International Journal of Historical Archaeology
https://doi.org/10.1007/s10761-022-00680-5

# Text Mining Oral Histories in Historical Archaeology

**Madeline Brown**[1] · **Paul Shackel**[1]

Accepted: 12 December 2022


## Abstract

Advances in text mining and natural language processing methodologies have the potential to productively inform historical archaeology and oral history research. However, text mining methods are largely developed in the context of contemporary big data and publicly available texts, limiting the applicability of these tools in the context of historical and archaeological interpretation. Given the ability of text analysis to efficiently process and analyze large volumes of data, the potential for such tools to meaningfully inform historical archaeological research is significant, particularly for working with digitized data repositories or lengthy texts. Using oral histories recorded about a half-century ago from the anthracite coal mining region of Pennsylvania, USA, we discuss recent methodological developments in text analysis methodologies. We suggest future pathways to bridge the gap between generalized text mining methods and the particular needs of working with historical and place-based texts.

**Keywords** Coal Mining · Oral Histories · Text Mining · Natural Language Processing · Labor History

## Introduction

In historical archaeology, text mining oral histories and interviews is a relatively new approach for retrieving important contextual information related to the communities we work with and study. In conjunction with qualitative interpretation of texts typi-

✉ Madeline Brown
mtbrown@umd.edu

Paul Shackel
pshackel@umd.edu

[1] Department of Anthropology, University of Maryland, College Park, Maryland, USA

Published online: 13 January 2023

cally conducted in historical archaeology, text mining and natural language processing (NLP) hold the potential to enhance and complement these methods. Here we outline several key benefits of developing text mining approaches for working with oral history texts: (1) rapid and efficient analysis of large volumes of data, (2) reproducible workflows, (3) reducing the potential for observer bias, and (4) structured analysis of subgroup differences. We do not suggest that text mining replace traditional oral history methods used in historical archaeology, but rather be incorporated as part of a multi-modal toolkit. These methods are then applied in the context of coal mining oral histories from the anthracite region of Pennsylvania to demonstrate one pathway for incorporating mixed qualitative and quantitative approaches to oral history texts.

Natural language processing (NLP) allows large volumes of text to be analyzed based on their structure, meaning, and linguistic attributes. Common NLP and text mining approaches include word frequency and bigram analysis, parts-of-speech (POS) tagging, sentiment analysis, and topic modeling (Coden et al. 2005; Silge and Robinson 2017). However, many NLP methods assume the availability of large volumes of data and cannot account for historical differences in word usage, slang, or other linguistic factors that may interest historical archaeologists and other humanities scholars (McGillivray 2021). Consequently, developing data processing and analysis pipelines specifically tailored to the needs of historical archaeologists has the potential to transform how text mining is applied in the field.

Treating oral histories as texts presents both benefits and limitations for interpretation and analysis (Boyd and Larson 2014a). For example, converting recorded audio or video interviews into text can reduce some of the available contextual information from the interview (e.g., pauses, gestures, etc.). However, much can be learned from approaching oral histories as text when the researcher aims to compile large volumes of information or rapidly compare accounts of the same event across multiple sources (Boyd and Larson 2014a). Text analysis tools support searching with regular expressions, extracting text strings, identifying central topics or themes, and other interpretive steps. Searching with regular expressions allows researchers to identify instances of multiple forms of words or concepts (e.g., finding all occurrences of strings beginning with "archaeo" or "archeo" in the same search or any dates based on the pattern of numbers in a string). Additionally, text mining allows for rapid identification of the most frequently occurring words and groups of words across texts and the ability to compare texts that may differ on key variables (e.g., author, date, place of origin).

Contemporary research increasingly calls for transparency and public sharing of data and analysis workflows. Similarly, research reproducibility can be enhanced by publishing the code used to analyze data alongside articles and reports. Reproducing qualitative text interpretation depends heavily on a researcher's expertise and experience working in particular historical and cultural contexts. While there is no replacement for the context-specific domain expertise of a trained specialist, there is room for lowering the entry bar for the general public and nonspecialist researchers to engage with oral histories. Publishing not only primary datasets but also analytic datasets (e.g., lexicons, sentiment tags, and stopword lists), code, and analysis pipelines is one way to move toward this goal. This potential is demonstrated in several existing projects. For example, NLP is already applied to archaeological gray litera-

ture to extract metadata in the "Archaeotools" project (Jeffrey et al. 2008, 2009). This project demonstrates the utility of NLP tools for rapidly identifying the *what, where, when, who*, and type of *media* in archaeological records (Jeffrey et al., 2008). NLP and text mining toolkits have also been developed for the fields of classic literature and crisis management (CLTK 2022, CrisisLex 2022; Olteanu et al. 2014).

In addition to analyzing the implied or understood meaning of historical texts, text mining approaches can also be used to investigate linguistic structure and syntax. Notably, parts of speech (POS) tagging can be used to specifically examine verbs, nouns, or other parts of speech as they are used in a text. Amrani et al. (2008) describe a text mining workflow for PoS tagging and analyzing archeological texts, wherein specialized words that appear in archaeological texts are used to create PoS tagging lexicons. Expanding this method for the field might support creating a publicly available database of domain-specific lexicons for oral histories from different geographic and temporal contexts, which could improve access to insights and analyses based on these texts. Structured text analysis or "corpus linguistics" can identify insights from data that might remain unnoticed through qualitative analysis alone or which might be inadvertently overlooked (Sealey 2010). Archaeologists can also identify geographic places and events from historical texts (Hestia Project n.d.; Liceras-Garrido et al. 2019; Murrieta-Flores and Gregory 2015). As existing studies show, creating lexicons of places, people, and other key terms of analytic interest can improve the utility of text mining methods for working with domain- or field-specific texts (Coden et al. 2005; Liceras-Garrido et al. 2019).

