes of collected documents under the title *The Territorial Papers of the United States*. The serial set goes through Wisconsin's entrance into the union. After this volume was completed, the project of publishing more volumes was abandoned. Still, this comprehensive collection of communications between officials in the nation's capital and the territorial residents describes the actual discourse of territorial governance as the American empire expanded, adding new lands and new state governments, which is why Congress saw fit to commission the collection.

All 28 volumes were published in the 20th century, so they have the advantage for text analysis in that the typeface is standard and uniform. Digitizing offers several advantages, because so far, these documents have only been available in hard copy form. Each volume isolated a different territory and each of the 28 volumes was indexed separately. The project to digitally analyze *The Territorial Papers of the United States* is important for several reasons.

First, these texts contribute to our understanding of American identity at a critical period of the nation's history as a settler colonial state. This period is revered and mythologized when the American people contemplate the origins of the nation. Since this collection is extremely large, digital methods are the only ways to cull the letters for common themes. In fact, digital methods offer the prospect that one can discover the relative order of significance of these themes. With digital methods, one can also discover dynamic changes occurring simultaneously or successively in different regions of the nation's borderlands. This analysis should yield answers to the interesting and classic questions: How do borderlands become integrated into the nation-state? As the states were formed, what did the American people want from their government and conversely, what

did the government want from its people in the process of bringing them into the nation-state? Stated another way, how did the American people engage the process as they constituted new state governments for themselves? Since these texts concern expansion of a major empire, these results contribute to the general understanding of empires in world history and settler colonialism in particular.

Second, this cumulative correspondence spanning six decades provides evidence about gradual changes in legal identity as the nation building takes place. These letters occur in the period of national expansion between the nation's initial constitution and its significant reformation by successive amendments during Reconstruction after the Civil War. The Congresses' Reconstruction Amendments—the Thirteenth, Fourteenth, and Fifteenth Amendments—were greatly influenced by the process of national expansion that had reconfigured the nation.

Many describe the territorial ordinances like the Northwest Ordinance as a proto-constitutional document (Duffey, 1995; Lawson & Seidman, 2004). As the major law governing the territories and shaping their entrance into the membership as states, the Ordinance provided both substantive and procedural legal rules. The Ordinance outlined the general path by which regions could petition to be recognized as states and enter the union on an equal basis with other states. Although the Northwest Ordinance was the operative law, it contained aspirational and open-textured language to be filled in by the nation in the course of settlement (Frymer, 2017). The Ordinance's meanings were subject to negotiation between the government and those residing in the territories. As frontier residents petitioned for self-governance, respect from the national government, and inclusion in the national populace, the terms of inclusion and treatment were continually being negotiated in the discourse contained in these documents. Many of these concerns gave rise to the terms *citizenship*, *equal protection*, and *privileges and immunities* in the Fourteenth Amendment.

Third, these texts represent evidence of an important narrative for American identity. Stories of origin are always instrumental in creating and reinventing the narrative of national identity. Frontier stories occupy a significant place in American imagination. With the opportunity to analyze such a large corpus of papers meticulously curated by trained historians, one can revisit some conventional beliefs about how settlement took place. By disaggregating the place of origin of the communications, one can reveal the relative differences in significance of various themes for the settlement people as compared with their thematic significance for the national leaders. Moreover, where certain topics are conspicuously absent from the corpus, it may indicate that

they were not significant subjects of interest in the communication between the territories and the national capital about the national expansion project.

The volumes of the *Territorial Papers* are suitably susceptible to textual analysis techniques. These documents were curated and transcribed by experts in the field. Again, because they were printed in the 20th century, these documents are suitable for OCR, making the OCR product more reliable.

However, there are also numerous difficulties. One of the greatest obstacles to textual analysis of the *Territorial Papers* is spelling irregularities. Not only did spellings of common words change over the six decades, many of the correspondents, particularly frontier residents, were only partially literate, and hence, they improvised spellings. We have made no attempt to standardize these spellings in the corpus, because even these slight changes in spellings may prove an interesting subject for later inquiries about the basic literacy of the correspondents.

