Using Natural Language Processing to Organize and Analyze Oral History Projects

by Christopher M. Pandza

A thesis submitted to the
faculty of Columbia University
In partial fulfillment of the requirements for the
Degree of Master of Arts in Oral History

New York, New York

May 2023

Contents

Introduction

Over the course of a typical 90-minute oral history interview, a person speaking at 150 words per minute can generate 13,500 words of unstructured or semi-structured language. Transcribed, such an interview might be represented by 54 double-spaced pages. At this rate, even a modestly sized project with 20 interviews might produce 270,000 words—almost five novels worth of text. But unlike novels, oral history interviews are not neatly structured into chapters but rather follow the idiosyncratic richness of human memory recall and storytelling.

These rich documents offer insight not only into past events but how events are remembered (or forgotten), how narrators relate to broader historical and social forces, and how narrators make sense of their lives in totality. In addition, oral history interviews—often to the embarrassment of interviewers—can be delightfully transparent in revealing the conditions of their own production. Oral history interviews make partially visible how interviewers and narrators navigate co-authorship in relation to social, cultural, and technological bounds, as well as to each other.

Due to an interview transcript's high text volume and loose structure, bringing a large number of interviews into conversation with each other or systematically analyzing them as a whole is a difficult task for any one researcher. A researcher attempting to analyze a collection might manually code interviews or rely on a team to code interviews in a systematic way. However, associated costs and inter-coder reliability issues make this type of analysis challenging, if not impossible, through conventional means.

However, this challenge doesn't stop researchers from approaching systematic analysis. For example, in "Editing Procedures in Studs Terkel's Oral Histories", Lucie Kučerová analyzes Studs Terkel's editing process by comparing four original transcripts with Terkel's edited essays.[1] In

"Golden Door Voices: Towards a Critique of the Ellis Island Oral History Project", Mario Varricchio attempts to understand cultural biases in the Ellis Island Oral History Project by manually coding and analyzing 30 of the collection's nearly 2,000 interviews.[2] Varrichio concludes, as many oral history researchers do, that conclusive proof of his work could only come from an out-of-scope systematic analysis of the entire corpus.[i]

Outside of oral history, other fields are approaching challenges related to unstructured human language using natural language processing (NLP). Natural language processing is an artificial intelligence (AI) field that deals with understanding and generating human language. NLP can be thought of as an umbrella for many distinct tools that interface between human language and machines. Businesses, for instance, use NLP tools to transform vast pools of unstructured customer data into actionable insights. A customer service department might use NLP to quickly sort and assign customer complaints to different service centers. Social scientists use NLP to measure sociocultural phenomena and developments in discourse. For example, in "The Semantic Structure of the Conformist and Dissenting Bible, 1660-1780", the authors used a suite of NLP tools to map the Protestant Bible's semantic structure against different sects' uses of the Bible in sermons.[3] Researchers in the humanities, who are perhaps epistemologically similar to oral historians, are also finding use cases for NLP. For example, in "Improving Measures of Text Reuse in English Poetry: A TF–IDF Based Method", the authors examine the influence that English romantic poets had on Irish poet William Butler Yeats by measuring text reuse.[4] NLP has also been adopted by artists in innovative ways. For example, Nisga'a poet Jordan Abel uses NLP to create poetry that challenges problematic representations of Indigenous peoples by remixing works like *The Last of the Mohicans.*[5] Even if you have never encountered the term NLP, it's likely that you

[i] Corpus (plural corpora) is a term commonly used in linguistics to describe a collection of texts.

are already using NLP. NLP powers virtual assistants like Apple's Siri, translators like Google Translate, text message autocomplete tools, and many other consumer applications.

Our discipline is well-positioned to benefit from NLP methods given the nature of the transcripts we produce and handle. One might imagine that there is unbridled enthusiasm around using NLP in the discipline—perhaps even enthusiasm that needs curtailing. However, a cursory search for the most popular text analysis tools in *The Oral History Review* yields only passing references to these methods. In my search for practitioners in the field processing text with machines, I've only found a handful of examples: *Let Them Speak*, an interactive text indexing project at Yale Library's Fortunoff Video Archive for Holocaust Testimonies,[6] and *N'TOO*, a chatbot art installation by Stephanie Dinkins.[7] More recently, in "Audio Segmenting and Natural Language Processing in Oral History Archiving", Holly Anne Rieping experiments with NLP tools that segment interviews and create metadata.[8] The two other NLP-enabled oral history projects I've encountered are ones I have a direct hand in producing: the Obama Presidency Oral History project's digital archive, and my own work with AI as part of my studies at Columbia.

Why has oral history not been so interested in using NLP? Part of the reason might be that there is a dearth of oral historians with quantitative backgrounds—many oral historians come from qualitative training in the humanities, anthropology, and the arts. This explanation, however, fails to explain the explosion of NLP in the arts and humanities.[9]

Perhaps NLP's speed and distance-reading feel inherently antithetical to oral history's epistemological and ethical tenets. Oral historians conceptualize their work as slow work—media that is produced and consumed through deep, close listening. As Sheftel and Zembryzki note, "relationship building, interviewing, and careful analysis, all of which are at the core of oral history,

take time… slowness allows us to consider the impact, context, trajectory, and implications of our work."[10] Interview events are embedded in relationships, ethics, commitments, and responsibilities. Treating narrations as machine inputs appears to be in conflict with oral history's ethical and epistemological tenets.

But is NLP inherently incompatible with the tenets that define oral history? Without considering NLP in relation to oral history practice, our field risks neglecting a set of methodologies that could create new opportunities for practitioners to organize, analyze, and otherwise enable their work. Engaging directly with these tools will help us understand if, when, and how there are benefits to incorporating these tools into our practice proactively and on our own terms. If oral historians are to adopt NLP-based tools into our discipline, the following issues must be approached:

1. What tools are available, and how do they work?
2. How do these tools create value for oral historians, and what are their limitations?
3. How do these tools interact with oral history's epistemological and ethical frameworks?

In this paper, I begin to answer these questions by using several NLP tools and techniques to organize and analyze the Ellis Island Oral History project. My two main areas of demonstration will be how NLP can be used to organize a large archive and how NLP can be used to analyze the conversations that compose a large archive. I will not endeavor to catalog the entire universe of NLP tools at an oral historian's disposal or exhaust the epistemological and ethical questions that these tools raise. Instead, in order to advance an overdue discussion in our field, I will demonstrate, using a large oral history project, a number of ways that oral historians can think about using NLP in service of their practice.

When looking for a corpus to explore, my two main criteria were scale and personal interest. Scale is relevant to NLP work as some tools perform better with larger corpora; I set a threshold of at least 200 interviews for my analysis. My second criterion was personal interest; the corpus should cover a topic area that I find energizing. I found the perfect candidate for this work in the Ellis Island Oral History project, which contains nearly 2,000 interviews related to a topic by which I'm energized: immigration. I've become interested in how immigrants tell their life stories in relation to immigration, both through listening to my parents' immigration stories and the stories of immigrants in the immigration center at which I volunteer.

Most importantly, the project has specific design characteristics that are interesting entry points for analysis. What makes the Ellis Island Oral History project so compelling as a candidate for analysis is its composition. The project has a singular area of focus: preserving the stories of those touched by Ellis Island. However, this focus was executed through nine iterations of project designs over many decades. With varying interview lengths, interview guides, interviewers, and locations, these series can demonstrate how project design influences what is covered in an interview, as well as how it is covered.

This is useful because the implications of project design decisions are tacitly understood by oral historians but not typically evaluated at a project's completion. For example, oral historians generally assume it's easier to conduct a rich interview with more time than less,[11] using an interview guide can lead to overly rigid agendas or forced chronologies,[12] and interview agendas are ultimately the product of co-authorship by interviewers and interviewees.[13] Owing to its composition and scale, the Ellis Island Oral History project represents rich grounds for evaluating some of the tenets we tacitly understand as oral historians.

Using NLP, I compare the EI and KECK series to better understand relationships between project design and interview content. In particular, I ask two questions:

1. The Ellis Island Oral History project divides its interview collection into two categories, “immigration experience only” and “life narrative” interviews. Considering the topics discussed in the collection, is this distinction mea?
2. If there are distinctions between the “immigration experience only” and “life narrative” interviews, what factors cause those distinctions?

To answer these questions, I use two main tools. GPT-3, a language model released by OpenAI in June 2020, and the Structural Topic Model (STM) package, a Latent Dirichlet Allocation (LDA) topic model created by Molly Roberts, Brandon Stewart and Dustin Tingley,[14] which derives from David Blei’s landmark 2003 paper “Latent Dirichlet Allocation”.[15] Using these tools, I develop and evaluate NLP-based methods for oral historians.

The Ellis Island Oral History project

Project overview

Ellis Island is a federally-owned island in New York Harbor that was the gateway for millions of immigrants to the United States from 1892 to 1954.[16] Conducted between the 1970s and 2023, the Ellis Island Oral History Project represents a long-standing effort to preserve the memories of those whose lives touched and were touched by Ellis Island. The collection's narrators are primarily immigrants to the United States who passed through Ellis Island, but also include Ellis Island staff, members of the US Coast Guard, and descendants of immigrants. With roughly 2,000 interviews, this project is one of the world's largest oral history projects, which makes it appropriate for NLP analysis and which, in turn, can demonstrate the utility for NLP for large oral history archives.

The Ellis Island Oral History project is produced and maintained by the National Parks Service (NPS) and The Statue of Liberty—Ellis Island Foundation, Inc., which was founded in 1982 to restore and preserve Ellis Island and the Statue of Liberty.[17] This oral history project is part of the Foundation's efforts to transform Ellis Island and the Statue of Liberty into tourism and research destinations, including opening the Ellis Island Immigration Museum in 1990, The American Family Immigration History Center in 2001, and the Statue of Liberty Museum in 2016.[18] The Foundation is also engaged in a larger effort to tell, "the entire story of American immigration," with Ellis Island at its center.[19]

The Ellis Island Oral History project is made up of nine series of interviews conducted from 1973 to the present (see Appendix A for summary).[20] The first interviews were conducted by Margo Nash, who was hired by the NPS to head the creation of Ellis Island oral histories in the summer of 1973. Nash later set up the Allee, King, Rosen, and Fleming Incorporated (KECK) series, a series of 200

interviews conducted by a consulting firm for the purpose of producing quotations to supplement exhibits at the upcoming Ellis Island Immigration Museum. The single largest series, at nearly 1,400 interviews, is the Ellis Island (EI) series. The EI series was headed by Paul Sigrist from 1989 to 1999 and began recording in 1990. The primary interviewers for this series were Janet Levine and Paul Sigrist, and from 1991 onward, Janet Levine was the only oral historian on staff and was responsible for day-to-day activities. For this analysis, I will focus only on comparing the two largest series, EI and KECK, as working with larger series will allow for more reliable results, and these two have distinct design features.

A 2005 internal document titled "The Complete Ellis Island Oral History Collection" written by Paul Sigrist and Janet Levine outlines the specifics of each series and the differences between them.[21] Sigrist and Levine broadly categorize the series into two types: "immigration experience only" and "life narrative". "Immigration experience only" interview series, such as the KECK series, are described as dealing primarily with an immigrant's decision to leave for America, the voyage to America, processing at Ellis Island, and adjustment to life in the United States. In contrast, "life experience" series, such as the EI series, are described as "more all-encompassing", and aim to create a more comprehensive picture of an immigrant's life before and after immigration. The EI series is further distinguished from its predecessors by its use of the "Oral History Form", a questionnaire used to source and conduct interviews. Interviewers used the Oral History Form flexibly, rather than prescriptively.

