And the Rest is History: Measuring the Scope and Recall of Wikipedia's Coverage of Three Women's Movement Subgroups

Laura K. Nelson
Rebekah Getman
Syed Arefinul Haque

Abstract

Narrating history is perpetually contested, shaping and reshaping how nations and people understand both their pasts and the current moment. Measuring and evaluating the scope of histories is methodologically challenging. In this paper we provide a general approach and a specific method to measure historical recall. Operationalizing historical information as one or more word phrases, we use the phrase-mining RAKE algorithm on a collection of primary historical documents to extract first-person historical evidence, and then measure recall via phrases present on contemporary Wikipedia, taken to represent a publicly-accessible summary of existing knowledge on virtually any historical topic. We demonstrate this method using women's movements in the United States as a case study of a debated historical field. We found that issues important to working-class elements of the movement were less likely to be covered on Wikipedia compared to other subsections of the movement. Combining this method with a qualitative analysis of select articles, we identified a typology of mechanisms leading to historical omissions: paucity, restrictive paradigms, and categorical narrowness. Our approach, we conclude, can be used to both evaluate the recall of a body of history and to actively intervene in enlarging the scope of our histories and historical knowledge.

Keywords

sociology of history, computational text analysis, phrase mining, women's movements, Wikipedia

Word Count*:* 12,930 (inclusive)

Acknowledgements: A version of this paper was presented at the Culture and Politics Workshop at the University of North Carolina; the authors extend deep gratitude to the participants for their helpful comments, particularly extended comments from Andrew Perrin and Neal Caren. We also thank Zobayer Hasan for his helpful suggestions regarding the processing and searching of large-scale text using Elasticsearch.

Author Bios

Laura K. Nelson is an assistant professor of sociology at the University of British Columbia. She studies culture, gender, social movements, and institutions, primarily using computational methods. She has published in the *American Journal of Sociology*, *Sociological Methods & Research*, *Gender & Society*, *Mobilization: An International Quarterly*, and *Poetics*, among other outlets.

Rebekah Getman is a Ph.D. candidate in the sociology department at Northeastern University. She studies health, institutions, gender, and social movements using qualitative and computational methods, and is currently studying information flow about childbirth between patients and experts during COVID. She has published in the *American Journal of Public Health* and *Health, Education, and Behavior*.

Syed Arefinul Haque is a Ph.D. candidate in the Network Science Institute at Northeastern University. He studies the diffusion of ideas and the sociology of knowledge. Currently, he is using network science and natural language processing techniques to quantify gender and racial diversity in researchers and experts, and how ideas related to gender diversity move from one organization to another. He has recently published articles in journals such as *Nature Communications* and *Clinical Microbiology and Infection*.

Writing history is the process of transforming a series of facts into historical patterns, including sequences, processes, and narratives (McNeill 1986). In order to identify interpretable patterns, scholars must decide what information is important enough to include in a historical account. While any individual historical account must omit more historical information than it includes, the general assumption is that, as the collective scholarship on a topic accumulates, historians will iteratively assemble all relevant information about that topic into interpretable narratives and patterns. Historically, of course, this has not been the case. Women's history, working class history, ethnic studies, and other subdisciplines emerged precisely because information that gets included in a body of knowledge is imbued with power dynamics and intentional and unintentional biases (Dill 1979; Foner 2003; Lerner 1975; Schwartz 2003). While scholars have persuasively demonstrated systematic biases in historical work, particularly conspicuous absences (see also Dunbar-Ortiz 2014; Zinn 2015), it has thus far been difficult to measure and confirm precise patterns of omissions in historical work at scale.

Measuring the *recall* of a body of history—what information is included in a collection of histories among all available historical information—requires two methodological advances. First, we need a method to measure historical ground truth, or a range of first-person, on-the-ground empirical evidence about a historical event or period that could be included in secondary histories. This in turn requires a way to empirically operationalize historically-relevant information from this primary evidence. Second, we need a method and data source to measure whether or not that information appears in a body of historical work. In this paper we leverage Wikipedia as the largest and most accessible public summary of verifiable knowledge to propose a general approach and a specific method to measure the recall of historical knowledge. To demonstrate our approach we present a targeted case study comparing what information from a slice of the U.S. women's movement—a diverse and contested topic in history—is included in contemporary Wikipedia, and what information is omitted.

We use the term *ground truth* to simply mean information provided by direct observation rather than inference. While it is impossible to establish the full and complete "truth" of an historical event or topic, it is possible to assemble direct, first-person empirical evidence of an event or topic. Following the practice of historians, we constructed first-person empirical evidence via primary archival material, assuming that while archives can never represent the entirety of the truth about an event or topic, they do present a curated slice of on-the-ground empirical information. To construct primary empirical evidence of the women's movement, we used a small but comprehensive collection of primary source documents from three diverse subsections of this movement, compiled by the editors of the digital library *Women and Social Movements in the United States, 1600-2000*: (1) the *Equal Rights Journal* representing the National Woman's Party, a professional, mainstream women's movement organization; (2) meeting notes and minutes from the National Consumers' League, an example of a working-class women's organization; and (3) a collection of writings by Black women suffragists, representing the intersection of race and gender.

We operationalized information as one or more word phrases, and we used the phrase mining algorithm RAKE (Rapid Automatic Keyword Extraction) to extract key phrases from our women's movement subcorpus. Using Elasticsearch, an open-source database allowing for efficient search across large amounts of data, we then identified which of these phrases were present in a recent dump of English-language Wikipedia, comparing the presence and absence of phrases across the three movement subsections. We followed this quantitative analysis with a targeted qualitative analysis of frequently used phrases from the women's movement data that were omitted from select Wikipedia articles, to both validate our approach and to better understand patterns in historical omissions.

We found that, while virtually all extracted phrases (~95%) appeared somewhere on Wikipedia, only 50-60% of the phrases more distinctive to the women's movement (operationalized as phrases with three or more words) were present, and there was much less coverage of the working-class subsection of this movement compared to the other two subsections. Analyzing these missing phrases across all three subsections in more detail, we developed a typology of mechanisms of factual omissions. First, in some cases there was simply a *paucity* of details. This was the case, we found, with the National Consumers' League and working-class women in general. Second, *restrictive paradigms* can lead to the association of an organization or phenomenon with particular exemplar issues, producing a myopic focus not supported by the primary material. This was the case, we found, with the National Woman's Party, which has become too tightly coupled with suffrage and the Equal Rights Amendment in its collective history. Third, *categorical narrowness* arises when a certain category, such as feminism, is defined so narrowly that many issues important to historical actors never make it into historical accounts. We found this was the case with the writings of Black women suffragists, where issues around Jim Crow Cars, domestic work, and health and sanitation have been decoupled from historical accounts of Black women's activism, even as these issues were core to their work.

Methodologically, we show how phrase mining, long popular in natural language processing (NLP) pipelines, can be leveraged for social science and historical research. Many popular text analysis techniques, such as topic modeling and word embeddings, are designed to operationalize themes or relationships between words and phrases in a corpus, but do not allow scholars to measure the development and diffusion of specific ideas or concepts within and across texts. Our approach demonstrates how NLP tools such as phrase mining and fuzzy string matching can be used to identify and track discrete ideas and concepts in collections of texts.

Measuring specific ideas rather than themes or semantic relationships can be particularly useful when analyzing domains such as social movements, where framing is crucial and impact often comes via introducing carefully crafted phrases (e.g. Black is beautiful), concepts (e.g., sexual harassment), and ideas (e.g., women's rights) into mainstream discourse.[1] Applied to the domain of public history, our method and approach can be used to both evaluate the recall of a body of history and to actively intervene in enlarging the scope of our histories, with implications for historians, historical sociologists, and the sociology of knowledge more broadly. While narrating history includes much more than merely selecting facts, the selection of information is a crucial first step to historical interpretation.

BACKGROUND AND OVERVIEW

[1] Science and entrepreneurship are two other popular domains for phrase mining applications. In these domains impact is also often measured by the growth and spread of ideas or concepts.

History as a Site of Struggle

The “1619 Project” by journalist Nikole Hanna-Jones was published in *The New York Times Magazine* in 2019 in order to reframe “the country’s history by placing the consequences of slavery and the contributions of black Americans at the very center of our national narrative” (Silverstein 2019). Reconfirming the power of historical narratives, this publication sparked a backlash that reached the highest office in the United States, with former president Donald Trump threatening to investigate public schools teaching the essays from the project (Liptak 2020) and, a year later, multiple states introducing bills to cut funding for schools providing lessons derived from the essays (O’Kane 2021).

Textbooks are another site of public and political contention about history, as state legislatures debate and determine what should be included in textbooks for public schools (see, e.g., Moreau 2004). In January 2020 *The New York Times* investigated the impact of these decisions, highlighting differences between versions of the same textbooks used in public schools in California and Texas. The California textbook, for example, annotated the Second Amendment to indicate that it has allowed for some gun regulations, while in the section on the Harlem Renaissance, the Texas textbook includes a sentence indicating that some have “dismissed the quality of literature produced” (Goldstein 2020). *The New York Times* study suggests that these small differences in what is included and how can substantially alter the way we understand and convey history (see also Polletta et al. 2011; Schwartz 2003).

Most debates about history, of course, do not become national discussions or spur changes in public policy. Yet even those confined to the halls of academia are consequential.

Academic debates about history include not just the accuracy of information included in published histories, but also historical omissions. Women's history, for example, developed as a concerted movement in the United States in the 1960s and 1970s to counter the absence of information about women in historical scholarship (e.g., Davis 1976; Dill 1979; Riley 1988; Scott 1986). Ethnic studies and working class history movements similarly focused historical lenses on populations traditionally left out of historical scholarship to counter perceived biases (Dunbar-Ortiz 2014; Zinn 2015).

In short, those across the political spectrum, and activists and historians alike, recognize the significance of the information we include and omit when we tell history. The intellectual and political importance of these questions, and the often intractable debates about how history is told, reveals an underlying challenge: is there a way to determine all information that could be included in our collective history, in order to identify what is systematically missing from historical work? In other words, if all histories are by necessity partial, can we better identify the precise contours of their partiality? As more knowledge is digitized and made available, and as knowledge is increasingly summarized in various digital formats and databases, we have a new opportunity to contemplate the possibility of measuring the recall of historical scholarship. One resource in particular has brought us closer to this potential: Wikipedia.