Developing reproducible workflows for text mining oral histories has the potential to inform how community oral history projects are analyzed and communicated to the public. For example, if an oral history corpus includes perspectives from distinct social groups, then sentiment analysis and topic modeling might offer insight into how the same event or community is experienced differently by these groups. Because the meaning and sentiment of words change over time and in different cultural contexts, creating place-based and historical sentiment tagging lexicons may improve the ability to quantitatively analyze historical texts (Hamilton et al. 2016; n.d.; McGillivray 2021). Discipline-specific tools for text mining and natural language processing are being developed for classical languages (Classical Language Toolkit 2022) and may be meaningfully adapted for historical archaeology use cases in the future. Burns (2018) notes that although stopword lists exist for historic languages, there is a need to create new lists with contemporary coding methods and reproducible workflows in mind.

Applying text analysis methods to oral histories offers the ability to quickly identify key topics within the histories, including events, social or political issues, and distinctions in perceptions or memories across subgroups of interviewees. For example, Rieping (2022) presents a case study comparing results from topic models and summary keywords (e.g., from the transcription service Otter.ai) that were applied to texts from the MIT Black Oral History Project. With this approach, Rieping (2022) identified potential key topics within the interviews, such as the interviewee's goals, childhood, or advice (see table 3.6). This case demonstrates how text mining approaches can be meaningfully applied to large oral history projects to identify central narrative themes.

Text mining also offers tools for analysis across interviewees or text sources. This might enable the examination of differences in experiences, language, or other factors along dimensions of community and individual identity, such as gender, race, place of residence, or occupation. Silge (2017) demonstrates this potential in the context of gender by analyzing which words directly follow pronouns in a collection of nineteenth-century novels. Further, by including the code scripts used to generate this analysis, other researchers can build on these results using their own datasets (Silge 2017). In another study working with oral histories from the "Millennibrum" Project in Birmingham, UK, Sealey (2010) initially identified differences in word frequencies among male and female interviewees based on metadata demographic attributes and then delved deeper into additional variables that might contribute to how gendered language differences are expressed in these oral histories. In addition to examining differences with a priori demographics, topic modeling allows for examining emergent subgroups based on the text itself. For example, text analysis might reveal important distinctions along latent attributes related to identity, community membership, or life histories across oral history recordings.

Our ongoing work with the anthracite region oral histories explores the process of creating a custom lexicon or semantic tagging library to improve text interpretation. Creating context-specific dictionaries for tagging words may be more widely accessible than machine learning approaches, particularly for archaeologists with previous experience coding qualitative data. In this case study, we approach coal mining oral histories as text, and further, as texts that have the potential to be analyzed using contemporary data science approaches. Our goal is to use existing archived oral histories gathered 50 years ago and begin mining these texts to help identify significant trends in these acquired stories that may otherwise be overlooked. Mining text is an important methodology for historical archaeologists to help reconstruct the lives of communities and highlight portions of their lives that might otherwise be overlooked. The example of text mining presented here comes from a set of oral histories from the anthracite coal mining region of northeastern Pennsylvania, where the industry was dying, and communities remembered work and the struggle to survive during the industry's decline.

## Case Study Background: Anthracite Region of Pennsylvania, USA

Anthracite coal was discovered in northeastern Pennsylvania in the late 1760s, although large-scale extraction began in the 1820s, and the boom started in the 1840s. New transportation systems, such as railroads and canal systems, allowed for transporting large quantities of coal to the east coast. As a result, anthracite coal is often credited with igniting the industrial revolution in the United States, helping American industries become international manufacturing leaders (Palladino 2006).

The growing anthracite industry attracted a new immigrant workforce. The first coal miners to the anthracite region came from England, Wales, and Germany. By the 1840s and 1850s, they migrated from Ireland. In the 1880s, many coal mine owners, also known as coal operators, started recruiting workers from eastern and southern Europe. The newcomers were described as Polish, Slovak, Ruthenian, Ukrainian,

Hungarian, Italian, Russian, and Lithuanian (Blatz 2002: 27; Palladino 2006). As economic and political conditions in their home countries deteriorated because of widespread famine and dealing with oppressive feudal-like systems, many of these new immigrants were easily enticed to the coalfields by recruiters. Family members also participated in chain migration (Greene 1968: 25–26; Miller and Sharpless 1998: 172–173). This large-scale migration to the region created a ready workforce, although there were often more available workers than jobs. Surplus labor allowed the coal operators to keep wages relatively low with the threat that there were more available hungry men willing to move into the labor system (Roller 2015, 2018). The coal operators also believed that the new, ethnically diverse workforce would make the Irish coal workers' efforts to organize very difficult (Barendse 1981: 7–8, 24–28; Brooks 1898; Greene 1968; Miller and Sharpless 1998: 170–173; Roberts 1970 [1904]).

The newest immigrants to the region encountered the growing US national xenophobia (Dublin 1998; Roller 2015; Shackel 2023). The racialized inequalities the new immigrants faced were justified and seen as part of the natural order and justified through religion and ideology, empirical science, and formal science. With the increased in-migration from eastern and southern Europe during the late nineteenth century, social scientists developed evolutionary hierarchies. These new immigrants were placed below those of Western and Northern ancestry (Omni and Winant 1983: 51; Orser 2007: 9; Smedley 1998: 694). As a result of the scientific racism, the new immigrant miners received about 20% less pay than the "English speakers," those with ancestral roots in Western and Northern Europe. This pay inequity was justified by the coal operators because of the newcomers naturalized racialized status (Blatz 1994; Galtung 1990; Wallace 1987). Language became a significant structural barrier to advancement. In the late nineteenth century, a new Pennsylvania law required that miners take the mining license exam in the English language. This goal of this law was to prevent many of the foreign-born workers from advancing their skills and competing with the "English speaker." As a result, it kept their family income relatively lower than those who could take the English exam (Aurand 2003; Novak 1997). The *Report on Immigration* (US Senate 1911) has a section that ranks coal workers' ability for specific tasks based on nationality. It places the eastern and southern Europeans at the bottom of the scale of each occupation and they do not even qualify for supervisory or technical positions (Roller 2015b).