This distinction coincides with broader historical shifts in the field. Launched in 1989, the EI series' development coincided with a building interest in narrative forms, especially life narratives, as a mode of inquiry. Described by Ken Plummer as the "narrative turn", this shift occurred in many disciplines, including anthropology, psychology, sociology, and oral history.[22] Narrative became an

important concept in emerging thinking around oral history, as oral history distanced itself from simple testimony and the accounting of facts. Alessandro Portelli distinguishes narrative from simple testimony:

> Oral testimony has been amply discussed as a source of information on the events of history. It may, however, also be viewed as an event in itself and, as such, subjected to independent analysis in order to recover not only the material surface of what happened, but also the narrator's attitude toward events, the subjectivity, imagination, and desire that each individual invests in the relationship with history.[23]

Within this context, the EI series resembles an attempt to move beyond extracting simple testimony about immigration and toward a life narrative approach that would be familiar to many oral historians today.

Accessing the archive

The Ellis Island Oral History project can be accessed in several ways, including by visiting Ellis Island to explore interviews on dedicated computers or reading a number of excerpts that are used to describe exhibit stations. End-users may also access the project online, interact with excerpts on the National Parks Service's website,[24] or read derivative works such as Peter Morton Coan's *Ellis Island Interviews: In Their Own Words*.[25] But users accessing the Ellis Island Oral History Archives through the National Parks Service's dedicated website face several challenges that my research addresses.[26] First, not all interviews hosted on the site have an available transcript. Second, search functions are limited. Users may filter by first and last name, country of origin, and topic, though not by their own search terms. In addition, not all interviews are properly tagged, meaning that users filtering by the aforementioned dimensions receive incomplete results. For example, when filtering interviews by the topic "Voyage to America"—a topic one would expect to be present in nearly every interview—the site returns only two results. Gaps in tagging reflect how the tags were created: through discontinuous volunteer work rather than a well-resourced

endeavor.[27] Moreover, it is also not clear *where* in the interview this topic appears; to find instances where "Voyage to America" is discussed in an interview, one would have to read the interview in its entirety. As I will show, NLP methods allow me to address these concerns.

I contacted the National Park Service to gain direct access to the archive and its ancillary files and they provided the following digital files:

1. A folder with all Ellis Island transcripts and miscellaneous related files (2,397 total files).
2. "The Complete Ellis Island Oral History Collection", a 2005 five-page internal document written by Sigrist and Levine that describes each of the project's nine oral history series.
3. The Oral History Form used to locate and conduct EI Series interviews.

Two NLP use cases for oral historians: Metadata mining and Topic modeling

The files provided by the NPS required considerable pre-processing to be organized, cleaned, and otherwise prepared as inputs for further analysis. Through this pre-processing work, I develop two use cases for oral historians: metadata mining and topic modeling.

[1] Metadata mining

Notably, there is no metadata sheet describing and summarizing the collection's contents. In practice, missing metadata means that there is no efficient way for a researcher to organize or locate transcripts by country of origin, interviewer name, age at immigration, narrator gender, or other dimensions that might be pertinent to research. Metadata is critical for organizing and making sense of a collection, but as is especially the case with older oral history archives, metadata can be broken, inaccurate, incomplete, or missing.

Missing metadata limits discoverability and, consequently, potential engagement with interviews. For example, a researcher listening to a full-length interview might want to learn more about the experiences of immigrants who survived pogroms in Russia. To find more such narratives, the researcher would have to individually open each interview and read a header to determine the narrator's country of origin. Then, the researcher would have to read the interview to determine if pogroms were discussed. If the researcher later wanted to examine these narratives in conversation with testimony from survivors of pogroms in other regions, they would have to restart the process. Though this kind of engagement with the text might lead an end-user to serendipitous findings, this process's labor intensity might mean that the opportunity to locate narratives and engage more deeply with the corpus is lost.

The Ellis Island Oral History project does not have a master sheet that attaches metadata to each interview. Instead, each interview's metadata is stored in a header at the top of each interview (Figure 1). This metadata header includes named elements including the interview date, the interviewer's name, the narrator's name, and the narrator's country of origin.

Figure 1: Example of an Ellis Island Oral History interview header[28]

EI-613
MARY CATHERINE THERESA BOYLE KELLY
BIRTH DATE: JANUARY 21, 1911
INTERVIEW DATE: MAY 2, 1995
RUNNING TIME: 23:14
INTERVIEWER: PAUL E. SIGRIST, JR.
RECORDING ENGINEER: SAME
INTERVIEW LOCATION: HOMESTEAD HEALTH CENTER
STAMFORD, CONNECTICUT
TRANSCRIPT PREPARED AND REVIEWED BY: PAUL E. SIGRIST, JR., 4/1998

IRELAND, 1929
AGE 18
PASSAGE ON "THE BALTIC"

A conventional way to extract this metadata would be to manually open each file and record the desired elements. However, assuming recording and double-checking each header takes a human 5 minutes to complete, it would take a full-time worker almost five weeks to extract all possible metadata from all documents in the Ellis Island corpus. Extraction in this manner is tedious and expensive (which is perhaps why nobody has attempted to do it).

A more automatic way of extracting this information would be to query the text for specific patterns. For example, when looking for a narrator's birthdate, one could write a script to extract every instance of "BIRTHDATE:" along with the date that follows. However, many hands produced these transcripts over many years and consistency therefore varies. For example, "BIRTHDATE:" sometimes appears as "BIRTH DATE:" or "DATE OF BIRTH:". Moreover, dates are stored inconsistently in a number of formats; May 1st, 1909 could be stored as "May 1, 1909", "05/01/1909", "1 May 1909", or any other imaginable format. As such, writing code to extract predefined elements for this task is complicated, iterative, and risks missing large swaths of data. In addition, each interview also contains valuable elements that are easily interpreted by a human but not easily extracted based on patterns. For example, in reading the first page of Ms. Kelly's interview, one could discern the narrator's gender from the narrator's name and the pronouns by which they are addressed. In addition, one could discern that Ms. Kelly is indeed an immigrant rather than a worker on Ellis Island, which would be an important distinction for researchers.

Turning away from both manual and pattern-based approaches, I opted to use a language model. A language model is commonly defined as a probabilistic mechanism for generating text.[29] In plain language, a language model is a tool that can understand human language in terms of probabilities, which it is taught to do by being exposed to millions of examples of text—commonly books, newspaper articles, and language from the internet. Through these examples, language models learn what words are most likely to come next in a given sentence. Language models are typically general-purpose and are used for tasks including text summarization, classification, and translation. At the time of writing, the most commonly known language model is OpenAI's ChatGPT. Importantly, in contrast to the aforementioned pattern-based approaches, language models are better suited to process semi-structured or unstructured text—typos and all.

Implementation

For this task, I opted to use GPT-3, a popular language model produced by US-based artificial intelligence firm OpenAI. As a general-purpose model, GPT-3 is able to perform many tasks reasonably well out of the box, such as summarizing text, writing poetry, or powering chatbots. GPT-3 can also be retrained, or "fine-tuned", to perform specific tasks. I chose to fine-tune GPT-3 to extract metadata from the roughly 2,000 transcripts I acquired from the NPS so that I could guide the extraction toward a format that would be useful in an archive (i.e., individual columns for defined elements).

This process involved several steps. First, I isolated the first page (which includes the metadata header) of 100 random interviews. I selected these interviews using a random number generator in Microsoft Excel. Of these 100 interviews, I manually extracted 12 points of metadata from the headers as a comma-separated list.

In addition to extracting named elements from the metadata headers, I also performed a few interpretive tasks. For example, gender, number of narrators, and type of narrator (i.e., immigrant vs. non-immigrant), are not explicitly named in the interviews' metadata headers but are interpretable to a human. Accordingly, I included these elements in my extraction.

The elements I recorded were:

1. Interview code number
2. Interview date
3. Interviewer name
4. Interview duration

5. Narrator name
6. Narrator country of origin
7. Narrator date of birth
8. Narrator gender
9. Narrator age at immigration
10. Flag if narrator is a non-immigrant

Next, I fed these 100 documents and related extractions into GPT-3 in a process called “fine-tuning”, whereby GPT-3 learns how to replicate the work I have conducted. This exercise outputs a new, “fine-tuned” model that can better perform the task at hand.

I tested the fine-tuned model's performance by manually exposing it to 50 new interviews and comparing its results to additional manual extractions. In practice, this meant pasting a header into the model, and evaluating the string of named elements it extracted to ensure consistent accuracy and order. Because the results were accurate, I wrote a Python script to automatically process every interview in the corpus through the fine-tuned model and to record all results into a master table.[30] The result of this process is a table with columns that represent each piece of extracted metadata. For example, returning to Mary Catherine Theresa Boyle Kelly’s interview, the eight columns of metadata are stored as follows:

Code	Narrator	Interviewer	Int. date	Country	Birthdate	Age at Imm.	Gender
EI-613	Mary Catherine Theresa Boyle Kelly	Paul Sigrist	May-95	Ireland	Jan 1911	18	F

In addition to the columns directly extracted by GPT-3, I created several derivative columns. First, in instances where duration was not explicitly named in a header, I estimated interview duration using document length as an approximation. Second, I assigned each interview to its corresponding

series by referencing the first two characters of each interview code (i.e., EI = Ellis Island, K = KECK, etc.). These additional data are useful for further analysis.

I then sampled 100 interview extractions using a random number generator and compared these extractions against their source interview files. I found that, on average, extractions performed by the fine-tuned model were inaccurate less than 1% of the time across all fields. Most instances of error occurred when the model handled interviews with unusual formats, such as an interview with three narrators. Considering that manual data entry typically has an error rate of around 1%,[31] I found this result to be acceptable.

Surprisingly, the most challenging part of this process was not learning how to work with GPT-3 but figuring out how to accommodate differences in encoding and file types that have emerged as the collection has been handled over the years. Documents were saved in .rtf, .doc, .docx, and .txt formats with different encodings, some of which were rare or unreadable in a modern word processor or with a standard Python package. For the files that could not be read, I manually opened, inspected, and saved them in legible formats. Ultimately, around 50 interview files were not easily salvageable.

Uses for oral historians

The main benefit of using the method I've developed, or any related method using a language model like GPT-3, is resource savings. Based on the time it took to manually code the training set, working at a consistent pace with no breaks, it would have taken over 60 hours for me to manually extract the metadata. Outsourcing this work would be costly. At the time of writing, the average hourly wage for a data entry clerk is $20. The Oral History Association (OHA) endorses the American Folklore Society's guidance to offer double a salaried worker's hourly wage to contract employees, meaning that this data entry project could cost $2,400 to complete in accordance with industry standards.[32] Processing via GPT-3 reduces the resource intensity of metadata extraction significantly. It took three hours to manually code 100 interviews, ten minutes to train GPT-3 to do the same, and just under an hour and a half to process the entire 2,000-document corpus. The total processing cost for my work was under $100 based on OpenAI's per-token pricing scheme.[33]

Given that projects often have pre-defined, limited budgets, saving on metadata extraction means spending and labor can be allocated in other higher-value areas, such as caption creation, public programming, or translation. Moreover, existing oral history projects that have not invested in metadata at the outset may find an opportunity to create metadata for their projects using NLP, improving discoverability and accessibility. In some instances, using NLP to assist with metadata creation can mean the difference between having metadata or none at all.