Wikipedia

At its most basic, Wikipedia is a digital encyclopedia with the lofty goal of imagining a "world in which every single person on the planet is given free access to the sum of all human knowledge" (Slashdot 2004). Now in its twentieth year, Wikipedia has become much more than an encyclopedia, and its importance and impact is difficult to overstate. Wikipedia is a collective: anyone can write or edit Wikipedia articles, though everyone is expected to strictly adhere to the collectively created *Manual of Style* (Wikipedia 2020b). As of the end of 2020, English Wikipedia has over 40.5 million users with a registered username and just over 130,000 of these actively and regularly edit articles. It has 1.7 billion unique visitors monthly and contains over 55 million articles in more than 300 languages, including over six million articles in English (Wikipedia 2020a). Wikipedia is more than a popular website and an example of non-hierarchical production on the internet. It is one of the main ways knowledge today is collectively constructed, and is "increasingly recognised as a global consensus view about people, places, events and things around the world" (Vrandečić and Ford 2020).

In addition to page views, Wikipedia is now thoroughly ingrained in the world's information-gathering workflow (Orlowitz 2020). Wikipedia results are often among the first results for Google searches, Google often excerpts Wikipedia in a "knowledge panel" included at the top of search results, and Apple's *Siri* and Amazon's *Alexa* use Wikipedia on both their back- and front-ends. Put simply, Wikipedia, "the virtual front page of every library" (Orlowitz 2020, 132), is the largest, most accessible, and potentially most influential collection of information we have ever seen. Because of its integration into large-scale information systems and its role as the initial source of information for those seeking to learn about a topic, we consider Wikipedia to be one of the most important sources of data for those seeking to understand the state of existing knowledge on virtually any subject.

Three of its guiding principles are important for the way we are using Wikipedia: verifiability, notability, and no original research. Anyone reading Wikipedia can *verify* that the information comes from a reliable, independent, published source with a reputation for fact-checking and accuracy (Wikipedia 2021b). While non-academic sources may be considered reliable, peer-reviewed academic sources are almost always the preferred source, particularly for topics in medicine, science, and history (ibid.). For a topic to receive its own page on Wikipedia, it must also be *notable*, meaning it has received "significant coverage" in reliable sources and the coverage has to be in enough detail that *no original research* is needed to extract content (Wikipedia 2020c). According to Wikipedia, original research includes any analysis or synthesis of published material that reaches a conclusion not directly stated by one of the reliable sources (Wikipedia 2021a).

While Wikipedia is thorough and accurate, there are still known biases in what gets included on Wikipedia and what does not. First, the majority of its articles are English-language, and its roots as an English-language project introduces cultural and ingroup biases into Wikipedia articles (Callahan and Herring 2011; Hecht and Gergle 2009; Oeberst et al. 2020). Scholars are working to discover and analyze language-specific differences in the representation of knowledge across Wikipedia (e.g., Bao et al. 2012). Gender and racialized biases have also been extensively documented, including biases in who contributes to Wikipedia (Hargittai and Shaw 2015), who gets biographical pages (Adams, Brückner, and Naslund 2019; Reagle and Rhue 2011), how the notability guideline is applied (Tripodi 2021), and how men and women are described (Graells-Garrido, Lalmas, and Menczer 2015; Wagner et al. 2016).

Because of the notability and verifiability guidelines, however, Wikipedia Executive Director Katherine Maher opined that, while Wikipedia is a work in progress and they are continually trying to improve, many biases uncovered on Wikipedia mirror biases in society, and in particular, biases in published work (Maher 2018). Regardless of where the biases originate—whether in the writing and editing process, or as a reflection of broader societal biases—Wikipedia remains the first source of information for students and others seeking information on a topic, particularly in English-speaking countries. Investigating the biases and omissions in Wikipedia helps elucidate the kinds of information readers are ingesting when they seek this initial information on a topic, as well as potential biases in the broader established knowledge of a topic.

We use Wikipedia in conjunction with primary data sources to identify what information is omitted from published (and accessible) historical knowledge. We propose an approach and method to identify and explain potential historical omissions on Wikipedia using a targeted topic as a case study: three subsections of the women's movement in the United States between 1899 and 1935.

Women's Movement History as a Site of Struggle

Two features of the U.S. women's movement make it an important case study to better understand systematic biases in historical omissions. First, as an extensively researched movement with substantial archives, there exists a wealth of both primary material, which we used to identify the range of possible information that ought to be included in a comprehensive collective history, and peer-reviewed secondary material for Wikipedia articles to cite. Second, like many movements, its narrative history is perpetually contested, primarily who is included as central actors in this movement and how this history is framed.

From its early moments, participants of the women's movement—particularly its middle- and upper-class members—have taken great care in documenting and writing its history. Early leaders of the national women's suffrage movement, such as Elizabeth Cady Stanton and Susan B. Anthony (Stanton et al. 1881), as well as local activists and clubwomen (e.g., Davis 1922) extensively documented and narrated the history of the first-wave feminist movement, providing primary material for future historians. While the history of this movement was largely ignored after 1920, one of the major campaigns of the second-wave feminist movement (~1964-1984) was to uncover, retell, and re-record the history of women's movements, including their own. Similar to the first wave, second-wave activists wrote articles and books on the history of both the first and second waves as part of their activism, again providing primary material and narratives for future historians (e.g., DuBois 1971; Evans 1980; Firestone 1968; Flexner 1970; Morgan 1970).

The history as written by second-wave activists was quickly contested, however, with Black women and other women of color in particular criticizing the excessive focus on white activists (hooks 2000; Lorde 1984). Scholars have since written alternative histories of this movement, centering the working class (Cobble 2005; Milkman 1985; Orleck 1995) and non-white groups (Cahill 2020; Giddings 2007; Jones 2020; Orleck 2015; Parker 2020; Roth 2004; Ware 2019), whose activism peaked at different times than predominantly white women's activism and focused on different issues and solutions.

The debates about the history of this movement mirror many debates happening within history as a discipline more broadly, particularly around the role and coverage of women, race, and class in history, and the intersection of these three categories in historical work. The competing histories are not just about distinct narratives of the same information; the narratives are built around foundationally different key events, issues, and solutions. Identifying the information narratives are built on is thus a first step toward measuring differences in historical interpretation.

Primary documents are the most common data historians use to construct the facts and information that form the backbone of historical narratives. Archives and museums that collect and preserve historical documents are not politically neutral (Autry 2013; Risam 2018), and there are many social processes influencing who gets to produce primary documents, and of those, what gets preserved over time (Brown and Davis-Brown 1998; Smallwood 2016). While imperfect, archived primary documents are still the main, and in many cases the only, material historians use to construct empirical evidence to support their historical narratives. We used primary sources produced by activists and participants in the women's movement to construct first-person empirical evidence, compared this against the summarized secondary information provided via Wikipedia, to measure historical recall.

DATA

Women and Social Movements in the United States

We constructed historical first-person evidence using primary documents from the digital Alexander Street Press library *Women and Social Movements in the United States, 1600-2000* (WSM). In addition to scholarly articles and interpretations, WSM contains a curated collection of approximately 120,000 pages of primary source collections pertaining to women and social movements in the United States, including publications from national and international commissions and conventions, as well as from a variety of grassroots actors. Like virtually all archival collections, the WSM library is an expert-curated collection; the documents and issues included in the collection were chosen by the WSM editors for their importance to women's movements, but also to capture diverse issues and voices from this movement. We treat this library as a small but systematic and diverse collection of primary, empirical accounts of subsections of U.S. women's movements.

Because this collection included many types of documents, from official reports to personal correspondence, and there was uneven historical coverage across the years, we chose a smaller subcollection from the WSM library that represented diverse subsections of the movement but also had significant temporal overlap. Our final primary corpus from WSM included all of the documents published between 1899 and 1935 from three primary source collections in the WSM library: *Writings of Black Women Suffragists* (WBWS)*,* the *Equal Rights (journal)* (ERJ), and the *National Consumers' League* (NCL). WBWS represents the intersection of gender and race, the ERJ represents the perspective of liberal professional women, or the middle-class mainstream of the movement, and the NCL represents the intersection of gender and class. We limited the dates to 1899-1935 for two reasons. These dates cover the peak of the first-wave (woman suffrage) movement (1900-1920) plus the following fifteen, post-first-wave and early New Deal years (1920-1935). This allowed us to include the first wave proper as well as the immediate "doldrum" years following the ratification of the 19th amendment—the years that historians have shown to be important for issues around class and perhaps race (see, e.g., Cobble 2005). Second and more practically, these dates cover the bulk of the overlapping dates within the three chosen WSM collections. We use the word subcollection to refer to our curated collection of documents from each of these larger collections and the word subcorpus to refer to each group's specific collection of documents.

Despite being excluded from large parts of the early women's movement by white women and predominantly white organizations, Black women played key roles in the first-wave movement, including the suffrage movement (Buechler 1986; Cahill 2020; Giddings 2007; 2009; Hendricks 2013; Jones 2020), most influentially through the National Association of Colored Women (NACW), founded in 1896. Through the Black women's club movement, they organized for jobs and education and for access to services such as after-school programs, health and sanitation, and old-age homes. Through interracial and Black-only suffrage organizations, they organized local and national campaigns for woman suffrage and for increased representation in local governments.

Our WSM subcollection includes 697 documents from the WBWS collection published between 1899 and 1935. The types of documents in this subcorpus include political essays on specific topics, current events, and debates, including "How Enfranchisement Stops Lynchings," "Race Prejudice and Southern Progress," and "The Humor of Teaching"; letters between activists; and reports and summaries of events and meetings. The documents in this subcorpus comprise close to 1.2 million words, with an average of 1,755 words per document (see Table 1).

The *Equal Rights* journal was a weekly magazine published by the National Woman's Party (NWP). The NWP was founded in 1916 by white suffragists Alice Paul and Lucy Burns. It was formed out of the Congressional Union, founded in 1913, as a more radical and confrontational alternative to the National American Woman Suffrage Association. Women affiliated with the Congressional Union—and later the NWP—organized the large and innovative suffrage parade in Washington D.C. in 1913 as well as the silent sentinels—the first group to picket at the steps of the White House, and one of the only protests at the White House to continue through World War I (Adams and Keene 2008; Lunardini 2000). After the ratification of the nineteenth "woman suffrage" amendment in 1920, the NWP shifted its focus to passing the Equal Rights Amendment and other issues around equal rights, including the right for women to sit on juries, equal pay, and the right to retain nationality upon marriage.