Because these new immigrants were not seen as equals, they more frequently faced extreme physical, nutritional, and mental hardships as they dealt with substandard housing, dangerous living and working conditions, and frequent encounters with undernourishment. They also faced harassment and verbal abuse from the established population, and their economic survival was always in jeopardy. Many of the miner workers were constantly in debt to the company store, a form of debt-peonage, which kept them living in substandard conditions (Daniels 1972; Ranson and Sutch 1977). Many miners saw this situation as slavery and earning slave wages. These conditions associated with poverty most likely have had a long-term effect on the general health and well-being of the population (Shackel 2016, 2019).

It was not until the Hazleton Area Strike in 1897 that the United Mine Workers of America recognized the value of incorporating the newest immigrants into the union.

While the miners were defeated in this strike, during the next major strike in 1900 with a more inclusive membership, the coal workers achieved some hard-earned benefits, such as higher wages and freedom from the company store. By World War I, about 180,000 coal workers extracted 100 million tons of anthracite coal annually (Shackel 2018). However, after the war, oil and natural gas gained a more significant market share, and the anthracite industry gradually began to decline (Rose 1981: 77). By 1922, coal production had dropped nearly 40%. While there was a slight uptick in coal extraction during World War II, the industry further declined in the 1950s. Many men were unemployed or "gone to New Jersey" for work (Shackel 2023; Wolensky 2020). Towns were depopulated, and some of the smaller patch towns were abandoned. Today, only a few hundred people work in the anthracite coal industry in northeastern Pennsylvania. The region continues to struggle to find an economic engine to replace coal. Instead, low-skilled work is abundant in the newly developed fulfillment centers. Unemployment remains relatively high, and opioid addiction continues to rise, afflicting a community that feels forgotten (Bradlee 2018; Shackel 2023; Silva 2019).

## Methods

We analyzed transcripts from the Scranton Oral History Project for this text mining case study. In the 1970s, historians and folklorists focused on collecting oral histories of the men and women who worked in and around the vanishing anthracite coal industry in northeastern Pennsylvania. These interviews were collected by the Pennsylvania Historical and Museum Commission and are now on file at the Pennsylvania State archives. In addition, we have made the converted plain text files of the interview transcripts available online in the GitHub repository associated with this study (Brown and Shackel 2022). The Scranton interviews contain 26 transcripts analyzed in their entirety. The interviews generally ranged from 30 to 45 min each, ranging from 3,000 to 5,000 words. Many of the narrators reflect on their life experiences in the coal region from the beginning of the twentieth century. Many of these accounts detail experiences during the industry's collapse after World War I. Our research team brings together distinct subject expertise sets: Shackel has extensive experience working with oral histories from the anthracite region of Pennsylvania, while Brown has experience working with text analysis across numerous domains.

As part of our goal of increasing reproducible research in historical archaeology, we describe the data processing and analysis steps in detail and make a reproducible example of code used for this analysis available online (Brown and Shackel 2022). The Scranton interviews are archived by the Pennsylvania State Archives as pdf files. To work with these interviews, we first converted the files into text (by opening the files with a word processor) and standardized them in terms of font, format, and text size. Next, we removed any extra characters that appeared as a result of the file conversion process and added line breaks whenever the speaker in the transcript changed. These files were then saved as plain text files and imported into R for analysis (R Core Team 2022).

The first step of data wrangling was to remove the first few lines of the transcripts, which contained metadata (such as the setting and a description of the interview) rather than the primary interview data. Texts were then standardized to remove numbers and apostrophes and convert all terms to lowercase (this avoids counting "coal" and "Coal" as separate terms). In text mining, individual terms or words are often referred to as tokens. The next step in the data wrangling pipeline is to unnest tokens into individual and pairs of words (bigrams) and remove stopwords using tidytext's stop_words lexicons: onix, SMART, and snowball (Silge 2016). Stopwords are common words that appear in the text, such as *the*, *it*, or *as*, which do not add additional interpretability to the content of a text. These words may be useful for some linguistic analyses. However, in this case, we are most interested in the broader themes and sentiments discussed in the interviews rather than the structure of language, so we removed common stopwords. After these general text processing steps, we further removed certain slang terms, names, or common words that did not influence the meaning of the text (e.g., *interviewed, huh*, or *yea*). These stopwords are listed in the online GitHub repository associated with this study. At this point, the text data are considered cleaned, tidy, and ready for further analysis (Silge and Robinson 2017).

The process of removing terms from oral history texts involves both automated and subjective steps. An analysis might involve systematically excluding numbers from the analysis, while also deciding to remove particular filler words or abbreviations on a case-by-case basis. For example, in the Scranton oral history transcripts, we noted that the interviewer's and narrator's names appeared frequently. However, in our analysis, these names add minimal insight into the meaning of the texts and thus were removed from the data. Any time data are removed from a raw dataset, there is potential for introducing subjective or systematic bias. We address these concerns by describing the data processing pipeline, publishing our stopword list, and sharing our data analysis code.

Text mining approaches include a wide range of methods, from frequency analysis and natural language processing to network analysis and topic modeling. In this paper, we focus on word frequency and n-grams, while also discussing the future potential of sentiment analysis, and tagged lexicon analysis (Silge and Robinson 2017). Data analysis was conducted with R (R Core Team 2022), including the following packages: tidyverse (Wickham et al. 2019), textclean (Rinker 2018), tidytext (Silge 2016), and cleanNLP (Arnold 2017).