Limitations and considerations for oral historians

Part of what makes GPT-3 so efficient is that its processing occurs on powerful cloud-based servers instead of an end user's machine. Though cloud processing improves access for users with low computational power, it also creates points through which data can be compromised—both through

transmission to OpenAI and through handling at OpenAI. Though, at the time of writing, OpenAI's known data policies and procedures are not particularly unscrupulous,[34] sharing information with a third-party always represents a relinquishing of data control. Moreover, as the organizations that operate models like GPT-3 merge, divest, and evolve, the policies that govern user data also evolve.

For projects that have already gone public, there is little additional privacy and security risk in using cloud-based models like GPT-3. As noted, language models are trained on large collections of text, often including the internet, so it is possible that language models are reading digitally-hosted interviews. Though this reality may make practitioners uncomfortable, engaging directly with these models does not always increase the likelihood that your project will be read by a machine.

When dealing with closed oral history projects or narrations that contain politically or personally sensitive data, GPT-3 may not be an appropriate choice. As part of the process of informed consent, depending on the sensitivity surrounding an interview and its planned release date, it may be appropriate to discuss using cloud-based NLP products like GPT-3 with narrators, as cloud-based processing represents a stage in the production process where information is being shared beyond project staff. Alternatively, an oral historian might consider using a language model that can run locally, such as GPT-Neo or GPT-J. As the corpus I am working with contains interviews that have been public for decades, using GPT-3 was ultimately appropriate for my research.

Another limitation of using sophisticated language models is their interpretability. Though GPT-3 exhibits human-like performance on many tasks, it's not currently possible for end-users to readily understand *why* GPT-3 makes the decisions that it does. For example, in training GPT-3 to extract gender from the Ellis Island interviews, I can theorize why GPT-3 makes the assignments that it

does, but I ultimately don't know. Practitioners should take steps to evaluate the model's performance by checking the model's results against comparable human outputs.

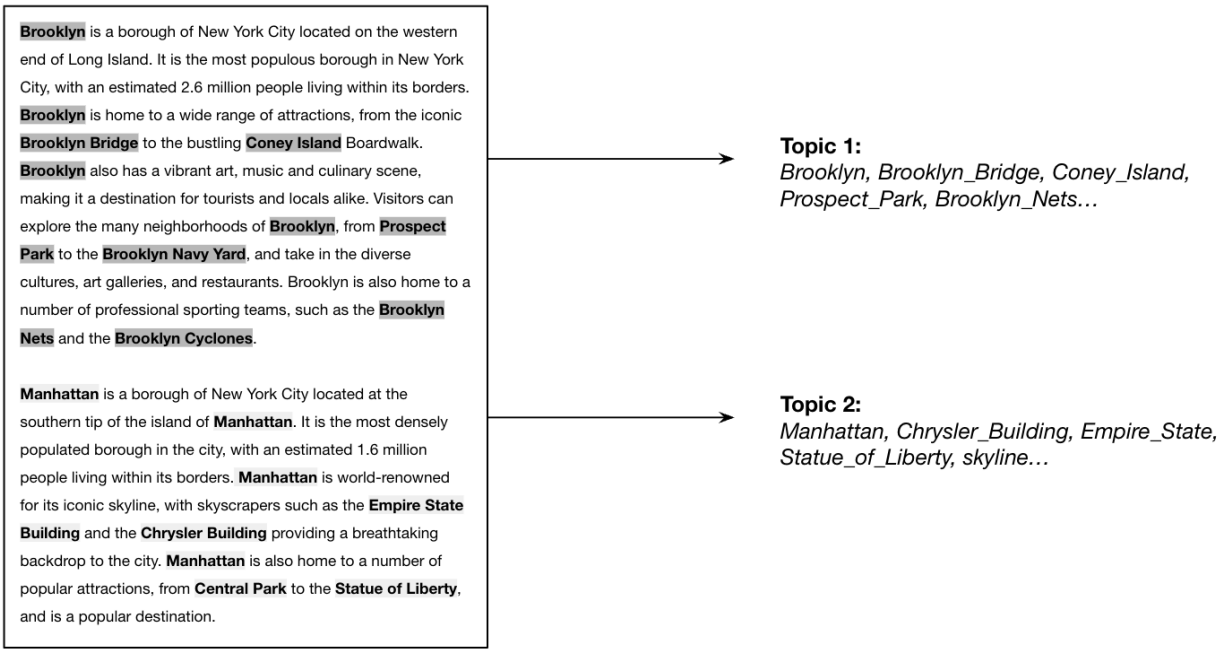
[2] Topic modeling

The metadata discussed thus far help us understand who is involved in each interview, but not what the interviews are about.

One of the most popular NLP methods for understanding the contents of a corpus is topic modeling. Topic modeling is a type of statistical NLP used to identify latent thematic structures in a corpus. In plain language, topic modeling is a set of techniques that allow researchers to identify topics in a corpus and assign topics to each document. Topic modeling is, typically, endogenous, meaning topics are generated using only the corpus under consideration rather than some external reference point or knowledge.

"Topics", as output by a topic model, can be thought of as mixtures of words that tend to co-occur in a corpus. For example, consider a topic model trained on 100 New York City travel guides (Figure 2). The model may identify one topic as composed of the words, "Brooklyn", "Brooklyn_Bridge", "Coney_Island", "Prospect_Park", "Brooklyn_Nets"; a human might interpret this topic to be about Brooklyn. A second topic identified by the model might be defined by "Manhattan", "Chrysler_Building",'Empire_State","Statue_of_Liberty","skyline"; a human might interpret this topic to be about Manhattan.

Figure 2: Topic identification using a topic model



In addition to identifying the topics in a corpus, topic models assign topics to each of the corpus's constituent documents. Sometimes this assignment is a single topic, but some of the most popular models assign a mixture of topics to each document.

But how does the model know how many topics to identify? The number of topics that a topic model identifies is often determined by humans. There are no hard and fast rules for how many topics a model should be tasked with identifying; training a model to find 10, 20, 50, 100, or 1000 topics is common. Generally, a lower number of topics yields more latent topics, and a higher number of topics yields more specific topics. For example, a five-topic model trained on New York City travel guides might produce one topic for each of the city's boroughs. With this model, each guide evaluated would be represented by a mixture of each borough (Figure 3). A hundred-topic model

may yield topics at the neighborhood level in addition to the borough level. It might also produce topics related to restaurants, modes of transportation, and landmarks.

Figure 3: Hypothetical New York City travel guides and their topic mixes

Topic modeling was initially conceived of as a tool for categorizing documents but has evolved to be used as a tool for discovering and analyzing latent themes and structures within a corpus. Evaluating New York City travel guides with such a model, a researcher might examine how prominent each borough is across different kinds of guides (Figure 3). Alternatively, a researcher might be interested in examining how each borough's prominence in guides has changed over time.

Given that I want to understand how interview topics vary by interview series, topic modeling is a suitable approach. To analyze the Ellis Island Oral History's topic structure, I opted to use an LDA topic model.

LDA topic modeling

A well-studied approach to topic modeling is called Latent Dirichlet Allocation. LDA is a generative probabilistic model that simultaneously tries to determine word lists that define *n* number of topics in a corpus, as well as the proportion of topics within each constituent document.[35] In plain language, LDA topic modeling works by looking at many examples of documents, recording which terms appear in each document in a matrix, and measuring how likely these terms are to co-occur. An algorithm then identifies topics (again, the number of which is specified by the end user) and creates lists of words that define each topic. Topic modeling conceptualizes a corpus's constituent documents as made up of mixes of topics in varying proportions.

Implementation

To measure the Ellis Island corpus's topics, I opted to use the Structural Topic Model (STM) package developed by Molly Roberts, Brandon Stewart and Dustin Tingley. STM is considered a standard LDA topic model by practitioners in many fields, and its limitations are well documented, tested, and understood. STM is based on David Blei's 2003 "Latent Dirichlet Allocation", and though LDA technology is now 20 years old, it still performs well compared to emergent models.[36] LDA topic models are also known to perform especially well with long documents like novel chapters, but have also been used to analyze documents as short as tweets.

Compared to regular LDA topic models, structural topic models consider metadata in their calculations, such as publication date, an author's political affiliation, a text's genre, or any other chosen dimension. For example, a researcher analyzing gubernatorial speeches throughout the COVID-19 pandemic might use topic modeling to track dominant speech topics as the pandemic

progressed. The researcher may also consider a governor's party membership to examine how party affiliation relates to messaging priorities. For the Ellis Island Oral History project, using metadata can enable us to compare the agenda of interviews based on the interviewer, narrator, country of origin, gender, or other factors. For example, immigrants arriving at Ellis Island were received in a large hall. The project's interviewers, who are deeply familiar with the site, refer to this area as the "Great Hall" or "Registry Hall". Narrators, who usually spent less than a day on Ellis Island and spoke limited English, sometimes refer to this area as merely "a big room".

> SIGRIST: Do you remember being in the Great Hall here at Ellis Island?
>
> IOZZIA: It was a big, big room. Which room it is over here, I don't know. I can't figure out which one it is.[37]

Structural topic models can be trained with these categories of speakers in consideration and better accommodate differences in language choice when different authors discuss the same topic.

To take advantage of these benefits, I initially intended to train STM on a diarized version of the corpus, assigning each turn of speech to its own document.[ii] My intention was to take advantage of STM's ability to measure how different speaker categories (i.e., narrator vs. interviewer) talk about the same topics using different language. However, in practice, I found that training STM on a diarized corpus tended to produce topics that were too specific to one speaker category to draw meaningful comparisons. Two reasons underlie this performance issue. First, LDA models generally perform better on longer documents. Turns of speech, especially those from interviewers, can be quite short—sometimes one word in length (e.g., "Understood."). Second, interviewers use

[ii] Diarization is the process of separating multiple speakers in a recording.

procedural language (e.g., “go on”, “tell me about”, “thank you”) that tended to be overrepresented in my results. As such, I opted not to take advantage of STM’s metadata capabilities in this way and instead used STM as a regular LDA model.

Working with a corpus in an LDA topic model requires a significant amount of pre-processing. Pre-processing involves “cleaning” the text by removing inaccuracies and noise, thereby improving performance by allowing tools to focus on terms that add to a sentence’s meaning. To pre-process the corpus, I used Quanteda, a common package designed to facilitate quantitative text analysis in R.

A standard first step in preprocessing is converting the corpus into *tokens*. Tokens can be thought of as a list of units of a specified length—for example, characters, words, bigrams (groups of two words), sentences, or paragraphs. Most LDA analyses are based on tokens that roughly correspond with words. For example:

> APPLEBOME: Before the boat came into the harbor and you got onto Ellis Island, do you remember seeing the Statue of Liberty?[38]
>
> *"applebome","before","the","boat","came","into","the","harbor","and","you","got","onto","ellis","island,","do","you","remember","seeing","the","statue","of","liberty"*

A common next step is to remove stopwords—common words that do not usually add meaning to a sentence, such as “the” or “a”. Removing stopwords from text can improve a topic model’s accuracy, as stopwords are often not meaningful on their own. Removing stopwords also makes text easier to process and can reduce the size of text data sets, allowing for faster analysis. For the purposes of this work, I used a standard English stopword list that comes bundled with Quanteda.