The NWP published *Equal Rights* as a weekly magazine from 1923 to 1954. It served as a resource to keep the membership informed on the status of issues and bills affecting women and championed by the NWP. Our subcollection includes all issues published between 1923 (the journal's first year) and 1935 contained in the WSM library, ranging in our subcorpus from a low of 26 documents per year (in 1929) to a high of 51 issues (in 1933). The documents in this subcorpus include over 3.4 million words, with 8,350 average words per document (see Table 1).

The NCL was founded by social reformers Jane Addams and Josephine Lowell in 1899 to focus on issues affecting working-class women. Its first general secretary, and arguably its most important member, was influential feminist and labor activist Florence Kelley (Sklar 1995). The organization used a mix of advocacy to push for government legislation, and consumer activism to promote an ethical marketplace, to achieve better working conditions and wages for women (and at times all) workers, and to promote better food and safety standards for consumers. Their early work focused on the harsh conditions that American workers faced, including advocating against sweatshop labor, for maximum hours and minimum wages, and for protective legislation for women workers. In the 1920s and 1930s they lobbied most strongly for maximum hours and minimum wage legislation for women workers. The organization continues today, focusing on occupational safety and consumers' rights (Storrs 2000).

Our subcollection includes 100 documents from the WMS NCL collection published between 1899 and 1935. The documents in this subcorpus are focused primarily on NCL meetings, including meetings notes, proceedings, minutes, and resolutions, as well as a few bulletins and invitations for NCL events. Our NCL subcorpus comprises 657,743 words, with an average of 6,577 words per document (see Table 1).

Table 1. Description of documents from three selected women and social movements subcorpora: 1899-1935

Primary Source Set	Document Count	Total Word Count (Inclusive)	Mean per Document Word Count	Dates Covered	Description
Writings of Black Women Suffragists (WBWS)	697	3,407,150	1,755	1899-1935	Essays, letters, and meeting notes
Equal Rights (journal) (ERJ)	408	1,191,764	8,351	1923-1935	Weekly journal of the NWP
National Consumers' League (NCL)	100	657,743	6,577	1903-1935	Meeting notes and proceedings

Wikipedia

Our data for Wikipedia were collected via the Wikimedia Foundation's Wikipedia dump—a collection of almost all of Wikipedia's data created and released to the public two times every month.[2] We downloaded the XML file from the August 20, 2020 data dump, including a snapshot of all articles on Wikipedia at the time of the dump but not the revision history or the talk pages.[3] We converted the XML file into JSON format for text analysis, preserving some of the original XML metadata.[4] Our Wikipedia data include a total of 15,403,173 English-language Wikipedia articles.

METHODS

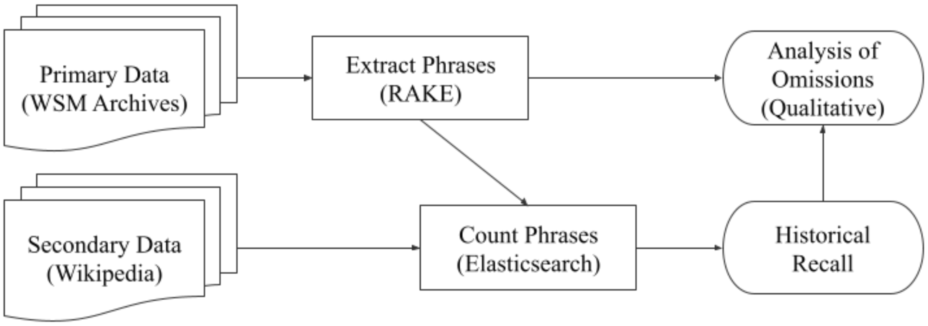
Our analysis proceeded in three steps: (1) we extracted discrete information from our primary data (our first-person empirical evidence) using phrase mining; (2) we calculated historical recall by identifying which of those phrases were present in our Wikipedia data; and (3) we analyzed phrases that were not present in select Wikipedia articles qualitatively, to identify mechanisms of omission. Figure 1 diagrams our general analytic steps and specific data and methods used. Based on our knowledge of how these subsections of the women's movement have been covered in histories, we expected the ERJ, as the mainstream organization now unquestionably at the center of the women's suffrage and women's rights movement, to be the most well-covered in Wikipedia, followed by the NCL, as a largely white yet working-class organization, followed by WBWS, as historians have argued that the roles of women of color in the women's movement have been either ignored or underplayed in histories of this movement.

[2] The data are available to download by anyone here: https://dumps.wikimedia.org/backup-index.html.

[3] Wikipedia has several namespace or page categories, such as *Article*, *Talk*, *User*, *User Talk*, *Template* etc. Among them, we selected the namespace *Article*.

[4] Specifically, we kept page_title, page_text, page_text_format, namespace, last_edit_date, last_edit_comment, last_edit_contributor.

Figure 1. General analytic steps (and particular data and/or methods used in this paper) to measure historical recall


Primary Data (WSM Archives)
Extract Phrases (RAKE)
Analysis of Omissions (Qualitative)
Secondary Data (Wikipedia)
Count Phrases (Elasticsearch)
Historical Recall

Note: This figure shows our general analytic steps, from data collection through analysis, to measure historical recall. The specific data and/or methods used in this paper for each step are in parentheses below the general step. Others using this approach will use different data and may use different methods in each step.

Conceptually, we first needed to operationalize and identify all relevant information from our primary corpus. In NLP, *information* is defined as sequences of characters that make up words, and sequences of words that make up phrases (a phrase can be a single word). Phrase mining, a subfield in NLP, aims to extract *quality* phrases from text: one-or-more word phrases that are complete semantic units and that denote entities—named entities, things, ideas, or concepts—recognizable as important to human readers (Cao et al. 2020; Shang et al. 2017). We operationalized relevant information as these one-or-more word phrases present in our primary corpus.

Many phrase mining techniques tend to work best on contemporary, domain-specific documents, and often require significant domain knowledge and hand-coding of documents or phrases (Wan and Xiao 2008; Witten et al. 1999; Zhang et al. 2008). Recent phrase mining techniques have proposed more domain-agnostic and fully automated approaches to phrase mining by using Wikipedia to provide lists of candidate phrases (Shang et al. 2017). We sought a method that is both domain agnostic, in that it can be applied to almost any historical time period, and, with a goal of evaluating omissions in existing knowledge, a method that does not rely on expert knowledge or lists, including Wikipedia.

RAKE is a well-known statistical keyword extraction method that is unsupervised, domain-independent, and nearly language independent, making it ideal for our purposes (Rose et al. 2010).[5] Unlike machine learning methods, the RAKE method uses the same, deterministic and interpretable steps to extract phrases from any document. Its purely statistical approach, however, does come at the cost of precision. Many of the phrases the algorithm identifies will be false positives: common words or idiosyncratic phrases that are not relevant to the field or corpus. Because we are aiming to identify *all* possible information that could be included in historical accounts of a topic, we see RAKE's algorithmic preference for recall over precision as a benefit for our use.

We implemented the RAKE algorithm on each of the documents in our corpus using the Python package python-rake 1.5.0,[6] removing digits from the text but doing no other

[5] This approach will not work on languages, such as Mandarin, that do not separate words via white space.
[6] https://pypi.org/project/python-rake/

pre-processing steps prior to extracting the phrases (we did additional cleaning on the extracted phrases and via our phrase matching process, detailed below).[7] We specified that phrases had to have at least three characters, a maximum of five words, and occur in at least one document two times to be included as candidate phrases.[8] This resulted in a total of 32,295 unique phrases of between one and five sequences of words.

Two of the authors systematically hand-coded each phrase for its relevance to the women's movement for a large set of these phrases (~20,000 phrases). Both authors found both true and false positives, as expected. Both also agreed, however, that hand-selecting true positives from these extracted phrases was prohibitively time consuming, but more importantly, we found it was difficult to devise criteria for what constituted a true positive. Does *women* constitute a true positive? It is a common word, but it is also a central concept to the women's movement. What about *night*? Here again, this is a common word but also central to the movement—many of the labor campaigns were about night work. We thus determined that hand-selecting true positives from the list would introduce its own potential biases into the process of constructing our collection of primary historical information.[9]

Instead, we opted for minimal hand cleaning, even as it came at the expense of some precision, making the entire process more reproducible and scalable. From the extracted phrases,

[7] We did not remove punctuation, stop words, or stem or lemmatize words, standard preprocessing steps in many text analysis applications. First, the RAKE algorithm relies on punctuation and stop words to identify key phrases. Second, the definition of a phrase for phrase mining purposes includes, among other criteria, a sequence of words (including single words) that form a complete semantic unit in the context of a corpus (Shang et al. 201, 3). Stemming words in a phrase would violate that definition, as many stemmed words are not real words (e.g., *jumping* is often stemmed to *jumpi*). Phrase mining papers typically do not stem or lemmatize words, opting to instead preserve the original semantic units (e.g., Cao et al. 2020; Shang et al. 2017). We follow that practice here.

[8] Phrases (including one-word phrases) that occur only once are often misspellings or typos, or lists of names of those present in meeting notes, or other idiosyncratic words.

[9] Others have similarly found that single-word phrases are quality phrases and should not be discarded, e.g., Zhang et al. 2008.

we replaced common punctuation marks that were not used in the RAKE algorithm to differentiate phrases, and we removed the gendered marital titles—Mr., Mrs., and Miss.—preceding full names (e.g., miss ida b wells became ida b wells).[10] Our final, cleaned phrase list, we believe, has high recall: it includes nearly all of the possible information present in these primary documents.

To account for the distributional properties of natural languages—for example, many words and phrases will appear in virtually all documents, regardless of topic—and to verify our phrase extraction method picked up quality phrases from the primary data, we used the Brown Corpus, an electronic collection of text samples of American English compiled in 1961 (Kucera, Francis, and Carroll 1967), to establish a baseline for how many English-language phrases we expect to appear by chance on Wikipedia. From the Brown corpus we used the same RAKE method to extract a random sample of key phrases, stratified to match the number of one-, two-, three-, four-, and five-word phrases identified in our primary data. If a similar proportion of phrases extracted from our primary data and those extracted from the Brown corpus appear on Wikipedia, this would suggest that either our method or our data are essentially picking up random noise—phrases that occur in written material regardless of topic. If a higher proportion of the phrases extracted from our primary data appear in Wikipedia compared to the phrases extracted from the Brown corpus, this would suggest that, unlike random phrases, the phrases we identified in our primary data are capturing information notable enough to merit mentions in Wikipedia, verifying our information-extraction method is identifying actual historical *information*, not simply linguistic noise.