## Challenges of Working with Historical and Place-based Texts

In analyzing oral histories using new and emerging methods, challenges may arise related to the incompatibility of archival methods and contemporary analysis workflows. In this paper, the original raw data archives were stored as pdf files, which had to first be converted to plain text files before they could be analyzed. Moreover, once the transcripts were converted to plain text, it was apparent that various irregularities in the text format existed. For example, not all lines began with standardized names (e.g., sometimes the interviewer or interviewee was given a title [e.g., Mrs.], and other times they were not). These irregularities may be idiosyncratic to particu-

lar datasets. However, they will be critical to document and develop strategies for addressing if historical oral history transcripts are to be easily incorporated into data science workflows.

Text tagging and structured documents can make rapidly subsetting and searching texts more efficient. For example, the Cambridge Greek Lexicon uses xml tagging to annotate definitions in a standardized and easily searchable way (Faculty of Classics 2022). This methodology might be extended and applied for streamlined natural language processing of other texts of interest to historical archaeologists. In the case of oral history transcripts, using html tags or simply the same string pattern each time an individual begins speaking would allow a rapid selection of interview sections. Depending on the research question, it might be important to differentiate between words spoken by the interviewer and interviewee in an oral history transcript. Differentiating sections of unstructured text remains a challenge for working with archival data, but one which may be addressed with natural language processing tools.

Adding to the complexity of working with transcripts, several different transcription methodologies exist that enable a text-based recording of pauses, verbal emphasis, tone, and other information that might be lost if only the words are recorded (see Britt 2018). However, such annotated transcripts offer additional challenges for text mining oral history and warrant additional development of standardized methodologies and coding packages to efficiently process these data for analysis.

Text data often contains a lot of noise that may or may not be pertinent for answering particular questions. For example, oral history transcripts may contain the speakers' names – interviewer and interviewee – before each line. Consequently, analyzing the most frequently occurring words in the original transcripts might lead to the names of the participants being at the top of the list. For the purposes of text analysis, these terms can be considered *stopwords* and included in a custom stopword list for this text. The custom stopword list can then be excluded from the analysis. In addition to the names of participants, working with historical and place-based texts adds an additional layer of complexity in determining which words to include and exclude from the analysis. Standard stopword lists do not always account for regional or temporal differences in speech nor contain slang or abbreviations. For the anthracite region oral histories, numerous abbreviations, contractions, and names were excluded from text analysis.

## Results

The primary aim of this paper is to reanalyze existing oral history documents using new analytic tools. We frame our analytical methods using the "tidy text mining" approach described by Silge and Robinson (2017), wherein text data are cleaned and organized for systematic quantitative analysis. Here, we examine word frequencies and n-grams (single terms and groups of consecutive terms). We conclude by discussing the potential of these approaches to inform oral history text mining more generally.

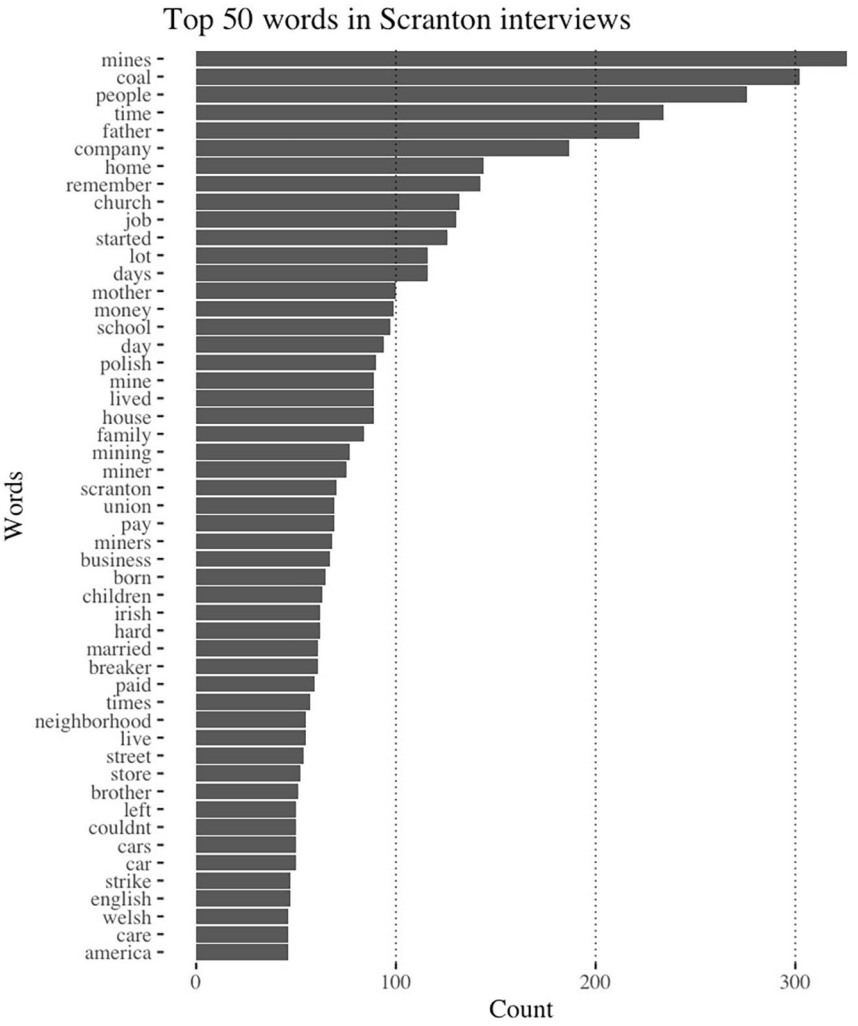


**Fig. 1** Most frequently occuring words in the Scranton Oral histories. This list does not include stopwords

## Which Words are Used Most Frequently?

The most common words appearing in oral histories can reveal insights into the types of questions asked by the interviewer and the topics seen as important by community members interviewed. Commonly agreed-upon stories in a community, a collective memory of the past, become apparent after text mining interviews. Figure 1 lists the words appearing most frequently in the 26 Scranton interviews. The most common word used is "mines," and there is also an emphasis on family with discussions that include terms like "father," "mother," "family," "born," "married," and "children." There is also some evidence of the ethnic tensions and identities in the community, as