Through my work, I found that oral history interviews require some additional pre-processing to arrive at coherent and interpretable topics. As such, I added the following steps.

[1] Removing transcription artifacts

Oral history transcripts typically have a significant amount of markup that helps humans interpret them. However, this markup can create negative interference for machines. Artifacts denoting throat clearing, mumbling, or coughing are not usually meaningfully correlated with the words a narrator uses, but LDA models do not know to ignore these artifacts. LDA models are instead agnostic to these artifacts, and consider them as part of the words that add meaning to the text. For example, without pre-processing, an LDA model may identify “coughs” part of a topic. A narrator who coughs frequently and speaks extensively about their childhood in Poland may encourage a model to identify “coughs” and “Poland” as part of the same topic. For this reason, I removed all standard non-speech markups from the corpus.

[2] Stripping proper nouns

Oral history interviews are typically diarized, meaning that they identify each speaker by their surname and a colon. These markings are not understood by the model as metadata and are instead considered meaningful words. Given that these identifiers are repeated throughout the corpus, and this metadata has been captured at the document level, they only serve to jeopardize the model’s ability to evaluate meaningful text. For this reason, I removed all diarized names from the corpus.

Still, non-diarized proper nouns can interfere with LDA modeling. For example, Maria is a common name referenced in interviews with Italian immigrants. In early rounds of modeling, I found “Maria” appearing inappropriately in otherwise coherent topics (e.g., “work”, “labor”, “union”, “maria”). There are a number of ways to remove these proper nouns. One method would be to use a named entity recognizer, a tool that recognizes the names of people, places, and things and allows one to strip them from the text. However, the wholesale removal of proper nouns could result in loss of meaningful terms, such as “New York” or “Ellis Island”. Moreover, stripping mixed-noun surnames like “Green” would also inadvertently strip the word green as a descriptor, which is consequential for compound terms such as “green card”. To remove interference without unnecessarily stripping data, I created a custom list based on common names in the corpus and the most common names in the United States during the nineteenth century.[39]

[3] Recombining select tokens

Splitting tokens at the word level can cause important multi-word terms to lose meaning. Common multi-word terms in the Ellis Island corpus include “New York”, “Ellis Island”, and “the Statue of Liberty”. Split into “new” and “york”, New York can appear in a topic model’s result as “new” or “york”, reducing the model’s interpretability. Furthermore, the model considers “new” in “New York” the same as “new” in contexts with completely different meanings, including “New Jersey”, “new shoes”, or “new world”. This lack of distinction, in addition to reducing interpretability, can reduce the model’s performance. As many key corpus terms (e.g., Ellis Island, Coast Guard, United States) are multi-word terms, I created a shortlist of tokens to recombine before processing in STM. To generate this list, I looked for common multi-word terms in the transcripts. When iterating through STM processing, I also used noise or unexpected terms as a starting point for identifying additional multi-word terms.

[4] Enhancing context through chunking

LDA topic models are sometimes categorized as a "bag-of-words" (BoW) approach. BoW describes how LDA treats tokens as if they were floating in a grab bag—out of order and stripped from their original context. To statistically analyze a corpus, an LDA topic model creates a matrix where each token is represented by a column, each document is represented by a row, and each cell represents a count of how many times a token appears in each document.

The primary limitation of BoW approaches is that they do not take into account word order or context. Though the recombination process I outline in step [3] helps introduce some context, as documents become longer, the "bag" of words becomes larger, and context is reduced. Accordingly, when analyzing literature with LDA topic models, humanists often process novels at the chapter-level, which are usually 1,000—5,000 words in length. In comparison, each oral history interview often produces ten times that amount. To add back in some contextual information and improve the model's performance, I "chunked" the corpus every 150 tokens. In practice, this means that a new document is created every 150 words.

Fitting STM and evaluating topics

With a preprocessed and chunked corpus, I ran the project's documents through STM. Given that I had a large, detailed corpus, I tasked STM with finding 100 topics. The final output was 100 topics, word lists that defined each topic, and a mixture of topics in each document chunk.

At first glance, most topics appeared to have a coherent meaning. Topic 10 ("ship", "class", "deck", "voyage", "cabin") is clearly about ships. Topic 11 ("school", "high", "teacher", "grade", "class") is clearly about school. Other topics, such as Topic 61 ("room", "looked", "waiting", "told", "asked"), might be recognizable as describing initial arrival and examination at Ellis Island, but only to someone who knows the Ellis Island corpus. Either way, simply imposing meaning on the word lists that a model outputs is insufficient.

The main risk with using topic models—using LDA or other tools—is a human tendency to impose meaning on model outputs where there may be none. Jonathan Chang et al. outline this issue in "Reading Tea Leaves: How Humans Interpret Topic Models" and propose using surveyed human evaluations to ensure topics and their constituents are consistent with human judgment.[40] The authors propose two complementary tests for measuring topic coherence: word intrusion and topic intrusion. Word intrusion tests involve picking a random topic from a topic model, taking the top five most common words, and adding in an unrelated "intruder" word. If the topic is coherent, a separate human evaluator should be able to identify the intruder. Topic intrusion works in a similar manner: a human evaluator is presented with a document, several high-probability topics assigned to that document, and one low-probability document. When evaluating my results with word intrusion, a human evaluator correctly identified 83% of intruder words. When evaluating for topic intrusion, a human evaluator correctly identified intruder documents 11% of the time.

Once I assessed the topics' coherence, I grouped several topics based on proximity to enhance interpretability. For example, I grouped topics related to ethnic groups, religion, and holidays together into a higher-order topic called "Ethnic Origins and Customs". I preserved both levels of topics in a table so that I could analyze findings at both levels (Table 1).

Table 1: Example of five topics and classifications

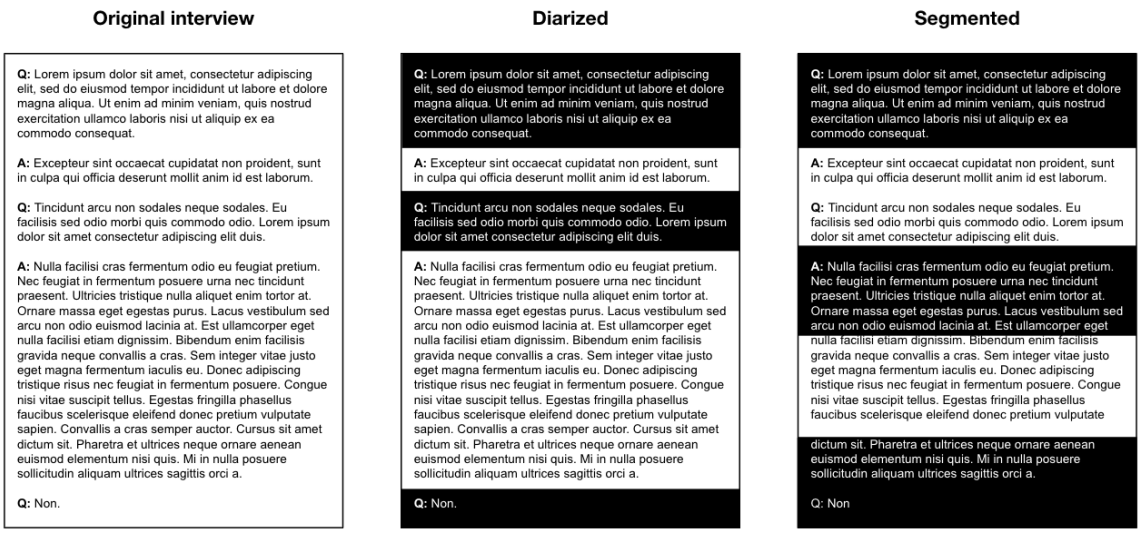
Highest probability words	Classification	Higher-order classification	Immigration-relation
ellis_island, new_york, statue_of_liberty, coming, oral_history	Ellis Island	Ellis Island	Immigration-related
age, spell, born, maiden_name, birth_date	Biographical information	Biographical information	Other
ship, class, deck, voyage, cabin	Voyage - Ship	Voyage	Immigration-related
school, high, teacher, grade, class	School	School	Other
christmas, special, tree, holidays, holiday	Holidays	Ethnic Origin & Customs	Other
…	…	…	…

For a full list of topics and topic classifications, see Appendix B.

Re-segmenting the corpus

Though I produced reasonable topic coherence with STM, "chunked" documents are not ideal for the purposes of analyses. Each chunk contains a mixture of interviewer and interviewee passages, and although each chunk is comparable in length, each interview is made up of a different number of chunks. As such, it's difficult to interpret or compare interviews to each other based on these chunks. To enhance the interviews' interpretability and comparability, I reprocessed the corpus in two ways: diarization and segmenting (Figure 4).

Figure 4: Diarized vs. segmented interviews



Diarization is the process of splitting a transcript based on speakers. Splitting a transcript based on diarization allows further processing to apply topics to specific turns of conversation. To perform this task, I used a pattern-based tool in R to split each interview into interviewer and interviewee segments. The output of this work was 373,534 documents, each representing an instance of speech. Then, I reprocessed this output through the STM model I previously created, which assigned the previously generated topics to these new documents.

Though diarization isolates turns in the text, it does not solve the issue of differing document lengths. When diarized, interviews are represented by segments that are unequal in length and number, and therefore not comparable between interviews. To create comparability, I standardized the interviews into points of “narrative time” using a technique developed in “Narrative Paths and Negotiation of Power in Birth Stories” by Maria Antoniak et al.[41] This technique involves splitting each document into an equal number of points so that a reader can compare narrative structures

between documents of varying lengths. With standardized points, each interview's development can be compared over what Antoniak et al. call "narrative time". I chose to split each interview into 20 data points. Oral history interviews are typically indexed at 5–10 minute intervals, corresponding with typical topic discussion lengths.[42] Segmenting interviews into 20 points, considering interviews are typically between 60 and 90 minutes in length, yields similar but slightly improved fidelity; each point represents 3-5 minutes of conversation on average.

Uses for oral historians

Topic modeling can enable oral historians to understand the contents of large corpora. In the simplest applications, a topic model could be used to tag and classify documents in a corpus. This tagging can occur at the document level (i.e., what's in this document?) or some smaller unit e.g., what's in this paragraph?). This form of classification can be used in a number of ways.

A researcher might use a topic model to quickly identify and classify sections of a corpus that address particular topics. For example, Topic 43, "Violence" could be used to locate violence-related push factors among immigrants from many different countries. A researcher querying the corpus this way might find unexpected connections, such as similarities in discussions of violence in Eastern Europe and Anatolia. Topic modeling, in this sense, is a companion for enhancing close reading and discovering interconnections.

An archivist or public humanities professional might use topic classifications to create new curatorial or user experience pathways. For example, the Ellis Island Oral History Project might use document-level topic assignments to tag interviews for their website, solving the topic tag gaps that the site currently exhibits. In addition, they may use paragraph-level topic assignments to allow users to find and compare specific instances in interviews where narrators discuss a topic, such as

their voyage to America. In both of these instances, topic models help end-users engage with and find interconnections between interviews. Finally, as will be explored in this paper, oral historians can also use topic modeling for quantitative analyses.