Furthermore, even with standardized print, OCR results were far from perfect and needed considerable additional cleaning. To produce our corpus, we followed a laborious iterative "data cleaning" process that adhered to principles of continuous quality improvement. For the past several years, our research assistants produced scans of the *Territorial Papers* and cleaned the basic documents. The cleaning included eliminating the modern editorial commentary of footnotes, headers, and footers. This work was followed by coding and still more preprocessing: stripping the metadata from each letter by marking the letter's author, the addressee who was the intended audience, and date. Although some of this work could be automated by programming, a considerable amount of subsequent cleaning was necessary. Producing a digitized version of the *Territorial Papers* has been a labor-intensive exercise.

For certain kinds of inquiry, such as keyword searches, the poor quality of OCR data can be quite troublesome. Topic modeling (a technique that we describe later in Chapter 9), with its goal of analyzing a large corpus of text for key language patterns and word clusters, on the other hand, may be less sensitive to poor text quality. Since topic modeling analysis looks for general patterns across thousands of pages rather than semantic accuracy in a handful of excerpts, the poor quality of the OCR in such a large corpus may not be as problematic. For discussion, see Young (2013) and Walker et al. (2010). More on that in Chapter 9.

Our digitized material contains 15 volumes and 8,500 entries, starting with the second volume covering the Northwest Territories with letters from 1789, and ending in 1848 with Volume 28 on Wisconsin, which was

the last volume Congress saw fit to subsidize and publish. This chronology and selection of region conveniently reflects the formation of these new northern and western states.

In this book, we analyze Volumes 2 and 3 (Northwest Territories); Volumes 7 and 8 (Indiana Territory); Volume 9 (New Orleans); Volumes 13, 14, and 15 (Louisiana and Missouri); Volumes 16 and 17 (Illinois); Volumes 10, 11, and 12 (Michigan); and Volumes 27 and 28 (Wisconsin). We refer to them as (sub) corpora: Corpus 1 (Northwest Territories) through Corpus 7 (Wisconsin), respectively.

Meta (covariate) data for each letter is extracted from the text document. The time when the letter was written is our most important metavariable, and generally, it could be extracted from the letters automatically, using appropriate computer code. Other important metavariables are (sub) corpus group (Corpora 1 through 7, spanning the Northwest Territories through Wisconsin), and the author and addressee of the letter. Unfortunately, the formatting of the letters is irregular, which made it very time-consuming to extract author and recipient. Great care was used to standardize the spellings of several thousand correspondents' names (e.g., there were numerous spellings of Governor Claiborne, often misspelled, with and without his first name and with and without his title).

In addition to letters, the *Territorial Papers* contain other documents significant to the historical process but styled differently than author/addressee. Nonetheless, these documents emanated from an entity, such as Congress, the Secretary of War, or the President, and had an intended audience, such as the people of a particular region or the nation at large. These documents include proclamations, memorials, petitions, militia orders, testimonials, editorials in newspapers, statutes, laws, and official reports. To code these, we developed a protocol to identify the authorial entities and their intended audiences. Statutes enacted by Congress are thus styled "From Congress to the World", since a statute is a communication at large. Similarly, executive proclamations styled "to all who draw near" are coded as "From Governor to the World".

2.4 Analyzing Text Data: Bottom-Up or Top-Down Analysis

A successful text analysis requires expertise from many areas. It requires a sound knowledge of the research question (content understanding), of the data and their limitations (data understanding, including data preparation and the continuous improvement of the text), of the modeling process (thorough understanding of the analytic/statistical models and

their assumptions, and familiarity with the software to implement the models), and of results interpretation.

A bottom-up approach starts with the data (in our case, the text) rather than from theoretically based hypotheses, and it mines the textual information for data insights that can help develop more narrowly focused theories to be tested subsequently on the available data. Any successful analysis of information (text or numeric data) relies on a continual iterative learning process. One can start from the data as we do here and use an exploratory data analysis (see Tukey, 1977) or bottom-up approach to gain information on patterns and relationships and to develop hypotheses; of course, any findings of the exploratory analysis needs to be incorporated into more formal models at the subsequent modeling stage. Or, one can start with a targeted analysis by fitting models that formalize those patterns, investigate whether such models actually fit the data at hand, and improve the models if there is lack of fit. At the beginning of an investigation where there is little theory about the discourse, the former approach with its wide lens is preferable. On the other hand, when theory has already been developed, one can adopt a tighter lens and start fitting more narrowly focused hypotheses to the data. It does *not* matter where one starts (from the data or from the model), as long as one follows a process that iterates between the data and the theory. George Box et al. (2005), in Chapter 1 of their book *Statistics for Experimenters*, characterize this very eloquently in reference to (statistical) learning as a feedback loop between data (facts) and conjectures (models).