[10] It was common practice from this era to include these titles when listing names, but this practice is not used on Wikipedia.

We then used Elasticsearch to measure whether the phrases from our primary material and the Brown corpus were present in the Wikipedia data. Elasticsearch is an open-source database based on the Apache Lucene library.[11] By using an inverted index, it allows advanced, rapid text searching over a large amount of documents. Elasticsearch itself has several text preprocessing pipelines. We used the default text preprocessing pipeline, the standard text analyzer, which removes most punctuation and converts the text into lowercase.[12] After pre-processing, we broke each of the extracted phrases into multiple terms by splitting them on spaces and/or hyphens. We then did a multi-term search query over all of the Wikipedia articles that we indexed in the Elasticsearch database, using the default value for fuzziness, a parameter that allows matches to terms that are as much as two edit distances away, depending on the size of the term[13] (e.g. *women* would match with *women* and *woman*, as *woman* is only one letter difference, or one edit distance, away from *women*).[14] We also allowed at least one intervening term appearing between the ordered terms coming from the phrase lists to ensure our search did not miss the mention of the phrases due to differences in stopwords or punctuation.[15] As one example, the way we implemented Elasticsearch matched the phrase *women's suffrage* to multiple versions of the phrase, including variations such as *Women's Suffrage*, *woman suffrage*, and *womens suffrage*.

[11] https://www.elastic.co/what-is/elasticsearch, accessed January 27, 2021.

[12] The preprocessing pipeline is based on Unicode specification: https://unicode.org/reports/tr29, accessed July 23, 2021.

[13] If the term has less than three letters it requires an exact match. If the term is between 3 and 5 letters it will match up to one edit distance. For terms more than 5 letters it allows for two edit distances. See https://www.elastic.co/guide/en/elasticsearch/reference/current/common-options.html#fuzziness, accessed July 23, 2021.

[14] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-span-multi-term-query.html, accessed January 27, 2021.

[15] Here again we did not remove stop words or stem or lemmatize words, as the goal of phrase mining is to identify corpus-specific complete semantic units, but we did allow for these fuzzy matching criteria to capture minor changes in language over time and across publication contexts.

To capture the different ways readers may search for and read information about a historical topic in Wikipedia, we searched for phrases in three sets of Wikipedia articles. First, we searched the full text across all Wikipedia articles in our data—the most comprehensive search for whether a phrase is present in Wikipedia. Second, if a word or phrase occurs in the title of a Wikipedia page it indicates that that concept is notable enough to merit its own article. To capture notable phrases, we searched across Wikipedia titles, using the metadata tag page_title. Third, if a casual reader wants to learn about a movement more broadly, they will likely read an article such as "Feminist movement" or "Black feminism," rather than searching for a particular organization or concept. To capture these articles, we searched for phrases in articles with *movement*, *history*, or *feminism* or *feminist* in the title (what we call history and movement pages below).

RESULTS

Primary Evidence: Women's Movement Discourse

Of the 32,295 unique phrases identified using the RAKE algorithm, 18,095 phrases occurred in ERJ, 13,249 occurred in WBWS, and 8,642 occurred in NCL. The extracted phrases were a mix of:

1. *movement actors*, including people (e.g., Ida B. Wells), organizations (e.g., National Woman's Party), organizational structures (e.g., women's committee), groups outside of movements (e.g., city council), and abstract groups (e.g., American Feminists)
2. *movement events* (e.g., Seneca Falls Convention)

3. *movement constituencies* (e.g., working women, married teachers, French women)
4. *movement grievances and targets* (e.g., race prejudice, maternal death, canneries)
5. *proposed laws, bills and acts* (e.g., Sheppard Towner Act, minimum wage law)
6. *general movement ideas* (e.g., social uplift), *solutions* (e.g., accident compensation), and *goals* (e.g., equal rights)
7. *general public sphere institutions* (e.g., Cornell University, A.M.E. Church).

Within this broad common structure, however, we found far more differences than similarities in the key phrases across our three groups. Of the 32,295 unique phrases, only 8% (2,724 phrases) occurred in all three collections. Of these shared phrases, 91% were one-word phrases (representing more common words), 9% (247) were two-word phrases, and only seven were three-or-more-word phrases (e.g., *new york city, minimum wage law*, *christian temperance union*). Table 2 shows the frequently used key phrases from those that occurred in all three subcorpora. The common areas of focus across these three groups centered on working conditions (e.g., *minimum wage*), *civil service*, *domestic service*, and education (e.g. *public schools*). The NCL mentioned other working-class issues, such as *working conditions* and *working hours*, more frequently than the other two groups; ERJ paid more attention to *equal pay* and *civil service*; and WBWS mentioned *high school*, *civil war*, *home life*, *world war*, and *christian temperance union* relatively more frequently than the other two groups.[16] WBWS was the only subgroup to frequently mention *colored women* as their constituency and ERJ to

[16] The Women's Christian Temperance Union was one of the early integrated women's organizations and was a key organization for the early Black women's movement, though many of their local unions remained segregated.

frequently mention *american women*; both ERJ and NCL frequently mentioned *working women* and *women voters* as their constituencies.

Table 2. Frequently used common phrases from three primary subcorpora from the women and social movements subcollection: 1899-1935

Writings of Black Women Suffragists	Equal Rights Journal	National Consumers' League
colored women	american women	supreme court
colored people	equal pay	minimum wage
young women	women voters	working conditions
public schools	working women	working women
high school	supreme court	minimum wage law
domestic service	young women	public opinion
civil war	civil service	women voters
home life	minimum wage law	working hours
world war	minimum wage	american federation
christian temperance union	public schools	public schools

Source: Three subcorpora from *Women and Social Movements in the United States, 1600-2000* and Wikipedia articles.

Compared to the 8% of phrases occurring in all three subcorpora, 75% of the phrases (24,347) occurred in only one collection: 7,911 phrases only occurred in WBWS (60% of all of the WBWS phrases, 25% of all of the phrases), 12,121 only occurred in ERJ (67% of ERJ phrases, 39% of all phrases), and 4,315 phrases only occurred in NCL (50% of NCL phrases, 14% of all phrases). Table 3 shows frequently used phrases unique to each group, suggesting a substantive difference in constituencies, ideas and solutions, and people across these three groups. WBWS uniquely mentioned *white women* and other racial markers, and they were the only subgroup to mention *booker* [T. Washington] and *frederick douglass* in their writings. *Lynching* was a unique focus of this group, as was *day nurseries* (used by working mothers), *uplift* (racial uplift was an important concept to this movement), and Black educational institutions, including the *tuskegee institute* and *howard university*. ERJ was the only group to mention the *equal rights amendment*, *jury service,* and the *inter-american commission* (more on these issues below). They were also the only group to mention the activist *alice paul* and to use the word feminist (*feminist movement*, *feminist*, and *feminism*). NCL uniquely focused on consumers and food safety (*consumer, label, pure food law, industrial poisons,* and *white list*), as well as sweatshop labor (*inspectors, sweating system*).

Table 3. Frequently used phrases unique to each of the three primary subcorpora from the women and social movements subcollection: 1899-1935

Writings of Black Women Suffragists	**Equal Rights Journal**	**National Consumers' League**
colored woman	maud younger	label
white women	equal rights amendment	florence kelley
lyching	alice paul	consumer
booker	alva belmont house	white list
tuskegee institute	international law	food committee
flesh	jury service	inspectors
frederick douglass	inter-american commission	industrial poisons
uplift	feminist movement	sweating systems
howard university	women lawyers	minimum wage boards
day nurseries	juries	pure food law

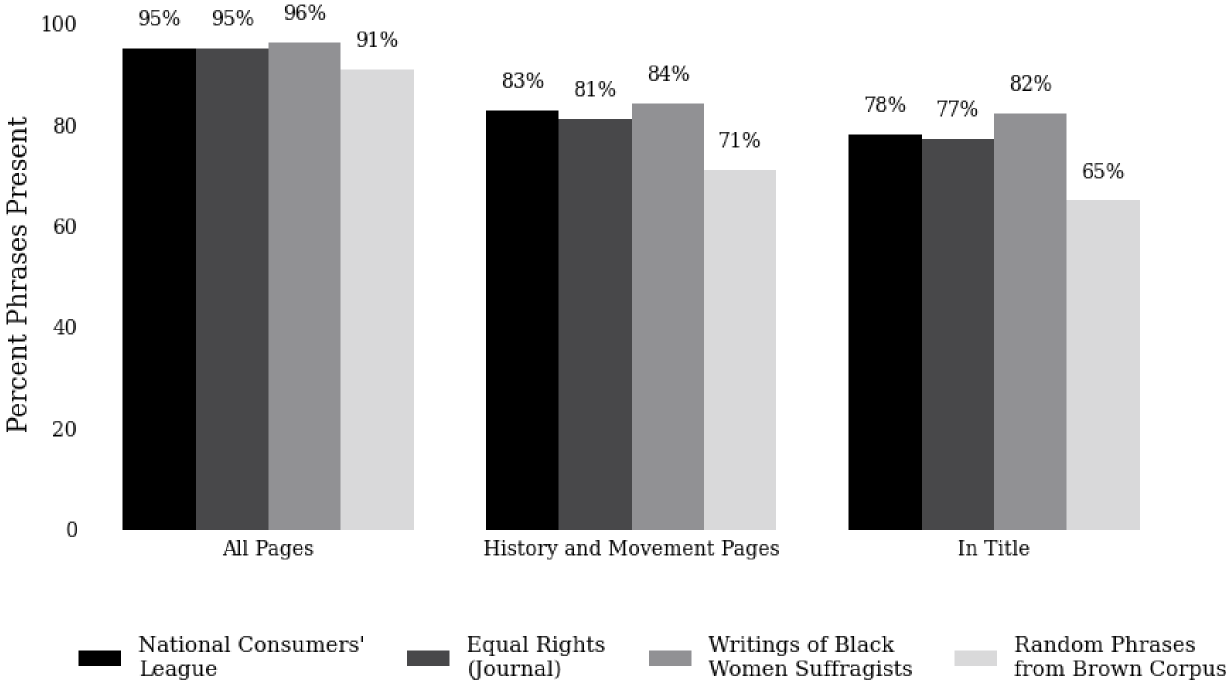
Source: Three subcorpora from *Women and Social Movements in the United States, 1600-2000* and Wikipedia articles.