Limitations and considerations for oral historians

Oral historians should consider several limitations and considerations when working with topic models. First, LDA topic models, such as STM, perform best with large corpora. Though there are no set rules for how large a corpus should be for LDA candidacy, my sense through experimentation is that for oral histories, 50–100 interviews might be the minimum corpus size. Most oral history collections are much smaller. However, emerging models like BERTopic or Top2Vec allow users to identify topics with much smaller corpora. These models achieve this ability through pre-training. Pre-trained models, however, are not endogenous, and carry the biases of the texts they were trained on.

As previously discussed, topic models also risk encouraging users to impose meaning where there is none. Combating this risk involves human evaluation and interpretation, which can only occur when interpreters know the corpus's domain well. Oral historians should use topic modeling to supplement, not supplant, closer engagement with corpora.

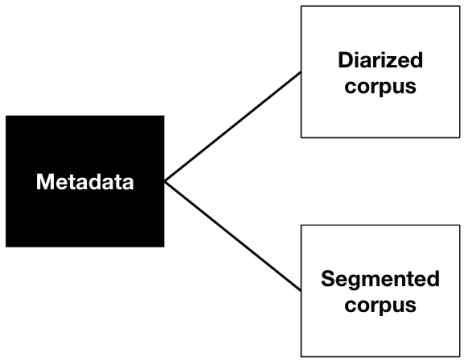
Oral historians may also be concerned by the amount of pre-processing that must occur for a corpus to be processed through a topic model. Slicing, dicing, and removing language from a transcript might appear to be a crude act of interpretation that jeopardizes the original transcript's meaning. However, the transcript itself is already an interpretation of the interview moment created by a transcriptionist, an audit editor, and other stakeholders. A topic model, like a transcript, is

merely another representation of the interview moment. Different representations of the interview moment—recordings, transcripts, dramatizations, and retellings—carry their own strengths and limitations. Topic modeling should be considered through this same lens. Topic models might not be used to understand the intricacies of an individual narrative, just as an individual narrative might not be used to make claims about the contents of an entire corpus.

Example analysis

After completing metadata extraction and topic modeling exercises, I stored the data from the metadata extraction and topic modeling exercises in spreadsheets. To enable cross-query, I connected these sheets to each other in a Microsoft Excel Data Model to create a small database (Figure 5). This database enables efficient querying and measurement of the corpus.

Figure 5: Ellis Island Oral History Project database



Research question #1

The Ellis Island Oral History project divides its interview collection into two categories, "immigration experience only" and "life narrative" interviews. Considering the topics discussed in the collection, is this distinction measurable?

Before reading transcripts from the EI or KECK series, I consulted Sigrist and Levine's documentation on the differences between each series. Levine and Sigrist describe EI's greatest contribution as shifting the project toward "life narrative" interviews with greater topical breadth than previous series, such as KECK.

KECK, by contrast, is described as an "immigration experience only" series, with little consistent focus on a narrator's life outside of their experiences as an immigrant and in relation to Ellis Island. Furthermore, a consultancy conducted KECK with a defined goal of producing site-specific museum narratives. Accordingly, in KECK, one would not expect to find narrators relating their lives beyond how they intersect with Ellis Island.

But how much of each interview should be dedicated to immigration-related topics? The Oral History Form that was used to conduct the EI series interviews suggests that roughly 30% of the interview should cover immigration-related topics (Appendix C).[iii] The balance of the form is related to pre-immigration and post-immigration life experiences.

Though there is no evidence of an interview guide used for the KECK series, it is reasonable to assume that KECK interviews would more heavily feature immigration-related topics. However, even a cursory dive into the KECK interview series reveals that most interviews span topics beyond

[iii] Determined by measuring the proportion of questions in the interview guide related to immigration.

"immigration experience only" topics. Consider the opening of KECK-65, an interview with Anna Perlstein, an immigrant to the United States who was born in Poland.[43]

DANE: This is Debby Dane and I'm speaking with Anna Perlstein on Thursday, October 24, 1985. We are beginning the interview at 2:35 PM and we are about to interview Anna Perlstein about her immigration from Poland, uh, in . . .

PERLSTEIN: Uh, from Israel.

DANE: From Israel in about 1913.

PERLSTEIN: I don't know. I couldn't tell.

DANE: This is interview 65. Mrs. Perlstein, can you tell me what day you were born and what town you were born in?

PERLSTEIN: I, I, I am not positive, but I always make it April 9. I was, I think I was born in Poland, Bialystock.

DANE: Do, do you remember what kind of town it was, what it did for what people did for a living?

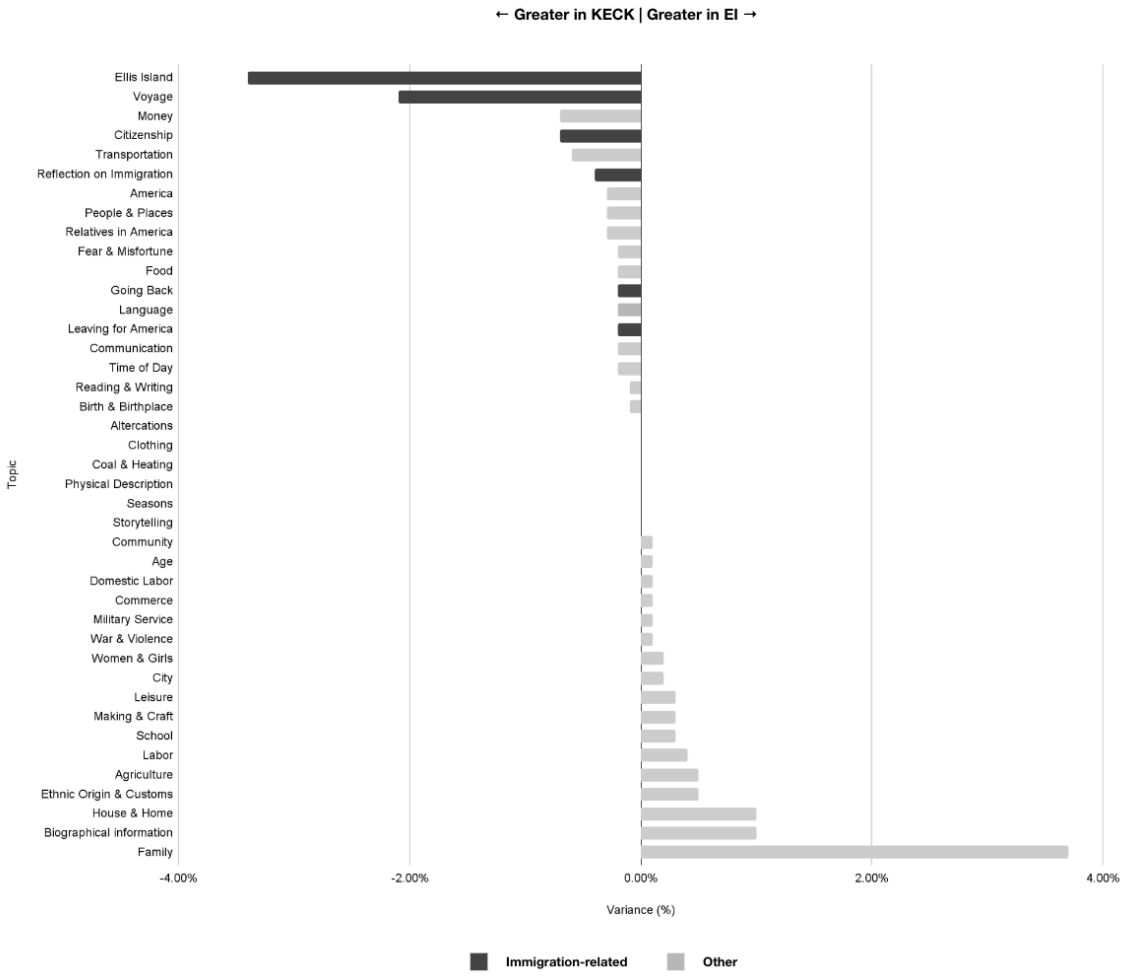
PERLSTEIN: How would I? I was only five years old when I left. I know my mother wasn't in very good circumstances. My father was in the United States and he sent her a little money whenever he could and then I had an aunt living in Warsaw, that's capital of Poland, and she had one son but he died and they wanted a child in her house and she want, she and her husband, so they came to Paris to, uh, where was I, in Bialystock, and asked my mother if she couldn't have one of the children live with her…

Perlstein goes on to discuss her life before immigration until roughly 40% of the way into the interview. Many, though not all KECK interviews appear to discuss the narrator's life outside of their immigration experience.

Comparing EI vs. KECK topic structures

If Sigrist and Levine's distinctions between KECK and EI are meaningful, KECK interviews should contain a greater share of immigration-related topics.

Figure 6: Variance in topic share, EI vs. KECK

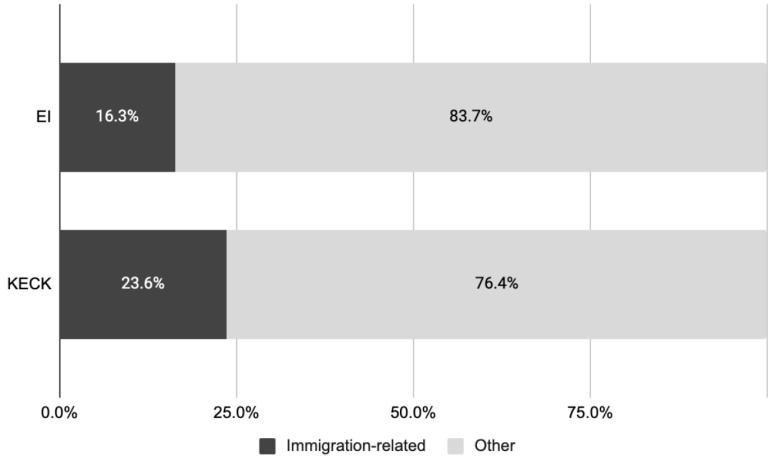


The chart in Figure 6 compares the overall topic mix of the EI and KECK series.[iv] For example, the topic "Ellis Island" is 3.3 percentage points less prevalent in EI series interviews than in KECK series interviews. This chart makes clear that the topics the EI series under indexes relative to KECK are generally related to immigration (shaded in dark gray). By contrast, topics related to domestic life, leisure, and family are more prevalent in the EI series.

[iv] For each topic, I subtract the average KECK topic share from the average EI topic share.

By grouping topics into a higher-order categorization (“immigration-related” and “other” topics) the overall difference becomes more apparent (Figure 7). Immigration topics account for 23.6% of KECK interviews, versus only 16.3% of EI interviews.

Figure 7: Proportion of immigration-related vs other topics, EI vs KECK



The difference between EI and KECK interviews is not as large as one might imagine based on the descriptors “immigration experience only” and “life narrative”. Furthermore, the share of immigration topics in *both* series is far lower than what the Oral History Form suggests for the EI series (30%).

Though this difference is smaller than anticipated, it is measurable and retains relative consistent when slicing the corpus by different variables, including country of origin, narrator gender, narrator age at immigration, and interviewer. For example, interviews by the top two interviewers in each series produce immigration-related topic shares close to the average of their respective series (Table 2).

Table 2: Immigration-related topic share, top two interviewers per series

EI	Share (%)
Average	16.3%
Janet Levine	15.6%
Paul Sigrist	17.4%

KECK	Share (%)
Average	23.6%
Nancy Dallett	23.2%
Dana Gumb	26.7%

Though the difference in immigration-related topic share between EI and KECK is smaller than anticipated, it is consistent and meaningful. But what drives this difference?