interviewees emphasized ethnic affiliations, such as "Polish," "Irish," and "English." These terms indicated a heightened awareness of collective identity and ethnic differences. Ethnic affiliation terms were used in reference and in comparison to each other to distinguish differences and rank. For instance, Stanley Guntack (interview, 1973), a second-generation worker of Polish descent, noted that his supervisors were Irish, and he felt that he was not treated fairly regarding work assignments and promotions. "Well, I'll tell you, they treated you all right when you were going for them, and to get something better you were always last. 'The next time we'll give it to you.' That's how they treated you." Also, John Parroccini, who worked in a Scranton mine, noted, "Well, years ago and sometime really you had to watch out, going out during the night. Different nationalities might attack you. But, I never had any trouble" (Parroccini interview, 1973). Regarding Slavic immigration, "Well the period was 1881, when my father was in Freeland. …My mother used to tell me that when it was time to go to work they would start gathering on their porch to go to work as a group because they were afraid that they would be attacked by someone. But when they traveled in a group, they felt protected. … In 1888 there was still trouble. The people didn't want the new immigrants coming in and taking their jobs. And it soon became a sport to hurt these people" (Bosak interview, 1973). Identifying the frequency of ethnic identifier terms in the oral histories helps direct subsequent qualitative analysis of these texts.

The terms *coal, people*, and *time* appear in the top five words. The use of the term coal is obvious since the interviews focused on life in the anthracite coal region, and questions tended to focus on the industry and the impact of this industry on workers and families. In some ways, "coal" might be considered a stopword for this text since it is the primary topic of the interviews. However, given the frequency of this term and our interest in disambiguating how *coal* is discussed, we further examine bigrams associated with coal later in the results. The word "people" is often used to describe a particular subset of people, such as, "the Polish are very religious people" (Serafin interview, 1973) and, "my mother worked for Jewish people" (Slavetskas interview, 1973). This again highlights the importance of ethnicity and immigrant origins in this region, collective identity, and ethnic differences. The term "time" is often used in reference to relationships, a distant event, or the length of an event. For instance, "But there was starvation, and it all depended on the strikes. 3 months is a long time. There was no aid from anyone" (Harding, interview, 1973).

### Which Pairs of Words Appear Together?

In addition to examining the top individual words in the oral histories, we also examined the most frequent pairs of words (bigrams). Pairs of words can yield interesting insights into multiword concepts, place names, or important events discussed in interviews. In these bigrams, coal companies and company stores appear at the top of the list (Table 1). We also see the appearance of the names of specific coal companies and mines, such as *Susquehanna Coal*, and the mention of the union - *United Mine (Workers).* These bigrams can be divided into several categories: (1) places, (2) identity markers, (3) events, and (4) coal mining experience.

**Table 1** Top pairs of words (bigrams) in Scranton oral histories

| Top 50 bigrams in Scranton interviews | | |
|---|---|---|
| Bigram | | Count |
| coal | company | 66 |
| mine | workers | 20 |
| world | war | 19 |
| coal | companies | 17 |
| Forest | City | 15 |
| Iron | Police | 15 |
| Glen | Alden | 13 |
| Hudson | Coal | 13 |
| United | Mine | 11 |
| Catholic | Church | 10 |
| company | store | 10 |
| chain | stores | 9 |
| Molly | Maguires[a] | 9 |
| St | Josephs | 9 |
| boarding | houses | 8 |
| grocery | store | 8 |
| soft | coal | 8 |
| boarding | house | 7 |
| company | stores | 7 |
| Lehigh | Valley | 7 |
| mining | industry | 7 |
| Polish | national | 7 |
| Scranton | Coal | 7 |
| section | forman | 7 |
| black | lung | 6 |
| breaker | boy | 6 |
| coal | mines | 6 |
| company | owned | 6 |
| hard | times | 6 |
| holy | trinity | 6 |
| national | church | 6 |
| Penn | anthracite | 6 |
| Rhode | Island | 6 |
| Roman | Catholic | 6 |
| street | car | 6 |
| Alden | Coal | 5 |
| Anthracite | Region | 5 |
| butcher | shop | 5 |
| contract | miner | 5 |
| due | bill | 5 |
| hundred | dollars | 5 |
| ice | truck | 5 |
| main | street | 5 |
| mining | company | 5 |
| Molly | Maguires[a] | 5 |
| polish | people | 5 |

**Table 1** (continued)

| Top 50 bigrams in Scranton interviews | | |
|---|---|---|
| Bigram | | Count |
| political | figures | 5 |
| Port | Chester | 5 |
| Providence | Rhode | 5 |
| religious | holidays | 5 |
| section | foreman | 5 |
| shaker | shoot | 5 |
| sheet | iron | 5 |
| silk | mills | 5 |
| speak | english | 5 |
| store | keeper | 5 |