This brief look at frequent key phrases provides a surprisingly reliable summary of the key issues, solutions, people, and institutions important to these movements. This exploration of phrases also confirms what scholars have long claimed: histories that focus on only one subsection of the larger women's movement—for example the national suffrage movement as led predominantly by professional white women—only cover a narrow part of the issues and concepts important to different sections of this movement. In other words, equal rights and feminism are not adequate stand-ins for the women's movement writ large. A collective history that is comprehensive ought to, in theory, cover all (or most) of the different key issues across all different subsections of this movement.

Historical Recall: Wikipedia Coverage

Figure 2 shows the percentage of phrases identified in the primary material that occurred in any Wikipedia article, in history and movement pages (defined above), and in article titles by subcorpus, as well as the random phrases extracted from the Brown corpus as a baseline. More than 95% of all phrases across all groups appeared somewhere on Wikipedia (compared to 91% of Brown phrases), between 81% (ERJ) and 84% (WBWS) showed up in history and movement pages (71% of Brown phrases), and between 77% (ERJ) and 82% (WBWS) showed up in Wikipedia titles (65% of Brown phrases). In short, the recall of Wikipedia is truly impressive.

Figure 2. Percent of key phrases in all Wikipedia articles, history and movement articles, and titles by subcorpus

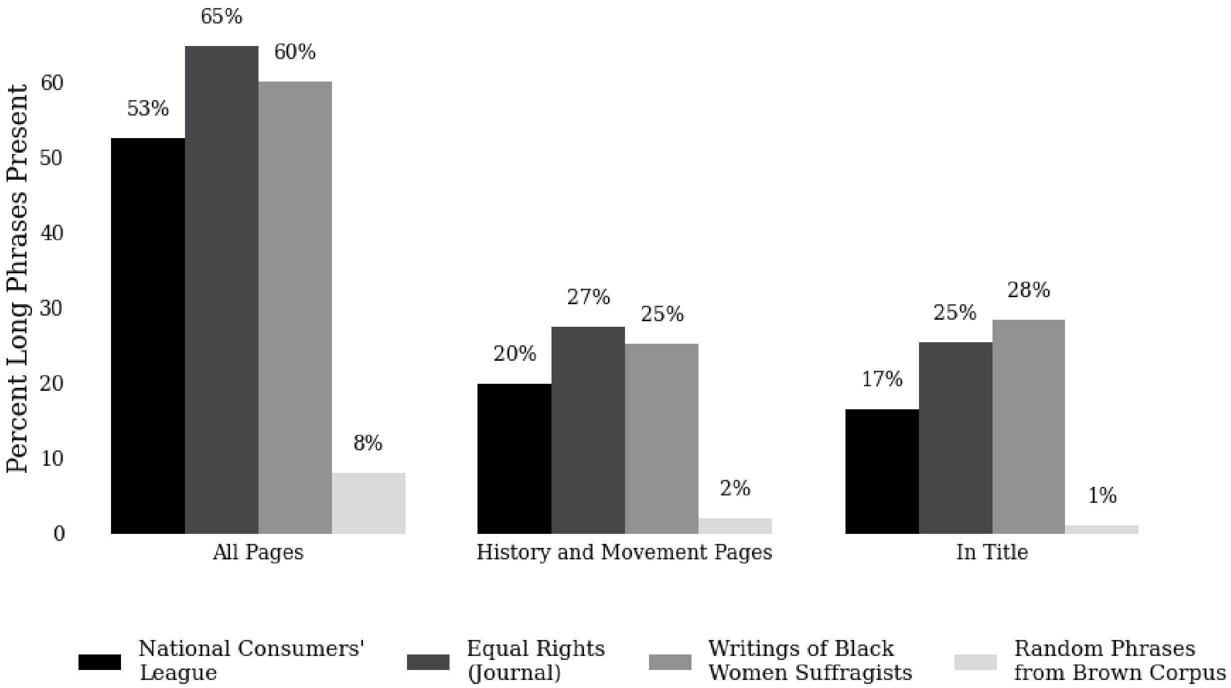


Source: Three subcorpora from *Women and Social Movements in the United States, 1600-2000*, the Brown Corpus, and Wikipedia articles.

The majority of this coverage, however (around 80%), were one- and two-word phrases: phrases important to the women's movement, but also phrases that commonly occur in language more generally (such as *women* and *night*). Of the phrases that did not occur on Wikipedia, between 73% (ERJ) and 80% (NCL) were three or more word phrases. These longer phrases capture information more distinctly related to the women's movement, evidenced as well by the low proportion of our random long phrases present on Wikipedia (see Figure 3). We found examining these longer phrases separately from the shorter, common phrases, revealed important patterns of historical omissions on Wikipedia.

Figure 3 shows the percent of long key phrases (three or more words) present in any Wikipedia article, history and movement articles, and Wikipedia titles, by subcorpus. Here we see a large difference between the proportion of phrases present from our primary data compared to the random phrases extracted from the Brown corpus. This finding suggests that the long phrases extracted from our primary data are capturing concepts from these women's movements that are notable enough to be included in Wikipedia, verifying our overall approach to extracting relevant primary information against which to measure historical recall. This figure also shows differences in rates of omission across the women's movement subgroups, with phrases from the ERJ and WBWS covered at similar rates, while the NCL phrases were more likely to be omitted from Wikipedia. Between 53% (NCL) and 65% (ERJ) of these long phrases appeared in Wikipedia articles (compared to 8% of Brown phrases), between 20% (NCL) and 27% (ERJ) appeared in history or movement pages (compared to 2% of Brown phrases), and between 17% (NCL) and 28% (WBWS) appeared in titles (compared to 1% of Brown phrases).

Figure 3. Percent of long key phrases (three or more words) in all Wikipedia articles, history and movement articles, and titles, by subcorpus



Source: Three subcorpora from *Women and Social Movements in the United States, 1600-2000*, the Brown Corpus, and Wikipedia articles.

Table 4 lists frequently used phrases that were *not* present in the history and movement Wikipedia pages, by subcorpus. This list includes names (*ida joyce jackson* [WBWS], *edith houghton hooker* [ERJ]), organizations and institutions (*frederick douglass center* [WBWS], *alva belmont house* [ERJ], *continental waist company* [NCL]), and, as we detail more below, key ideas and concepts important to the movement particpants (*jim crow car* and *domestic service occupations* [WBWS]; *equal nationality treaty* and *jury service bill* [ERJ]; and *minimum wage boards, pure food law,* and *federal child labor amendment* [NCL]). This last category in particular, the missing ideas and concepts, represent historical omissions that, we claim, lead to partial, and even inaccurate, historical interpretations of these movement subgroups. These missing long phrases can thus help guide historians to under-researched aspects of historical topics.

Table 4. Long phrases (three or more words) missing from history and movement Wikipedia pages from three primary subcorpora from the women and social movements subcollection: 1899-1935

Writings of Black Women Suffragists	**Equal Rights Journal**	**National Consumers' League**
frederick douglass center	edith houghton hooker	minimum wage boards
war camp community service	alva belmont house	pure food law
ida joyce jackson	lucretia mott amendment	minimum wage bill
phyllis wheatley home	equal rights treaty	forty-eight hour week
colored club women	special labor laws	industrial home work
jim crow car	hague nationality convention	federal child labor amendment
political equality association	equal nationality treaty	state factory inspector
miner normal school	equal nationality bill	continental waist company
high moral standards	equal nationality rights	eight hours law
domestic service occupations	jury service bill	misbranded malt liquors

Note: History and movement pages include pages with *movement*, *history*, *feminism*, or *feminist* in the title.
Source: Three subcorpora from *Women and Social Movements in the United States, 1600-2000* and Wikipedia articles.

Typology of Omissions

To explore how historians might use this method to identify both under-researched aspects of a historical topic and why certain aspects of this movement in particular are under-researched, we qualitatively compared select Wikipedia articles to the primary source material: we compared the article “National Consumers’ League” (Wikipedia 2020b) to the NCL subcorpus, the “National Woman’s Party” Wikipedia article (Wikipedia 2020a) to the ERJ subcorpus, the Wikipedia articles “African-American woman suffrage movement” (Wikipedia 2020a) and “Black feminism” (Wikipedia 2020b) to the WBWS subcorpus. We searched for and read the context around the most frequent phrases in the corpus, including the phrases omitted from Wikipedia, in both Wikipedia and the primary corpus, reading additional Wikipedia articles as needed for more context. This analysis produced a (partial) typology of omissions on Wikipedia: paucity, restrictive paradigms, and categorical narrowness.

Paucity: National Consumers’ League and Working-Class Women. Among the criteria for quality articles on Wikipedia are length, the presence of images and other media, and the number of sources listed (Wikipedia 2021c). Along all of these dimensions, the Wikipedia article on the NCL is low quality. It is brief, at around 1,380 words, it contains only ten references, and only one suggested reference for further reading. At the top of the article is an actual warning from Wikipedia: “This article includes a list of references, related reading or external links, but its sources remain unclear because it lacks inline citations. … (September 2020).” While the Wikipedia article provides bits of key information about the NCL, it does not go into depth about the many campaigns the NCL participated in and their importance to history.

One of the issues important to the NCL but omitted from the Wikipedia article, is the establishment of minimum wage boards. Kelley, who pioneered the use of sociological evidence in Supreme Court cases (Dreier 2012), published her research on minimum wage boards in the *American Journal of Sociology* (Kelley 1911), indicating its importance to her at the time but also to political and economic history more broadly. The NCL literature mentions minimum wage boards 46 times in our subcorpus, and in the 1910s, they had an entire special committee devoted to these boards, regularly reported on this work, and convinced states to appoint commissions to study their feasibility. Finding information on wage boards on Wikipedia, however, and their connection to Kelley or the NCL, is nearly impossible. Even the brief article on the Trade Boards Act of 1909 does not mention minimum wage boards, and it is never mentioned in the NCL Wikipedia article.