Research question #2

If there are distinctions between the "immigration experience only" and "life narrative" interviews, what factors drive those distinctions?

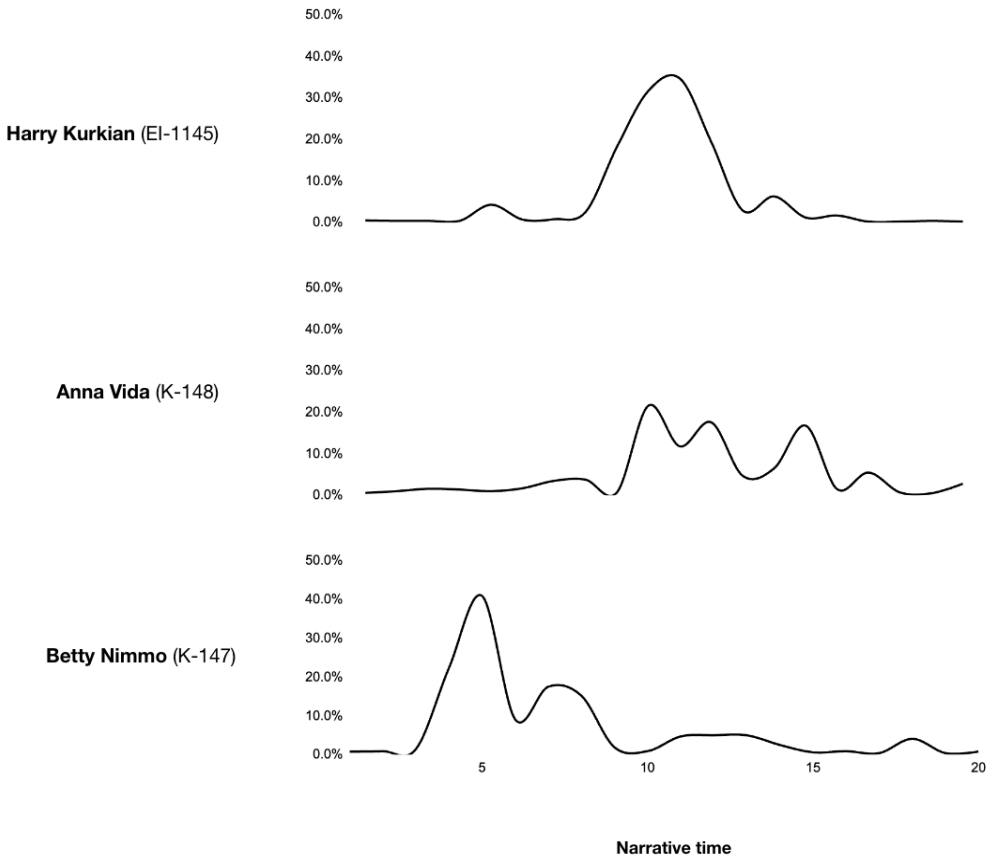
Oral historians tacitly understand that a number of factors are at play in determining an interviewer's agenda. An interview's duration, for example, can limit a conversation's breadth and depth. An interview's physical (or virtual) setting can also shape a conversation's boundaries. And of course, oral historians commonly accept that an interview is informed by its participants' subjectivities. As a co-authored document, an oral history interview's topics and agenda are the product of co-authorship between interviewers and narrators.[44] In "Living Voices: The Oral History Interview as Dialogue and Experience", Alessandro Portelli describes the oral history interview as an exchange between two people with different, legitimate agendas. "What the researcher is interested in hearing is not necessarily what the narrator is interested in telling," Portelli writes.[45] He adds:

> Often we interview people who have not had a chance to speak about themselves and be heard before; thus, they seize the opportunity not only to answer our questions but also to volunteer stories of their own. Our task is not merely to extract information, but to open up narrative spaces.[46]

This negotiation yields a document that is co-authored by those involved in its production and is unique to its participants.

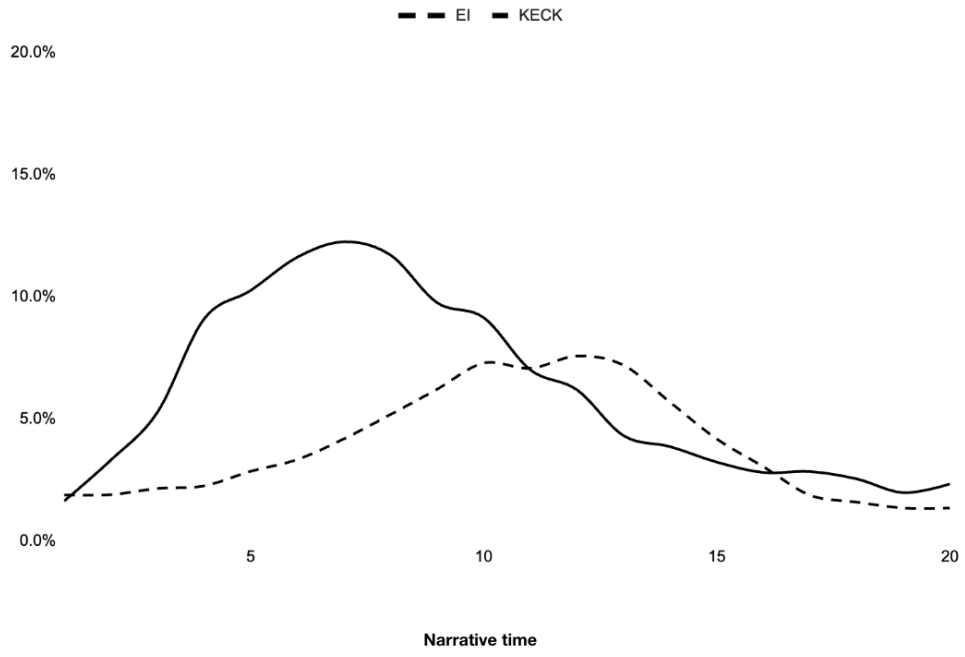
As one might expect, individual narrative time arcs in the corpus are highly varied. As I discussed earlier, I plotted each topic probability over 20 sequential interview segments. For example, Figure 8 shows the probability of the topic “Voyage” throughout the duration of three distinct interviews.

Figure 8: “Voyage” probability over narrative time



Yet, at the aggregate level, patterns emerge. Figure 9 shows the average probability of the topic "Voyage" throughout the KECK and EI series. And, most pertinent to the question at hand, immigration-related topics in KECK series interviews are, on average, 7.3% more prominent relative to EI series (Figure 7). Given the known variables that might influence the interview's agenda, which are actually driving these generalizable differences?

Figure 9: "Voyage" probability over narrative time, averaged by series



Perhaps one might hypothesize that the main drivers of average differences in the EI and KECK series are differences in each series' composition. For example, if one series' interviews are longer on average, it is reasonable to assume that participants had more air time to cover topics beyond their immigration experience. In addition, one might expect that other demographic features, such as gender or country of origin, could impact the overall topic structure. For example, though all

narrators relate their stories to Ellis Island and America, the push and pull factors leading narrators to America tend to vary by country of origin. This variance is significant enough that it can appear in topics. For example, Topic 45 (Appendix B), defined by the terms “women, men, killed, turks, police”, describes violence specific to Anatolia and the Balkans that would not reasonably be expected to occur in interviews with Irish immigrants, for example.

However, despite differences in corpus size, EI and KECK have roughly somewhat compositions (summarized in Table 3, Table 4, and Table 5). For example, narrator gender and country of origin are comparable, and in both series.

Table 3: EI vs. KECK interviews

Series	EI	KECK
Interviews	1,498	188
Interviewers	25	11
Countries represented	63	36
% Female	59%	61%
Average duration	1:05:25	0:47:39

Table 4: Interviews by narrator country of origin and series, EI vs. KECK

	Absolute		(%)	
Country of Origin	EI	KECK	EI	KECK
Italy	213	27	17%	16%
Poland	139	13	11%	7%
Germany	123	10	10%	6%
Russia	102	24	8%	14%
Ireland	56	5	4%	3%
Hungary	48	11	4%	6%
Austria	45	5	4%	3%
Turkey	41	6	3%	3%
England	36	9	3%	5%
Greece	29	5	2%	3%
Other	417	59	33%	34%

Table 5: Top interviewers by series

EI Interviewers	Abs	(%)
Janet Levine	805	64%
Paul Sigrist	362	29%
Kevin Daley	14	1%
Elysa Matsen	12	1%
Roger Herz	8	1%
Other	48	5%

KECK Interviewers	Abs	Share
Nancy Dallett	51	29%
Debby Dane	36	21%
Dana Gumb	28	16%
Edward Applebome	24	14%
Debra Allee	13	7%
Other	22	13%

However, one dimension through which EI and KECK differ more measurably is average interview duration. EI interviews are, on average, almost 15 minutes longer than KECK interviews. Longer interviews can allow for a greater number of topics to be discussed, lessening the percentage of each interview that is specific to immigration experience. Additionally, the EI series is dominated by fewer interviews.

To more directly understand the relationship between these known variables and the proportion of each series that is related to immigration, I ran a multiple linear regression analysis.[v] To perform this analysis, I assumed that immigration-related topic share is dependent on the following variables: series, narrator gender, interviewer, and narrator country of origin (Table 7). The following variables were used as controls: Series = KECK, Interviewer = Janet Levine, Country of Origin = Italy.

[v] Multiple linear regression is a type of linear regression that uses two or more independent variables to predict a single dependent variable.

Table 7: Multiple linear regression results

	Coefficients	Standard Error	P-value
Intercept	0.22	0.01	0.00
Duration	-0.04	0.05	0.40
Series: EI	-0.07	0.01	0.00
Narrator Gender: F	-0.01	0.00	0.00
Interviewer: Paul Sigrist	0.02	0.00	0.00
Interviewer: Nancy Dallett	0.00	0.01	0.95
Interviewer: Debby Dane	0.00	0.01	0.96
Interviewer: Other	0.02	0.00	0.00
Country of Origin: Russia	0.01	0.00	0.00
Country of Origin: Poland	0.01	0.00	0.00
Country of Origin: Ireland	0.02	0.00	0.00
Country of Origin: Hungary	-0.01	0.00	0.07
Country of Origin: Other	0.02	0.00	0.00

Of all variables considered, eight show statistical significance ($p \leq 0.05$): Series = EI, Narrator Gender = F, Interviewer = Paul Sigrist or Other, and Country of Origin = Russia, Poland, Ireland, or Other. Of the statistically significant variables, Series: EI has the most substantial coefficient at 0.07. This result suggests that, on average, an interview in the EI series spends about 7% less time on immigration compared to the KECK series, the reference category. In more concrete terms, this 7.00% reduction over the course of an hour of conversation is means 4.2 minutes fewer minutes spent on immigration. Similarly, an interview with a female narrator results in a 1% reduction compared to an interview with a male narrator.

Topic management in oral history

The multiple regression results suggest that compositional effects are important, and that series membership is the single most important variable in understanding the proportion of each interview related to immigration. But how are these differences enacted in the interview? How are topics being produced and managed differently in the EI series? Who controls the conversation?