Note: aMaguires was originally transcribed as Mcguires and as Mcguries

When we analyzed trigrams (three sequential terms), the top phrase found was "in the mines" – appearing 207 times – an order of magnitude higher than the next most common trigrams ("a lot of" [n=94], "there was a" [n=94], "did you have" [n=74]). This indicates that many of the sentences in the interviews are directly related to things that would have occurred or been experienced "in the mines." This trigram can be used to rapidly subset text to those experiences that occurred in mines. For instance, Tom Price, second-generation Welsh, noted, "You had to be Welsh to be a foreman in the mines at that time." (referring to the early twentieth century in the Scranton area) (Price, interview 1973). He also explained, "I had a brother that worked in the mines that started when he was 8 years old, his pail used to drag on the floor" (Price, interview 1973). Thomas Handing, second-generation Irish, exclaimed, "And most of the men that worked in the mines in those days worked 11 hours" (Handing, interview 1973). He also spoke about his family. "My older brothers were in the mines when they were 10 and 11" (Handing, interview 1973). While not all trigrams may be useful, subsetting based on trigrams can allow for rapid identification of sections of the interviews related to topics of interest that can then be further analyzed qualitatively.

## Words Associated with Coal

In the Scranton interviews, the term "coal" appears 319 times. To investigate patterns related to how the term "coal" is used in the oral histories, we extracted bigrams using this term. Looking at bigrams without stopwords that include "coal" shows the compound words that use coal and the verbs or adjectives used in conjunction with this term. The frequent use of the word "coal" is not surprising since these interviews focused on people's experiences with the coal industry. What is noteworthy are the particular patterns in the words linked to coal. We find that bigrams, including the word coal, overwhelmingly are characterized by terms about coal companies, both generally and referring to particular company names, as well as the terms *mines* and *mining*, discussing the mines in both verb and noun forms (Table 2).

The words most frequently adjacent to coal tend to be associated with the companies. For instance, a typical question for each interviewee was, "What coal company did your work for?" Furthermore, in this case, Stanley Guntack, who was described

**Table 2** Bigrams associated with "coal" appearing at least twice in Scranton Oral histories

| Coal Bigram | | Count | Coal Bigram | | Count |
|---|---|---|---|---|---|
| coal | company | 66 | coal | hole | 2 |
| coal | companies | 17 | coal | industry | 2 |
| Hudson | Coal | 13 | coal | miners | 2 |
| soft | coal | 8 | coal | people | 2 |
| Scranton | Coal | 7 | coal | town | 2 |
| coal | mines | 6 | hauled | coal | 2 |
| Alden | Coal | 5 | hauling | coal | 2 |
| coal | mining | 4 | loading | coal | 2 |
| Susquehanna[a] | Coal | 4 | lower | coal | 2 |
| Valley | Coal | 4 | mine | coal | 2 |
| coal | business | 3 | mining | coal | 2 |
| Moffitt | Coal | 3 | reading | coal | 2 |
| PA | Coal | 3 | | | |

Note: [a]Both 'Susquehanna' and 'Susquehannah' appeared in the original transcript. Here they are combined

as "a second generation Pole from Dickson City," answered, "The Hudson Coal Company" (Guntack, interview, 1973). In another case, Earl W. Lamb, who rose to the position of president of the Moffitt Coal Company, spoke about his experience and training as an engineer. He noted that soon after becoming an engineer, he left the Scranton Coal Company and moved to the Hudson Coal Company. He described the technological innovations in coal extraction, moving from primarily manual labor to using more technology and heavy machinery. He noted that in 1924, "Hudson [Coal] was quite advanced in those days, and mechanization was just a period when mechanization was looked on as a thing of what was going on" (Lamb, interview 1973).

The high frequency of bigrams associating coal with a company is a by-product of the interviewer asking questions about the company and the interviewee's career within the company that employed the former coal worker. It may also be a reflection of the importance that the interviewee placed on these companies. In the 1970s, these companies were going bankrupt and closing throughout the region, and they served as a reminder of the last good wage coal workers earned before having to find employment outside of the region.

Common, although less frequent bigrams, or second-tier bigrams, associated with the term "coal," includes an emphasis on the individual worker. For instance, the term "coal miners" was often used to describe individual work. One former miner explained, "there were runners and drivers and motor runners, coal miners and laborers. And it wasn't easy, and in those days it was hand mining. There wasn't as much dust in the mines today as there was before. It was hard work and it took the best out of you" (Davis, interview 1973).

These second-tier bigrams also include the type of work performed by the individual, "hauled coal," "hauling coal," and "loading coal" coal. As well as "mined coal" and "mining coal." Hauling coal is often referred to as individuals bringing the coal to markets. In reference to mining coal, Ben Grevera (interview 1973) spoke about how his father worked for Susquehanna Coal Company for 53 years. At the end of his career, when suffering from Black Lung Disease, he asked for a lighter job. The company responded. "They told him that they had paid him for what he had done. My old dad had to quit work. Later on, I went into coal mines. I mined coal and I mined in the gangways" (Grevera, interview, 1973).

Examining the frequently occurring bigrams, including the term "coal," reveals clear patterns across the oral histories. In this case, there is a strong emphasis on referring to specific coal companies, like Hudson Coal, Scranton Coal, Alden Coal, Susquehanna Coal, and Valley Coal. The focus on people and individual work plays a secondary, although important, role in understanding these interviews. The coal-bigram analysis notably highlights the two-word concepts that frequently appear in the histories, which may not be extracted by analyzing only single words. Pairing the quantitative identification of frequent bigrams with a close reading of these contexts for these terms enables a more detailed discussion of how these terms are used in context as key referents in the oral history narratives.

## Discussion

Analyzing word frequencies and n-grams in the anthracite region oral histories adds a new dimension of interpretation to these narratives. Though a seemingly simple method, word frequency analysis can be an important first step in examining the themes and content of an oral history corpus. This method can reveal the most frequently occurring terms and topics, which may or may not correspond to a researcher's initial or qualitative impression of important themes. This type of corpus-level analysis becomes powerful for guiding the close reading and interpretation of individual texts while keeping the overall content of the collection of texts in mind.