The protection of in-home workers (workers who completed work, such as sewing, in their homes) from exploitation as well as the Pure Food and Drug Act of 1906 are similarly not included in the Wikipedia NCL article. *Home work*, *pure food* and *food handler* were mentioned over 100 times each in the NCL documents, yet none of these issues are mentioned in the NCL Wikipedia article, and the article on the Pure Food and Drug Act on Wikipedia does not mention the work the NCL did to get the act passed. Without additional knowledge, casual Wikipedia readers may never know that NCL was involved in these campaigns.

The number of words in the NCL subcorpus (~650,000) was less than one fifth of the words included in the WBWS subcorpus, and thus the comparatively fewer details on the NCL may indeed be proportionate to their impact and/or recorded archives (though true impact is difficult to define). This paucity, however, confirms what historians have long claimed: there is a general inattention to working-class women in the women's movement and an inattention to the role of women in the labor movement more broadly (Cobble 2005; Milkman 1985; Orleck 1995). The method presented here can help identify (or confirm) broad areas where historians could do more work narrating, and it can also point to specific information (such as protection of in-home workers) that is conspicuously absent from published histories.

Restrictive Paradigms: National Woman's Party. Unlike the NCL, the NWP is a relatively well-known and well-researched mainstream organization. Their Wikipedia page is over 6,000 words, and it contains multiple images, comprehensive tables, forty-four notes, and ten links to further readings. The bulk of the article, however, suggests published histories of the NWP are artificially restricted, or limited to, the paradigmatic examples of suffrage and the equal rights amendment, to the detriment of a complete understanding of this organization.

For example, while the Wikipedia article devotes over 1,250 words to describe notable leaders of the organization, including listing the leaders from every single state in the United States, the article never mentions the words *jury*, *juries*, or *nationality*. Between 1923 and 1935, the Equal Rights Journal used the word *jury* or *juries* over 1,700 times. The right to serve on juries was an extensive, long, 50-state battle carried out by the women's movement after suffrage was won, led in part by the NWP (McCammon 2012). As an example of the breadth but also potential inconsistencies in Wikipedia, the page titled "Women in United States juries" does link to the National Woman's Party page, acknowledging their work on the jury movement. In other words, an informed reader might specifically search for this aspect of the NWP's work and find it on Wikipedia, but the casual reader would know nothing about it from reading only the NWP's page.

The NWP's work on the issue of equality in nationality is equally ignored on Wikipedia. In many countries during these years, including the United States, women lost their nationality upon marriage to a citizen of a different country, and had no control over their assets and children. The Convention on the Nationality of Women, adopted by the Pan American Union in 1933, was the first internal treaty ever adopted concerning women's rights—an important historical moment on its own terms. NWP member Doris Stevens worked extensively on this campaign, supported by the NWP. Between 1923 and 1925, the Equal Rights Journal mentioned *nationality* over 3,580 times, *married women* over 1,500 times, and the *equal rights treaty* 184 times—more attention than they gave the jury movement. Despite the historic importance of this treaty, for the international women's movement and for the living standard of women across the globe, *nationality* is never mentioned in the Wikipedia article on the NWP. There is a short Wikipedia article on the Convention on the Nationality of Women, but the article does not mention the NWP. To get to the NWP from the article on the convention, a reader would have to click on the Doris Stevens link, which then links to the NWP page.

In sum, the paradigmatic association between the National Woman's Party and suffrage and the ERA has produced notable blind spots and absences in the other important work done by the NWP, restricting our historical interpretation of this organization. The writings of Black women suffragists demonstrate a similar, but even more pernicious, type of omission: categorical narrowness.

Categorical Narrowness: Writings of Black Women Suffragists. The Wikipedia article "African-Amercian women's suffrage movement" is just over 3,000 words long, with twenty-two references and extensive links to further information. The article "Black feminism" is the most objectively high-quality article directly related to our corpus: it is a full 9,167 words, with multiple images, 102 references, and twelve books and articles referenced for further reading. Nonetheless, there are significant omissions in these two articles when compared to the WBWS subcorpus, rooted in a narrow or constricted idea of what should be classified as suffrage or feminist movements.

For example, domestic work was arguably one of the most important issues for Black women during the early 20th century. Domestic work was one of the only occupations open to Black women in both the north and the south, and it was rife with exploitation (Sharpless 2013; Williams 2002). This issue was prevalent throughout the WBWS subcorpus. In the WBWS subcorpus—a corpus selected and categorized by experts to represent the Black women's suffrage movement—the word *suffrage* was used around 340 times. The word *domestic*, alternatively, was mentioned 527 times in this subcorpus—55% more often than *suffrage*. Yet the word *domestic* is not mentioned at all in the Wikipedia article "African-American women's suffrage movement," and is mentioned only three times in the "Black feminism" article (one of these mentions is in relation to domestic violence). Even if the omission of *domestic* in the Wikipedia article on suffrage is not necessarily problematic, as that article focuses specifically on the right to vote, its absence from the "Black feminism" article is conspicuous, given its importance to early Black feminists.

General health concerns were another central issue for Black activists, suffragists, and feminists during this time. The words *health* and *hospital* were mentioned over 540 times in the subcorpus—close to 60% more often than *suffrage*—often mentioned alongside education and employment as the key issues at the center of the Black club movement (the main organizational center of Black women's activism during this time), particularly as hospitals and other public health programs almost always refused Black patients. Health, sanitation, and, specifically, tuberculosis (mentioned 100 times in the WBSW documents) were never mentioned in the two articles on Wikipedia.

A final example of a concern important to the first-wave Black women's movement, the issue of Jim Crow Cars, was covered on Wikipedia but not its relationship to the women's movement. This issue was particularly important to professional Black women. Segregated first-class cars were only open to white men and women, and second-class cars, open to both black and white people and which allowed smoking and drinking, exposed Black women to sexual harassment and assault that many white women could escape by buying first-class tickets (of course, women of all races who could not afford first-class tickets were also exposed to these threats). Jim Crow and segregation are not mentioned in the article on the African-American woman suffrage movement at all, and are mentioned twice, only briefly, in the article on Black feminism.

Like the issues around the jury movement and the equal nationality movement discussed in relation to the NWP, an interested and informed reader could find information on Wikipedia about issues left out of the main articles on Black feminists and suffragists. In our efforts, however, it was much more difficult to find information on the role of Black women in the domestic worker and anti-segregation movements compared to the jury and equal nationality movements. After much searching, we found an article called "Domestic worker," which is an impressive 9,267 words with 80 references. This article has a short section on Black domestic workers, but it does not mention the many women's organizations that fought for better working conditions for Black domestic workers. We could find no information on the role of Black *women* in the domestic worker's rights movement in our search of Wikipedia. The article "Jim

Crow laws," additionally, details at length segregation in trains. In the section on early attempts to break Jim Crow, however, women are not mentioned once, despite their role in the early lawsuits against Jim Crow Cars (Giddings 2009).

Similar to the NCL, there is a paucity of information on the role of Black women in important movements such as domestic-worker rights. Similar to the NWP, issues important to the early Black women's movement, such as employment, health, and segregation, were not included in the main pages for these movements. We call these omissions categorical narrowness, however, because we see a different mechanism at play in the case of WBWS. The jury movement is unequivocally seen as a women's issue—women's suffrage is linked from the Women in United States juries page, and the role of women's organizations such as the League of Women Voters and the NWP are mentioned on the page. The phrase *women's rights* is explicitly mentioned on the page discussing the Convention on the Nationality of Women. Domestic workers rights, particularly in the early 20th century, Jim Crow segregation, and health issues are simply not coupled with women's rights or feminism on Wikipedia. This is a result of categorical narrowness: issues that are not categorized as women's rights issues, despite their importance to the women's movement itself, represent another distortion of history.

DISCUSSION AND CONCLUSION

By comparing information extracted from primary historical evidence to Wikipedia, we provided a method and approach to measure the scope and recall of the largest and most popular and accessible English-language collection of historical knowledge. We found that over 95% of the key phrases used by the movement actors appeared on Wikipedia. Much of this recall, however, was one-word phrases that are simply commonly used in the English language. When digging into the missing 5%, we discovered rich data for interrogating patterns around omissions in our collective historical consensus. As expected, we found Wikipedia contains fewer details about working-class women compared to professional white women, but contrary to our expectations, we found similar rates of coverage between professional white women and Black women, perhaps because of important efforts by historians to recover and re-narrate the important roles played by Black women in the women's movement. Even for the groups with more comprehensive coverage, we identified patterns in what is omitted and why. In particular, when paradigmatic examples are too tightly coupled with an organization or topic, or categories are too narrowly (and ahistorically) defined, ideas important to historical actors are relegated as background noise, as well-meaning scholars transform facts into historical patterns.

Our research has three important implications. First, we found that phrase-mining primary historical texts is an effective method for identifying a broad range of relevant information related to a historical topic. While not the focus of this paper, the phrases themselves could be analyzed on their own to describe and explore the distinct issues, foci, constituencies, and institutions important to different movement groups. As sophisticated but complicated text analysis methods continue to be incorporated into sociology, such as topic models (DiMaggio, Nag, and Blei 2013; Mohr and Bogdanov 2013), word embeddings (Kozlowski, Taddy, and Evans 2019; Stoltz and Taylor 2021), and other machine learning and deep learning methods (Edelmann et al. 2020; Evans and Aceves 2016; Molina and Garip 2019), scholars would be wise to keep simpler and more interpretable methods such as phrase mining in their toolkit. In particular, unlike many machine learning methods which identify clusters of words that are assumed to represent abstract themes, phrase mining preserves the actual language used in primary material and is thus more appropriate for identifying and operationalizing concrete and discrete information conveyed in text, including specific ideas and concepts—a frequent task in many content analysis projects (see also Cao et al. 2020).

Second, our findings and proposed method can help guide historians toward important gaps in the historical record. In a more superficial way, this historical correction could start with Wikipedia. Informed Wikipedia volunteers could use these methods to target their search for sources, filling in omitted information. Small tweaks to what already exists on Wikipedia could further make it easier to find linked information on important concepts relevant to a historical topic. Adding the jury and equal nationality movement as examples of NWP campaigns, for example—and linking to relevant Wikipedia pages—would not deviate from the main narrative of their article, but would provide a more comprehensive account of this organization to casual readers.