Portelli asserts that "...the interview is two things at once: a tool for research, and the opening of a narrative space. And, given that ultimate control is in the stories narrators wish to tell, often the agendas that prevail are the narrators'."[47] It is seductive to imagine that a narrator's agenda is the one that most often prevails; this thinking certainly lessens the burden on oral historians. Reading through Ellis Island Oral History interviews, one could easily adopt this stance. Consider EI-306, an interview conducted by Janet Levine with Rose Halpern, an immigrant to the United States from Ukraine.[48] Levine asks Halpern about her grandparents, to which Halpern responds with a story about a pogrom. Levine attempts to turn the interview again toward Halpern's grandparents.

> **LEVINE:** Now, did you have grandparents?
>
> **HALPERN:** Yes, yes, yes. In Russia they warned, in Chesinifka where we lived. But my mother's, my mother's parents lived away, Terlitza, Terlitza. They lived in Terlitza, maybe a couple of hours, maybe about two hundred miles away from the Chesinifka. They lived there, my grandparents.
>
> **LEVINE:** Did you see them much?
>
> **HALPERN:** Yes. I'll come to that, darling. I'll come to my grandparents. When I first start you [sic] my experience, the first bad, very bad experience was when we had to leave there, Chesinifka, okay?

Halpern repeatedly resists Levine's attempts to manage topics.

Some might interpret Levine's attempts to shape the agenda as indicative of her power as an interviewer. Eva M. McMahan, for example, asserts that the balance of power in an interview setting tends toward the interviewer, who ultimately has more control over the interview's topic structure.[49] Mario Varicchio tends to share this vantage point in "Golden Door Voices: Towards a Critique of the Ellis Island Oral History Project". Though Varricchio acknowledges that the EI series interviewers maintain some flexibility with narrators, he argues that they tend to silence and foreclose topics in service of topics in their interview agendas.

How Sigrist and Levine conceptualized their roles as interviewers is more complicated than that. EI-469 is a recorded reflection by Sigrist on his early interviews. In this recording, Sigrist comments on the nature of topic management in Ellis Island interviews. Sigrist references EI-I, an interview with Italian immigrant Charles Crimi and the first interview he conducted for the EI series:

> Anyway, Mr. Crimi was quite a character, as I said. He had his own agenda. I began the interview by asking him what his name was and his birthday and he simply ignored me and, beginning with a line, and I'm paraphrasing, this may not be exactly right, "Let's brush away the cobwebs of history" and on he went.[50]

Referencing an interview where the narrator never quite arrived in America (the tape ran out before the narrator arrives in America in his own testimony), Sigrist remarked:

> I learned a lesson during that interview. If you've got someone who, who has a memory and the ability and the eloquence to tell a story such as Mrs. Bacos did, that you just simply let them do it and you accommodate yourself to them.[51]

It may not be useful to make generalizations about which parties tend to hold the most power in an interview setting. Assigning control to the narrator falsely reinforces that their story is inherent and that other factors are inert. Moreover, it neglects the reality that there are generalizable differences

in topic mixes between EI series. Furthermore, it fails to explain why Ellis Island should be a topic taking up double-digit percentages of the collections' interviews at all; most immigrants spent no more than a day on Ellis Island.[52] However, assigning control to the project and its interviewers neglects the vast variety of narratives that a project collects. Topic management between interviewers and narrators is situational, relational, and actively negotiated throughout the duration of an interview. But how can we observe or measure this negotiation?

In "A Conversation Analytic Approach to Oral History Interviewing", Eva M. McMahan describes in more concrete terms how topics are managed within an interview setting. By examining elite oral history interviews, McMahan develops a conversation analytic approach for understanding how questions and answers are used to explicate historical events.[53] McMahan argues that although question and answer form the basic unit of an interview, the third turn response, or question-answer-response pattern, is how topics are extended and ultimately managed.

Using McMahan's work as a framework, I analyzed the diarized Ellis Island corpus in two ways. First, I examined who *introduces* immigration-related topics in each interview. Second, I examined how interviewers and narrators either *extend* or *change* topics based on their responses to the speaker preceding them.

Table 8 demonstrates how I implemented this approach. Each row represents a turn in the conversation and is assigned to either the interviewer or narrator ("Speaker Label"). From each turn, I extracted the most probable topic ("Max Topic"). Each "Max Topic" is represented by a topic as numbered in Appendix B. For each row, I tracked whether or not the most probable topic was being introduced into the interview for the first time ("Introduction?") by comparing the row to the

rows above it. Then, for each turn, I tracked whether or not each speaker's most probable topic was congruent with the turn that preceded it ("Extend?").

Table 8: Example of a diarized and annotated corpus

Interview #	Speaker Label	Text	Max Topic	Introduction?	Extend?
132	Interviewer	Lorem ipsum dolor sit amet.	10	1	0
132	Narrator	Excepteur sint occaecat cupidatat non proident.	10	0	1
132	Interviewer	Sunt in culpa qui officia deserunt mollit anim id.	10	0	1
132	Narrator	Quis autem vel eum iure reprehenderit qui.	26	1	0
132	Interviewer	Nam libero tempore, cum soluta nobis.	10	0	0

This table represents a necessary simplification of McMahan's question-answer-response framework, as tracking conversation at a three-segment level of detail would require more sophisticated interview coding. In addition, this table does not account for extensions that the topic model cannot classify. For example, sometimes interviewers extend or change topics without using the language of the topic itself. For example, an interviewer may want to extend a narrator's reflection on Ellis Island (Topic 8) by saying "Great" (Topic 22), instead of "Great. Tell me more about Ellis Island" (Topic 8). This type of extension would not be properly captured. However, at a high level, this table is useful for understanding how topics are introduced and how turns in conversation align or misalign with the turns preceding them. In addition, these limitations apply to the entire Ellis Island corpus, meaning that EI and KECK can still be reasonably compared.

Topic introduction

Table 9: Topic introductions by topic type, series, and speaker

All topics	Interviewer	Narrator	Var
EI	47%	53%	-6%
KECK	50%	50%	0%
Immigration-related topics	Interviewer	Narrator	Var
EI	60%	40%	20%
KECK	65%	35%	30%
Other topics	Interviewer	Narrator	Var
EI	45%	55%	-10%
KECK	48%	52%	-4%

Table 9 summarizes the share of introductions in the corpus by topic type, series and speaker. In both series, interviewers and narrators each introduce roughly half of the interview's topics. When considering immigration-specific topics in isolation, EI interviewers introduce 20% more immigration-related topics than narrators. Similarly, KECK interviewers introduce 30% more immigration-related topics than narrators. Relative to all other topics, interviewers are more likely than narrators to introduce immigration-related topics in both series.

But how do narrators respond to immigration topics that interviewers introduce? How do interviewers respond to narrators who veer outside of their immigration-focused agenda?

Topic extension

Table 10: Topic extensions by topic type, series, and speaker

	EI Series		KECK Series	
	Interviewer	Narrator	Interviewer	Narrator
Total extensions	17,451	25,480	1,285	2,207
Immigration-related extensions	4,892	7,657	664	748
Immigration-related extensions (%)	28%	30%	52%	34%

The above table counts all instances of topic extension by topic type, series, and speaker. In both the EI and KECK series, immigration-related extensions by narrators are comparable and account for roughly 30–34% of all topic extensions. However, the share of immigration-related extensions is less comparable between the series when examining interviewers. KECK interviewer extensions are related to immigration topics 52% of the time, whereas only 28% of EI interviewer extensions are related to immigration. Interestingly, narrator behavior is comparable in both series; differences in topic duration appear to be driven by interviewers alone.

Topic introduction and extension data make visible differences in how the EI and KECK interviewers manage interviews. Namely, KECK interviewers are more likely to introduce and extend topics related to immigration.

These results demonstrate that the difference between the KECK series and EI series is smaller than one might imagine based on the distinctions “immigration experience only” and “life narrative” offered by Sigrist and Levine. Instead, the EI series and its questionnaire might codify existing practices from the KECK series that preceded it.

Considering all variables, the single strongest predictor of how related the interview is to immigration experience is which series the interview belongs to. A project's goals are expressed through a number of factors, including the conditions and training that inform how an interviewer conducts an interview. The EI series interviewers openly permitted themselves to explore topics beyond immigration, which is expressed in the series' topic mix. Though little is publicly known about the KECK series, it is reasonable to assume that interviewers were shaped by the fact that they were commercially contracted to produce specific outcomes for the Ellis Island Immigration Museum. More specifically, interviewers were tasked with producing audio clips related to the site itself. Ultimately, this analysis reinforces that project design matters. Though it is clear that interviewers operationalize the series' designs and objectives, it is less clear which policies or procedures guide their decision-making. Further analysis in this area might examine how other oral history projects' training procedures influence conformance to specific agendas.

There are many more questions one might explore with the dataset I produced to conduct this analysis. Does the EI interview guide's chronology influence an interview's chronology? Are male and female narrators asked similar questions? What can be said about an individual interviewer's interview style?

Looking forward

This work demonstrates only a handful of cases in which NLP serves an oral historian's work. The universe of NLP tools at a researcher's disposal is vast, and at the time of writing, it seems to grow broader every month. Moreover, these tools are beginning to be packaged in ways that are much more accessible than they were at the outset of this work, and in some cases, they do not require any code to be operated. When I began working with the Ellis Island Oral History project in

September 2021, OpenAI's GPT suite was a limited, access-controlled platform for researchers. And now, at the conclusion of this work in 2023, GPT has become a household name via its user-friendly derivative chatbot, ChatGPT.

Though NLP tools are becoming more accessible to the public, there are still barriers that oral historians may face in integrating them into their practice. Without a theoretical foundation in NLP, it may be difficult for oral historians to appreciate the inner workings of the tools at their disposal. Some measure of knowledge must be cultivated to thoughtfully integrate these tools—either through a practitioner's independent education or interdisciplinary collaboration. As tools become easier to use, coding is becoming less important, though some technical context is imperative for thoughtful interpretation. Though an oral historian may not know the details of a microphone's circuit board, they certainly know how to effectively use a microphone.

With each advance in NLP, public conversation seems to converge around two poles: at one end, tech-determinist optimism, and on the other, fear. These reactions are rooted in real opportunities and risks, but their intensity tends to conceal a more complex reality. Oral historians, having thoughtfully integrated many tools, including the tape recorder, digital video, Zoom, and the internet, into their practices, should consider NLP tools as they would any other technology: simply more tools with opportunities, risks, and constraints.

Endnotes

[1] Lucie Kučerová, “Editing Procedures in Studs Terkel’s Oral Histories,” *Discourse and Interaction* 4 (January 4, 2011): 51–66.

[2] Varricchio, Mario. “Golden Door Voices: Towards a Critique of the Ellis Island Oral History Project.” *Oral History Forum d’Histoire Orale* 31 (January 1, 2011): 4.

[3] Mark Anthony Hoffman et al., “The (Protestant) Bible, the (Printed) Sermon, and the Word(s): The Semantic Structure of the Conformist and Dissenting Bible, 1660–1780,” *Poetics* 68 (June 2018): 89–103, https://doi.org/10.1016/j.poetic.2017.11.002.

[4] Wenyi Shang and Ted Underwood, “Improving Measures of Text Reuse in English Poetry: A TF–IDF Based Method,” *Lecture Notes in Computer Science*, March 17, 2021, 469–77, https://doi.org/10.1007/978-3-030-71292-1_36.

[5] Jordan Abel, “Empty Spaces,” *Canadian Literature*, no. 230-1 (2016), https://doi.org/10.14288/cl.v0i230-1.189158.