Text mining can do more than confirm or refute the validity of narrowly focused theory-driven word occurrences. Text mining allows you to get to know the text more thoroughly. Rather than being a "one-step" hypothesis testing activity, text mining is a continual learning process. Both the data preparation and the modeling involve iterative, continual tasks. The preparation of the textual information involves multiple, iterative data cleaning steps—all very time-consuming and often requiring considerably more effort than the actual analysis and the modeling of the data. And the textual analysis iterates between exploratory analysis and thoughtful model construction. It would be a mistake to rule out a bottom-up approach as just a fishing expedition; much can be learned from the method.

Opportunities for text mining applications are growing rapidly. Gentzkow et al. (2017) discuss a wide variety of successful text mining applications, giving examples such as authorship problems, stock price predictions, media slant, and assessment of policy uncertainty. Text is inherently high dimensional, and this high dimensionality poses many

challenges for the statistical analysis of the textual information. Their paper discusses practical difficulties for textual analysis:

- Users not having easy access to clean raw text and struggling with data cleaning procedures
- Users struggling with the appropriate feature selection (words or *n*-grams)
- Stemming issues and questions about the appropriate unit for analysis (documents/letters or strings of consecutive words)

Their paper also discusses challenges for the statistical models such as convergence issues in the estimation of unobserved components models (e.g., the topic models considered in Chapter 9 of this book) and the efficient estimation of supervised classification models that relate the textual information to covariate (meta) variables. We touch on all these issues in later chapters.

2.5 References

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters* (2nd ed.). Wiley.

Duffey, D. P. (1995). The Northwest Ordinance as a constitutional document. *Columbia Law Review*, *95*, 929–968. https://doi.org/10.2307/1123211

Frymer, P. (2017). *Building an American empire: The era of territorial and political expansion*. Princeton University Press. https://doi.org/10.2307/j.ctt1vxm7rr

Gentzkow, M., Kelly, P. T., & Taddy, M. (2017). *Text as data*. https://web.stanford.edu/~gentzkow/research/text-as-data.pdf

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2018). *Congressional record for the 43rd-114th Congresses: Parsed speeches and phrase counts*. Stanford Libraries [distributor], 2018-01-16. https://data.stanford.edu/congress_text

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019, July). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica, 87*, 1307–1340. https://doi.org/10.3982/ECTA16566

Lawson, G., & Seidman, G. (2004). *The Constitution of empire: Territorial expansion and American legal history*. Yale University Press. https://doi.org/10.12987/yale/9780300102314.001.0001

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Walker, D. D., Lund, W. B., & Ringger, E. K. (2010). Evaluating models of latent document semantics in the presence of OCR errors. In *Proceedings of the 2010 Conference on Empirical Methods of Natural Language Processing* (pp. 240–250). Cambridge, MA.

Young, D. T. (2013). How do you measure a constitutional moment? Using algorithmic topic modeling to evaluate Bruce Ackerman's theory of constitutional change. *Yale Law Journal*, *122*(7), 1990–2054.

Appendix to Chapter 2: The Complete Congressional Record

Gentzkow et al. (2018) have made available *all* speeches in the U.S. Congress from 1873 to 2017 (43rd to 114th Congress) as transcribed in the *U.S. Congressional Record*. They obtained the digital text from HeinOnline who performed OCR on scanned print volumes. The digitized text of speeches as well as the metainformation for the speeches (name of the speaker, together with his or her chamber/state/party affiliation, and the date of the speech) for each of the 72 consecutive sessions of Congress are available on their Stanford University website.

Gentzkow et al. (2019) use the data to measure trends in the partisanship of Congressional speech from 1873 to 2017. They define partisanship as the ease with which an observer can infer a congressperson's party from his or her text utterances. Their findings are published in the July 2019 issue of *Econometrica*.

More details on their data and how to process it in R are shown on the website for Chapter 2 at https://www.biz.uiowa.edu/faculty/jledolter/analyzing-textual-information/.