On a deeper level, however, this is not a job for Wikipedia volunteers. Research that has not been published simply cannot be included on Wikipedia. Scholars can use this method to identify topics that could benefit from additional historical attention or, if the publications exist, improved accessibility, ensuring a field of knowledge is truly inclusive. In short, these findings suggest that both historians and Wikipedians can leverage increasing access to digital primary historical sources to help flag omissions or under-reported knowledge, and the statistical phrase mining of primary texts is a promising method to do so.

Third, via a qualitative analysis of select Wikipedia articles, we identified a typology of historical omissions—paucity, restrictive paradigms, and categorical narrowness. These *mechanisms of omissions*, we argue, clarify how and why some information about a historical topic is considered relevant, and why some information is omitted. While paradigms and categories are crucial for building and conveying patterns in history, they can also serve to distort and confine our historical understanding. Scholars should periodically reexamine categorical assumptions about a topic, including how topics are categorized in both primary and secondary collections. Using phrase mining on large amounts of primary source text can provide a helpful lens to question taken-for-granted knowledge, patterns, and categories related to historical topics.

Limitations and Further Research

We see this research as merely scratching the surface of how text analysis methods can be used to enhance our understanding of the comprehensiveness of historical fields, and our findings prompt more questions than they answer. For example, there was no reliable way to determine whether an article in Wikipedia was specifically about women or women's movements. One extension of this approach could include better classifying articles that are specifically about women or women's movements, narrowing the analysis of historical omissions to these more relevant articles. We did not, additionally, distinguish between whether the omissions we identified were due to a lack of reliable peer-reviewed publications documenting these aspects of the movements, or because the Wikipedia articles simply do not adequately cover existing publications. Future research could use other data sources, including *Scopus* or *Microsoft Academic Graph*, to identify published sources that do potentially cover these omitted details, identifying which details are missing from secondary histories, and which are simply omitted from Wikipedia. These existing large scientific databases, however, currently focus mainly on journal articles from the physical sciences, which is why we did not include these data here. Historians and other scholars working in book fields should build on existing scholarly databases to ensure they adequately cover social science and humanities scholarship as well as the physical sciences.

We used the frequency of key phrases in primary source texts as one measure of the importance of that phrase to movement participants. This is not the only way to measure importance. Movement coverage in newspapers, for example, is commonly used by social movement scholars to identify important features and successes of social movements. Future research could include newspaper data, for example from the newspaper database *Chronicling America*, to identify which of these phrases were covered in contemporary newspapers, and whether newspaper coverage is a reliable indicator of their inclusion in later historical accounts. Newspapers, of course, have their own set of biases, and should be used with caution when identifying ideas and facts relevant to a historical topic.

Finally, the primary corpus we used came from a curated library of documents, chosen by editors for their content and importance, and represents a very small slice of the primary record of this movement. Future research could expand this to other curated collections, but could also work to include more documents from those not typically included in archives of this era. This could include writings from LGBT activists, Native American and other indigenous women, and other races, ethnicities, and religious groups. Historians could also continue to work to make available non-traditional documents—oral histories, stories, and songs, for example—that better represent non-elite ways of recording information. Of course, information that was simply never recorded is still important to try to reconstruct, and will never be captured using quantitative methods such as this (Risam 2018).

Each of these choices—in particular what primary and what secondary material to include in the analysis—capture different types and moments of omissions and biases. Further research could extend this method to other primary and secondary sources, some of which we listed above, systematically capturing different ways and moments in which biases and omissions are introduced into the historical record.

Understanding what and who gets included in history, and what and who does not, is a long standing, ongoing concern for both historians and the public. Historians have done important work documenting and narrating diverse and complex historical topics such as the women's movement, but these histories can always be improved. The case study presented here suggests one way we can leverage new methods and data to better measure the scope of existing histories at scale, and can be used to guide historians and Wikipedians alike as they work to fill in gaps and omissions that can distort the way we remember history. Once this information is recorded in large information systems such as Wikipedia, the rest, as they say, is history.

Author's Note: Replication instructions can be found in Appendix A: Replication Material.

REFERENCES

Adams, Julia, Hannah Brückner, and Cambria Naslund. 2019. "Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the 'Professor Test.'" *Socius* 5, January.

Adams, Katherine H., and Michael L. Keene. 2008. *Alice Paul and the American Suffrage Campaign*. Urbana: University of Illinois Press.

Autry, Robyn. 2013. "The Political Economy of Memory: The Challenges of Representing

National Conflict at 'Identity-Driven' Museums." *Theory and Society* 42 (1): 57–80.

Bao, Patti, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. "Omnipedia: Bridging the Wikipedia Language Gap." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1075–84. New York, NY, USA: Association for Computing Machinery.

Brown, Richard Harvey, and Beth Davis-Brown. 1998. "The Making of Memory: The Politics of Archives, Libraries and Museums in the Construction of National Consciousness." *History of the Human Sciences* 11 (4): 17–32.

Buechler, Steven M. 1986. *The Transformation of the Woman Suffrage Movement: The Case of Illinois, 1850-1920*. New Brunswick, NJ: Rutgers University Press.

Cahill, Cathleen D. 2020. *Recasting the Vote: How Women of Color Transformed the Suffrage Movement*. Chapel Hill: The University of North Carolina Press.

Callahan, Ewa S., and Susan C. Herring. 2011. "Cultural Bias in Wikipedia Content on Famous Persons." *Journal of the American Society for Information Science and Technology* 62 (10): 1899–1915.

Cao, Hancheng, Mengjie Cheng, Zhepeng Cen, Daniel A. McFarland, and Xiang Ren. 2020. "Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora." *ArXiv:2010.06657 [Cs]*, October. http://arxiv.org/abs/2010.06657.

Cobble, Dorothy Sue. 2005. *The Other Women's Movement: Workplace Justice and Social Rights in Modern America*. Princeton, NJ: Princeton University Press.

Davis, Elizabeth Lindsay. 1922. *The Story of the Illinois Federation of Colored Women's Clubs: 1900-1922*. Chicago: Illinois Federation of Colored Women's Clubs.

Davis, Natalie Zemon. 1976. "'Women's History' in Transition: The European Case." *Feminist Studies* 3 (3/4): 83-103.

Dill, Bonnie Thornton. 1979. "The Dialectics of Black Womanhood." *Signs* 4 (3): 543–55.

Dreier, Peter. 2012. "Florence Kelley: Pioneer of Labor Reform." *New Labor Forum* 21 (1): 70–76.

DuBois, Ellen Carol. 1971. "Feminism Old Wave and New Wave." CWLU Herstory Project. 1971. www.cwluherstory.org/feminism-old-wave-and-new-wave.html.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41 (6): 570–606.

Dunbar-Ortiz, Roxanne. 2014. *An Indigenous Peoples' History of the United States. ReVisioning American History*. Boston: Beacon Press.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. "Computational Social Science and Sociology." *Annual Review of Sociology* 46 (1): 61–81.

Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42 (1): 21–50.

Evans, Sara. 1980. *Personal Politics: The Roots of Women's Liberation in the Civil Rights Movement and the New Left*. New York: Vintage Books.

Firestone, Shulamith. 1968. "The Women's Rights Movement in the U.S.A.: A New View." In *Notes from the First Year*. New York: New York Radical Women.

Flexner, Eleanor. 1970. *Century of Struggle: The Women's Rights Movement in the United States*.

New York: Atheneum.

Foner, Eric. 2003. *Who Owns History? Rethinking the Past in a Changing World*. New York: Hill and Wang.

Giddings, Paula J. 2007. *When and Where I Enter: The Impact of Black Women on Race and Sex in America*. New York: William Morrow Paperbacks.

———. 2009. *Ida: A Sword Among Lions: Ida B. Wells and the Campaign Against Lynching*. New York: HarperCollins.

Goldstein, Dana. 2020. "Two States. Eight Textbooks. Two American Stories." *The New York Times*, January 12. https://www.nytimes.com/interactive/2020/01/12/us/texas-vs-california-history-textbooks.html.

Graells-Garrido, Eduardo, Mounia Lalmas, and Filippo Menczer. 2015. "First Women, Second Sex: Gender Bias in Wikipedia." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 165–74. HT '15. New York, NY, USA: Association for Computing Machinery

Hargittai, Eszter, and Aaron Shaw. 2015. "Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia." *Information, Communication & Society* 18 (April): 424–42

Hecht, Brent, and Darren Gergle. 2009. "Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories." In *Proceedings of the Fourth International Conference on Communities and Technologies*, 11–20. C&T '09. New York: Association for Computing Machinery.

Hendricks, Wanda A. 2013. *Fannie Barrier Williams: Crossing the Borders of Region and Race*. Urbana, IL: University of Illinois Press.

hooks, bell. 2000. *Feminist Theory: From Margin to Center*. Cambridge, MA: South End Press.

Jones, Martha S. 2020. *Vanguard: How Black Women Broke Barriers, Won the Vote, and Insisted on Equality for All*. New York: Basic Books.

Kelley, Florence. 1911. "Minimum-Wage Boards." *American Journal of Sociology* 17 (3): 303–14.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84 (5): 905–49.

Kucera, Henry, W. Nelson Francis, and John B. Carroll. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Lerner, Gerda. 1975. "Placing Women in History: Definitions and Challenges." *Feminist Studies* 3 (1/2): 5-14.

Liptak, Kevin. 2020. "Trump Says Department of Education Will Investigate Use of 1619 Project in Schools." CNN. https://www.cnn.com/2020/09/06/politics/trump-education-department-1619-project/index.html.

Lorde, Audre. 1984. *Sister Outsider: Essays and Speeches*. Trumansberg, NY: Crossing Press.

Lunardini, Christine A. 2000. *From Equal Suffrage to Equal Rights: Alice Paul and the National Woman's Party, 1910-1928*. San Jose, Calif.: ToExcel Press.

Maher, Katherine. 2018. "Wikipedia Is a Mirror of the World's Gender Biases." *Wikimedia Foundation* (blog). October 18.

https://wikimediafoundation.org/news/2018/10/18/wikipedia-mirror-world-gender-biases.