We examined the potential for using natural language processing tools to wrangle oral history data and conduct term and n-gram frequency analyses. These analyses open the conversation about text mining in oral history, yet many other potentially productive applications exist. For example, topic modeling remains an area of potential innovation in oral history analysis, as it enables large quantities of text to be sorted into themes or *topics* for further analysis (Silge and Robinson 2017).

Text mining oral histories and interviews provides an additional avenue for understanding how a community views and understands its past. Working with aggregated texts from a particular social and historical context allows interpretation to follow multiple nested scales, from individual memory to the collective experience of a place or event. In this case, 26 interviews acquired in 1973 and transcribed in 1981 were re-examined using contemporary text mining tools. General patterns across the interviews are identified that both speak to common themes across interviewees related to family and coal companies, as well as individual experiences related to these themes and their experiences as immigrants or in relation to particular ethnic identities.

What communities remember and how we remember are important issues that allow us to see how public memory develops. In this case, the community consists of people connected to the anthracite coal industry in northeastern Pennsylvania. For the narrators, some of the most frequently used terms in the Scranton interviews are those related to family, such as "father," "mother," "family," "born," "married," and "children." These commonly used terms highlight the importance of family in the remembered daily lives of those living near and working in the coal mines.

As Sealey notes, text mining oral histories can reveal "linguistic patterns" and "patterns associated with the speakers' membership of various sub-categories"

(Sealey 2010: 1). Many of these interviews contained terms defining ethnic affiliation, which were used to describe the inter-group tensions in the community and at work. The bigrams and trigrams hint at ethnic tensions and identities as important facets of life and work around the mines. As Sealey (2010:1) observed in a different oral history study: "Each interviewee demonstrates the ever-present potential for linguistic creativity while simultaneously contributing to the collective entity that emerges as 'the discourse of life histories'" (Sealey 2010: 1).

We acknowledge that while there can be competing interests to control the collective meaning of a place, there can also be subordinated views that might not be represented in the dominant public memory. This is where individual interviews and oral histories can offer counter-narratives or distinct memories of a particular place and time. However, this exercise in text mining 26 Scranton oral histories expands the interpretation of this community and focuses on what the narrators believe is important to convey about their past. Text mining identifies what appear to be significant memories, such as the bigrams including the word "coal" that identify the many different coal companies in the region. These frequent references to coal mining companies may reflect their significance in the experiences of those living in coal mining communities. Text mining can also identify what might be a subordinate memory in the community, such as the bigrams – including the terms adjacent to the word "coal" – which identify tasks associated with the individual.

While text analysis of coal mining oral histories reveals themes related to family, ethnicity, and individual labor experiences, words associated with *coal* also tend to connect to specific companies, thereby situating these interviews in the particular capitalist landscape associated with this region and time period. Text mining and natural language processing approaches show promise for working with oral history texts of interest to historical archaeologists. Such methods allow researchers to efficiently and rapidly analyze lengthy collections of text, engage in reproducible research, reduce interpretation biases, and identify subgroup and latent patterns in texts. The end result is helping to create a better context for understanding the communities we study.

**Acknowledgments** The Scranton Oral History Project is transcribed and is on file at the Pennsylvania State Archives in Harrisburg, Pennsylvania. We are grateful to Aaron McWilliams and Suzanne Stasiulatis for providing guidance and access to this transcribed collection. Through the era of Covid-19 and the closing of the state archives, they were more than willing to locate and provide access to these oral histories. These data can also be found at the Anthracite Oral Histories Analysis Repository (URL: https://github.com/maddiebrown/OralHistories). This repository contains the text files for interview transcriptions from the Scranton Oral History Project. We also thank Catherine Gagnon for assistance with converting archival files into text.

## References

Amrani, A., Abajian, V., Kodratoff, Y., and Matte-Tailliez, O. (2008). A chain of text-mining to extract information in archaeology. In *3rd International Conference on Information and Communication Technologies: From Theory to Applications.* IEEE, Piscataway, NJ, pp. 1–5..

Arnold, T. (2017). A tidy data model for natural language processing using cleanNLP. *R Journal* **9**(2): 1–20. https://journal.r-project.org/archive/2017/RJ-2017-035/index.html.

Barendse, M. A. (1981). *Social Expectations and Perception: The Case of the Slavic Anthracite Workers*. Pennsylvania State University Press, University Park.

Blatz, P. K. (2002). Reflections on Lattimer: a complex and significant event. *Pennsylvania History* **69**(1): 42–51.

Boyd, D. A. and Larson, M. A. (2014). Introduction. In Boyd, D. A. and Larson, M. A. (eds.), *Oral History and Digital Humanities*. Palgrave Macmillan, New York, pp. 1–16.

Bradlee, B. Jr. (2018). *The Forgotten: How the People of One Pennsylvania County Elected Donald Trump and Changed America*. Little, Brown, Boston.

Britt, E. (2018). Oral history and the discursive construction of identity in Flint, Michigan. *Journal of Linguistic Anthropology* **28**(3): 252–272.

Brooks, J. G. (1898). Notes. *Yale Review* **6**: 306 – 311.

Brown, M. and Shackel, P. (2022). Anthracite Oral Histories Analysis Repository. http://github.com/maddiebrown/OralHistories; accessed November 2022.

Burns, P. J. (2018). Constructing stoplists for historical languages. *Digital Classics Online* **4**(2): 4–20.

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., and Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics* **38**(6): 422–430. https://doi.org/10.1016/j.jbi.2005.02.009.

Classical Language Toolkit. (2022). http://cltk.org/; accessed August 2022.