McCammon, Holly J. 2012. *The U.S. Women's Jury Movements and Strategic Adaptation: A More Just Verdict*. Cambridge ; New York: Cambridge University Press.

McNeill, WIlliam H. 1986. "Mythistory, or Truth, Myth, History, and Historians." *American Historical Review* 91 (1): 1–10.

Milkman, Ruth. 1985. *Women, Work, and Protest: A Century of US Women's Labor History*. Boston: Routledge & Kegan Paul.

Mohr, John W., and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41 (6): 545–69.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45 (1): 27–45.

Moreau, Joseph. 2004. *Schoolbook Nation: Conflicts over American History Textbooks from the Civil War to the Present*. Ann Arbor: The University of Michigan Press.

Morgan, Robin. 1970. *Sisterhood Is Powerful: An Anthology of Writings from the Women's Liberation Movement*. New York: Vintage Books.

Oeberst, Aileen, Ina von der Beck, Christina Matschke, Toni Alexander Ihme, and Ulrike Cress. 2020. "Collectively Biased Representations of the Past: Ingroup Bias in Wikipedia Articles about Intergroup Conflicts." *British Journal of Social Psychology* 59 (4): 791–818.

O'Kane, Caitlin. 2021. "Nearly a Dozen States Want to Ban Critical Race Theory in Schools." CBS. https://www.cbsnews.com/news/critical-race-theory-state-bans/.

Orleck, Annelise. 1995. *Common Sense & A Little Fire: Women and Working-Class Politics in the United States, 1900-1965*. Chapel Hill: University of North Carolina Press.

———. 2015. *Rethinking American Women's Activism*. American Social and Political Movements of the Twentieth Century. New York: Routledge.

Orlowitz, Jake. 2020. "How Wikipedia Drove Professors Crazy, Made Me Sane, and Almost Saved the Internet." In *Wikipedia @ 20: Stories of an Incomplete Revolution*, edited by Joseph Reagle and Jackie Koerner, 125–39. Cambridge: The MIT Press.

Parker, Alison M. 2020. *Unceasing Militant: The Life of Mary Church Terrell*. Chapel Hill: The University of North Carolina Press.

Polletta, Francesca, Pang Ching Bobby Chen, Beth Gharrity Gardner, and Alice Motes. 2011. "The Sociology of Storytelling." *Annual Review of Sociology* 37 (1): 109–30.

Reagle, Joseph, and Lauren Rhue. 2011. "Gender Bias in Wikipedia and Britannica." *International Journal of Communication* 5 (0): 21.

Riley, Denise. 1988. *'Am I That Name?'* London: Palgrave Macmillan UK.

Risam, Roopika. 2018. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." In *Text Mining*, 1–20. John Wiley & Sons, Ltd.

Roth, Benita. 2004. *Separate Roads to Feminism: Black, Chicana, and White Feminist Movements in America's Second Wave*. Cambridge, MA: Cambridge University Press.

Schwartz, Barry. 2003. *Abraham Lincoln and the Forge of National Memory*. Chicago: University of Chicago Press.

Scott, Joan W. 1986. "Gender: A Useful Category of Historical Analysis." *The American Historical Review* 91 (5): 1053–75.

Shang, Jingbo, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2017. "Automated Phrase Mining from Massive Text Corpora." *ArXiv:1702.04457 [Cs]*, March. http://arxiv.org/abs/1702.04457.

Sharpless, Rebecca. 2013. *Cooking in Other Women's Kitchens, Enhanced Ebook: Domestic Workers in the South,1865-1960*. Chapel Hill: The University of North Carolina Press.

Silverstein, Jake. 2019. "Why We Published The 1619 Project." *The New York Times*, December 20, sec. Magazine. https://www.nytimes.com/interactive/2019/12/20/magazine/1619-intro.html.

Sklar, Kathryn Kish. 1995. *Florence Kelley and the Nation's Work: The Rise of Women's Political Culture, 1830 - 1900*. New Haven: Yale University Press.

Slashdot. 2004. "Wikipedia Founder Jimmy Wales Responds." Slashdot. https://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds.

Smallwood, Stephanie E. 2016. "The Politics of the Archive and History's Accountability to the Enslaved." *History of the Present* 6 (2): 117–32.

Stanton, Elizabeth Cady, Susan Brownell Anthony, Matilda Joslyn Gage, and Ida Husted Harper. 1881. *History of Woman Suffrage*. Susan B. Anthony.

Stoltz, Dustin S., and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics*, May, 101567.

Tripodi, Francesca. 2021. "Ms. Categorized: Gender, Notability, and Inequality on Wikipedia." *New Media & Society*, June.

Vrandečić, Denny, and Heather Ford. 2020. "Automated Facts, Data Contextualization and Knowledge Colonialism: A Conversation Between Denny Vrandečić and Heather Ford on Wikipedia's 20th Anniversary." *Big Data & Society* (blog). December 10. https://bigdatasoc.blogspot.com/2020/12/automated-facts-data-contextualization.html.

Wagner, Claudia, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. "Women through the Glass Ceiling: Gender Asymmetries in Wikipedia." *EPJ Data Science* 5 (1): 1–24.

Wan, Xiaojun, and Jianguo Xiao. 2008. "Single Document Keyphrase Extraction Using Neighborhood Knowledge." In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, 855–60. AAAI'08. Chicago, Illinois: AAAI Press.

Ware, Susan. 2019. *Why They Marched: Untold Stories of the Women Who Fought for the Right to Vote*. Harvard University Press.

Wikipedia. 2020a. "Wikipedia:About." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=992972317.

———. 2020b. "Wikipedia:Manual of Style." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style&oldid=994999892.

———. 2020c. "Wikipedia:Notability." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Notability&oldid=995288718.

———. 2021a. "Wikipedia:No Original Research." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:No_original_research&oldid=998672086.

———. 2021b. "Wikipedia:Verifiability." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=999340133.

———. 2021c. "Wikipedia:The Perfect Article." In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Wikipedia:The_perfect_article&oldid=1003410614.

Williams, Fannie Barrier. 2002. *The New Woman of Color: The Collected Writings of Fannie Barrier Williams, 1893–1918*. DeKalb, Ill: Northern Illinois University Press.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. "KEA: Practical Automatic Keyphrase Extraction." In *Proceedings of the Fourth ACM Conference on Digital Libraries*, 254–55. DL '99. New York, NY, USA: Association for Computing Machinery.

Zhang, Ziqi, Jose Iria, Christopher Brewster, and Fabio Ciravegna. 2008. "A Comparative Evaluation of Term Recognition Algorithms." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Zinn, Howard. 2015. *A People's History of the United States*. Harper Perennial Modern Classics. New York: HarperPerennial.

Appendix A: Replication Material

Data

Women and Social Movements in the United States

Primary data were collected from the digital Alexander Street Press library *Women and Social Movements in the United States, 1600-2000* (WSM), found here: https://alexanderstreet.com/products/women-and-social-movements-library. Our data included all of the documents published between 1899 and 1935 from three primary source collections in the WSM library: *Writings of Black Women Suffragists,* the *Equal Rights (journal)*, and the *National Consumers' League* . These data are copyrighted and we do not have permission to distribute them, but they can be accessed through a subscription to the WSM library, or purchased in bulk from ProQuest.

Wikipedia

Wikipedia data were collected via the Wikimedia Foundation's Wikipedia dump: https://dumps.wikimedia.org/backup-index.html. We downloaded the XML file from the August 20, 2020 data dump, including a snapshot of all articles on Wikipedia at the time of the dump but not the revision history or the talk pages. Wikipedia has several namespace or page categories. Among them, we selected the namespace *Article*. We converted the XML file into JSON format for text analysis, preserving some of the original XML metadata. Specifically, we kept page_title, page_text, page_text_format, namespace, last_edit_date, last_edit_comment, last_edit_contributor.

Brown Corpus

We downloaded the Brown Corpus from the Python 3 package NLTK 3.5 (https://www.nltk.org/book/ch02.html).

Methods

Keyword Extraction

To extract key phrases from our WSM subcorpus we implemented the RAKE (Rapid Automatic Keyword Extraction) algorithm using the Python 3 package python-rake 1.5.0. On each text in our subcorpus as well as the Brown corpus we implemented the following steps:

1. Removed digits
2. Replaced newline characters with spaces
3. Implemented the RAKE algorithm using the following commands:

```
import RAKE
from nltk.corpus import stopwords
sw = stopwords.words('english')

Rake = RAKE.Rake(sw) #takes stopwords as list of strings

Rake.run(text, minCharacters = 3, maxWords = 5,
minFrequency = 2) #text is a string version of the document
```

On the resulting list of key phrases extracted from each text, we implemented the following steps:

1. Removed the following punctuation: `[ '/', ‘--', '*', '[', ‘]’ ]`
2. Removed `[‘mrs.’, ‘miss’, ‘mr.’]`
3. Removed leading hyphens (but kept hyphens between words)
4. Kept single word key phrases if they were at least 4 characters

Phase Matching in Wikipedia

To search for the extracted phrases in our Wikipedia data we used Elasticsearch, an open-source database based on the Apache Lucene library, found here: https://www.elastic.co/what-is/elasticsearch. We used the default text preprocessing pipeline, the standard text analyzer, which removes most punctuation and converts the text into lowercase. Elasticsearch’s preprocessing pipeline is based on Unicode specification: https://unicode.org/reports/tr29. After pre-processing, we broke each of the extracted phrases into multiple terms by splitting them on spaces and/or hyphens. We then did a multi-term search query (https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-span-multi-term-query.html) over all of the Wikipedia articles that we indexed in the Elasticsearch database, using the default value for fuzziness (https://www.elastic.co/guide/en/elasticsearch/reference/current/common-options.html#fuzzines). Specifically, if the term had less than three letters it required an exact match. If the term was between 3 and 5 letters we matched up to one edit distance. For terms more than 5 letters we allowed for two edit distances. We also allowed at least one intervening term appearing between the ordered terms coming from the phrase lists to ensure our search did not miss the mention of the phrases due to differences in stopwords or punctuation.

We employed this search pipeline on three sets of Wikipedia articles. First, we searched the full text across all Wikipedia articles in our data. Second, we searched across Wikipedia titles, using the metadata tag page_title. Third, we searched for phrases in articles with *movement*, *history*, or *feminism* or *feminist* in the title. Finally, we implemented the same pipeline on the Brown Corpus.