CrisisLex. (2022). CrisisLex.org. https://crisislex.org/; accessed August, 2022.

Dublin, T. (1998). *When The Mines Closed: Stories of Struggles in Hard Times*. Cornell University Press, Ithaca, NY.

Faculty of Classics. (2022). Tagging the Lexicon. https://www.classics.cam.ac.uk/research/projects/glp/tagging; accessed August 2022.

Greene, V. R. (1968). *The Slavic Community on Strike: Immigrant Labor in Pennsylvania Anthracite*. University of Notre Dame Press, Notre Dame, IN.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 595–605. https://doi.org/10.18653/v1/D16-1057.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (n.d.). SocialSent: domain-specific sentiment lexicons for computational social science. https://nlp.stanford.edu/projects/socialsent/; accessed August 2022.

Hestia Project. (n.d.). https://hestia.open.ac.uk/; accessed August 2022.

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S., and Zhang, Z. (2009). The archaeotools project: faceted classification and natural language processing in an archaeological context. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1897): 2507–2519. https://doi.org/10.1098/rsta.2009.0038.

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S., and Zhang, Z. (2008). When ontology and reality collide: the Archaeotools Project, faceted classification and natural language processing in an archaeological context. *Proceedings of the 36th CAA Conference*, pp. 285–290.

Liceras-Garrido, R., Favila-Vázquez, M., Bellamy, K., Murrieta-Flores, P., Jiménez-Badillo, D., and Martins, B. (2019). Digital approaches to historical archaeology: exploring the geographies of 16th century New Spain. *Open Access Journal of Archaeology and Anthropology* **2**(1). https://doi.org/10.33552/OAJAA.2019.02.000526.

McGillivray, B. (2021). Computational methods for semantic analysis of historical texts. In Schuster, K. and Dunn, S. (eds.), *Routledge International Handbook of Research Methods in Digital Humanities*. Routledge, London, pp. 261–274.

Miller, D. L. and Sharpless, R. E. (1998). *The Kingdom of Coal: Work, Enterprise, and Ethnic Communities in the Mine Fields*. University of Pennsylvania Press, Philadelphia, PA.

Murrieta-Flores, P. and Gregory, I. (2015). Further frontiers in GIS: extending spatial analysis to textual sources in archaeology. *Open Archaeology* **1**(1). https://doi.org/10.1515/opar-2015-0010.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: a lexicon for collecting and filtering microblogged communications in crises. *In Eighth International AAAI Conference on Weblogs and Social Media.*

Omi, M. and Winant, H. (1983). By the river of Babylon: race in the United States. *Socialist Review* **13**: 31–65.

Orser, C. E., Jr. (2007). *The Archaeology of Race and Racialization in Historic America*. University Press of Florida, Gainesville, FL.

Palladino, G. (2006). *Another Civil War: Labor, Capital, and the State in the Anthracite Regions of Pennsylvania, 1840–1868*. Fordham University Press, New York.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rieping, H. A. (2022). *Audio Segmenting and Natural Language Processing in Oral History Archiving*. Master's dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Rinker, T. W. (2018). *Textclean: Text Cleaning Tools Version 0.9.3*. Buffalo, New York. https://github.com/trinker/textclean
Roberts, P. (1970 [1904]). *Anthracite Coal Communities*. Arno Press and the New York Times, New York.
Roller, M. P. (2015). Diabolical consumerism: mass psychology and social production between the Gilded and the Golden Ages. In Leone, M. and Jocelyn K. (eds.), *Historical Archaeologies of Capitalism*. Springer, New York, pp. 25–34.
Roller, M. P. (2018). *An Archaeology of Structural Violence: Life in a Twentieth-Century Coal Town*. University Press of Florida, Gainesville.
Rose, D. (1981). *Energy Transition and the Local Community: A Theory of Society Applied to Hazleton, Pennsylvania*. University of Pennsylvania Press, Philadelphia, PA.
Sealey, A. (2010). Probabilities and surprises: a realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics* **31**(2): 215–235. https://doi.org/10.1093/applin/amp023.
Shackel, P. A. (2016). The meaning of place in the anthracite region of northeastern Pennsylvania. *International Journal of Heritage Studies* **22**(3):200–213.
Shackel, P. A. (2018). *Remembering Lattimer: Migration, Labor, and Race in Pennsylvania Anthracite Country*. University of Illinois Press, Champaign.
Shackel, P. A. (2019). Structural violence and the industrial landscape. *International Journal of Heritage Studies* **25**(7): 750–762. https://doi.org/10.1080/13527258.2018.1517374.
Shackel, P. A. (2023). *The Ruined Anthracite: Historical Trauma in Coal Mining Labor Communities*. University of Illinois Press, Champaign.
Silge, J. and Robinson, D. (2016). Tidytext: text mining and analysis using tidy data principles in R. *JOSS* **1**(3). https://doi.org/10.21105/joss.00037.
Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Sebastopol, CA.
Silge, J. (2017). Gender roles with text mining and N-grams. https://juliasilge.com/blog/gender-pronouns/; accessed August 2022.
Silva, J. M. (2019). *We're Still Here: Pain and Politics in the Heart of America*. Oxford University Press, Oxford.
Smedley, A. (1998). Race and the construction of human identity. *American Anthropologist* **100**(3): 690–702.
US Senate. (1911). *Reports of the Immigration Commission, Vol. 16: Immigrants in Industries*. Government Printing Office, Washington, DC.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43): 1686. https://doi.org/10.21105/joss.01686.
Wolensky, R. (ed.) (2020). *Sewn in Coal: An Oral History of the Ladies' Garment Industry in Northeastern Pennsylvania, 1945–1995*. Pennsylvania State University Press, University Park, PA.