[6] Yale University Fortunoff Archive, “Let Them Speak,” lts.fortunoff.library.yale.edu, accessed April 30, 2023, https://lts.fortunoff.library.yale.edu/.

[7] Stephanie Dinkins, “Not the Only One,” STEPHANIE DINKINS, accessed April 30, 2023, https://www.stephaniedinkins.com/ntoo.html.

[8] Holly Anne Rieping, “Audio Segmenting and Natural Language Processing in Oral History Archiving,” dspace.mit.edu, February 1, 2022, https://dspace.mit.edu/handle/1721.1/143185.

[9] Steven E. Jones, *The Emergence of the Digital Humanities* (Taylor & Francis, 2013).

[10] Anna Sheftel and Stacey Zembrzycki, “Slowing down to Listen in the Digital Age: How New Technology Is Changing Oral History Practice,” *The Oral History Review* 44, no. 1 (2017): 94–112, https://doi.org/10.1093/ohr/ohx016.

[11] “Principles and Best Practices,” Oral History Association, October 2009, https://oralhistory.org/about/principles-and-practices-revised-2009/.

[12] “Best Practices,” Oral History Association, accessed April 30, 2023, https://oralhistory.org/best-practices/.

[13] Alessandro Portelli, “Living Voices: The Oral History Interview as Dialogue and Experience,” *The Oral History Review* 45, no. 2 (2018): 239–48, https://doi.org/10.1093/ohr/ohy030.

[14] Molly Roberts, Brandon Stewart, and Dustin Tingley, “STM,” Structuraltopicmodel.com, 2018, https://www.structuraltopicmodel.com/.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3 (March 1, 2003): 993–1022, https://doi.org/10.5555/944919.944937.

[16] National Park Service, “Ellis Island Chronology,” Ellis Island, March 17, 2022, https://www.nps.gov/elis/learn/historyculture/ellis-island-chronology.htm.

[17] The Statue of Liberty—Ellis Island Foundation, Inc, “Mission and History,” March 16, 2020, https://www.statueofliberty.org/foundation/mission-history/.

[18] The Statue of Liberty—Ellis Island Foundation, Inc., “Ellis Island,” February 4, 2020, https://www.statueofliberty.org/ellis-island/.

[19] The Statue of Liberty—Ellis Island Foundation, Inc, “Mission and History”.

[20] Janet Levine and Paul Sigrist, “Description of Individual Interview Series That Comprise the Ellis Island Oral History Project Interview Collection,” June 2005.

[21] Janet Levine and Paul Sigrist, “Description of Individual Interview Series That Comprise the Ellis Island Oral History Project Interview Collection.”

[22] Ken Plummer, *Documents of Life: An Introduction to the Problems and Literature of a Humanistic Method* (London: George Allen and Unwin, 1983), 185–203, as cited in Mary Chamberlain, “Narrative Theory,” in *Handbook of Oral History* (Lanham, MD: AltaMira Press, 2006), 336–65.

[23] Alessandro Portelli, “Uchronic Dreams: Working Class Memory and Possible Worlds,” Oral History 16, no. 2 (1988): 46–56, https://www.jstor.org/stable/40179011, as cited in Mary Chamberlain, “Narrative Theory”.

[24] National Park Service, “Oral Histories for Your Classroom,” 2016, https://www.nps.gov/elis/learn/education/classrooms/oral-histories.htm.

[25] Peter Morton Coan, *Toward a Better Life: America’s New Immigrants in Their Own Words from Ellis Island to the Present* (Prometheus Books, 2011), as cited in Varricchio, Mario. “Golden Door Voices: Towards a Critique of the Ellis Island Oral History Project.”

[26] https://heritage.statueofliberty.org/oral-history-library

[27] Phone conversation with Matthew Housch, librarian and historian at Ellis Island (National Park Service) on May 27, 2022.

[28] Mary Kelly, EI-613, interview by Paul E. Sigrist, *National Park Service*, May 2, 1995.

[29] W. Bruce Croft and John Lafferty, *Language Modeling for Information Retrieval* (Springer Science & Business Media, 2013), IX.

[30] Python scripting supported by Kayla Pandza

[31] Monika M. Wahi et al., “Reducing Errors from the Electronic Transcription of Data Collected on Paper Forms: A Research Data Case Study,” *Journal of the American Medical Informatics Association* 15, no. 3 (May 1, 2008): 386–89, https://doi.org/10.1197/jamia.m2381.

[32] The American Folklore Society, “Position Statement: Compensation,” accessed April 30, 2023, https://americanfolkloresociety.org/our-work/position-statement-compensation/, as cited in Oral History Association, “OHA Statement on Freelance, Independent, and Contract Oral History Labor,” accessed April 30, 2023, https://oralhistory.org/oha-statement-on-freelance-independent-and-contract-oral-history-labor/.

[33] OpenAI, “Pricing,” n.d., https://openai.com/pricing.

[34] OpenAI, “Privacy Policy,” April 27, 2023, https://openai.com/policies/privacy-policy.

[35] David M Blei, Andrew Y Ng, and Michael I Jordan, “Latent Dirichlet Allocation.”

[36] Ismail Harrando, Pasquale Lisena, and Raphael Troncy, “Apples to Apples: A Systematic Evaluation of Topic Models,” ACLWeb (Held Online: INCOMA Ltd., September 1, 2021), https://aclanthology.org/2021.ranlp-1.55/, as cited in Maria Antoniak, “Topic Modeling for the People,” July 27, 2022, https://maria-antoniak.github.io/2022/07/27/topic-modeling-for-the-people.html.

[37] Antonina Iozzia, EI-477, interview by Paul E. Sigrist, *National Park Service*, June 8, 1994.

[38] Nancy Fuschetti, KECK-42, interview by Edward Applebome, *National Park Service*, March 6, 1906.

[39] Social Security Administration, “Top Names of the 1990s,” 2010, https://www.ssa.gov/oact/babynames/decades/names1990s.html.

[40] Jonathan Chang et al., “Reading Tea Leaves: How Humans Interpret Topic Models,” *Advances in Neural Information Processing Systems* 22 (December 7, 2009), https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

[41] Maria Antoniak, David Mimno, and Karen Levy, “Narrative Paths and Negotiation of Power in Birth Stories,” *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–27, https://doi.org/10.1145/3359190.

[42] Baylor University Institute for Oral History, “Time Coding & Indexing Oral Histories,” 2014, https://www.baylor.edu/content/services/document.php/66437.pdf.

[43] Anna Perlstein, KECK-65, interview by Debby Dane, *National Park Service*, October 24, 1985.

[44] Alessandro Portelli, “Living Voices: The Oral History Interview as Dialogue and Experience”.

[45] Alessandro Portelli, “Living Voices: The Oral History Interview as Dialogue and Experience”.

[46] Alessandro Portelli, “Living Voices: The Oral History Interview as Dialogue and Experience”.

[47] Alessandro Portelli, “Living Voices: The Oral History Interview as Dialogue and Experience”.

[48] Rose Halpern, EI-306, interview by Janet Levine, *National Park Service*, April 28, 1993.

[49] Eva M. McMahan, “A Conversation Analytic Approach to Oral History Interviewing,” in *Handbook of Oral History* (Lanham, MD: AltaMira Press, 2006), 336–83.

[50] Paul Sigrist, EI-469, National Park Service, n.d.

[51] Paul Sigrist, EI-469

[52] The Statue of Liberty—Ellis Island Foundation, Inc., “Overview and History”.

[53] Eva M. McMahan, “A Conversation Analytic Approach to Oral History Interviewing”.

Bibliography

Abel, Jordan. “Empty Spaces.” *Canadian Literature*, no. 230-1 (2016). https://doi.org/10.14288/cl.v0i230-1.189158.

Abrams, Lynn. *Oral History Theory*. Routledge, 2016.

Antoniak, Maria. “Topic Modeling for the People,” July 27, 2022. https://maria-antoniak.github.io/2022/07/27/topic-modeling-for-the-people.html.

Antoniak, Maria, David Mimno, and Karen Levy. “Narrative Paths and Negotiation of Power in Birth Stories.” *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–27. https://doi.org/10.1145/3359190.

Baylor University Institute for Oral History. “Time Coding & Indexing Oral Histories,” 2014. https://www.baylor.edu/content/services/document.php/66437.pdf.

Blei, David M, Andrew Y Ng, and Michael I Jordan. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (March 1, 2003): 993–1022. https://doi.org/10.5555/944919.944937.

Chamberlain, Mary. “Narrative Theory.” In *Handbook of Oral History*, 336–65. Lanham, MD: AltaMira Press, 2006.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. “Reading Tea Leaves: How Humans Interpret Topic Models.” *Advances in Neural Information Processing Systems 22* (December 7, 2009). https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

Dinkins, Stephanie. “Not the Only One.” STEPHANIE DINKINS. Accessed April 30, 2023. https://www.stephaniedinkins.com/ntoo.html.

Fuschetti, Nancy. KECK-42. Interview by Edward Applebome. *National Park Service*, March 6, 1906.

Halpern, Rose. EI-306. Interview by Janet Levine. *National Park Service*, April 28, 1993.

Harrando, Ismail, Pasquale Lisena, and Raphael Troncy. “Apples to Apples: A Systematic Evaluation of Topic Models.” ACLWeb. Held Online: INCOMA Ltd., September 1, 2021. https://aclanthology.org/2021.ranlp-1.55/.

Hoffman, Mark Anthony, Jean-Philippe Cointet, Philipp Brandt, Newton Key, and Peter Bearman. “The (Protestant) Bible, the (Printed) Sermon, and the Word(S): The Semantic Structure of the Conformist and Dissenting Bible, 1660–1780.” *Poetics* 68 (June 2018): 89–103. https://doi.org/10.1016/j.poetic.2017.11.002.

Housch, Matthew. Private conversation with Christopher M. Pandza. Phone, May 27, 2022.

Iozzia, Antonina. EI-477. Interview by Paul E. Sigrist. *National Park Service*, June 8, 1994.

Jones, Steven E. *The Emergence of the Digital Humanities*. Taylor & Francis, 2013.

Kelly, Mary. EI-613. Interview by Paul E. Sigrist. *National Park Service*, May 2, 1995.

Kučerová, Lucie. “Editing Procedures in Studs Terkel’s Oral Histories.” *Discourse and Interaction* 4 (January 4, 2011): 51–66.

Levine, Janet, and Paul Sigrist. “Description of Individual Interview Series That Comprise the Ellis Island Oral History Project Interview Collection,” June 2005.

McMahan, Eva M. “A Conversation Analytic Approach to Oral History Interviewing.” In *Handbook of Oral History*, 336–83. Lanham, MD: AltaMira Press, 2006.

National Park Service. “Ellis Island Chronology.” Ellis Island, March 17, 2022. https://www.nps.gov/elis/learn/historyculture/ellis-island-chronology.htm.

———. “Oral Histories for Your Classroom,” 2016. https://www.nps.gov/elis/learn/education/classrooms/oral-histories.htm.

OpenAI. “Pricing,” n.d. https://openai.com/pricing.

———. “Privacy Policy,” April 27, 2023. https://openai.com/policies/privacy-policy.

Oral History Association. “Best Practices.” Accessed April 30, 2023. https://oralhistory.org/best-practices/.

———. “OHA Statement on Freelance, Independent, and Contract Oral History Labor.” Accessed April 30, 2023. https://oralhistory.org/oha-statement-on-freelance-independent-and-contract-oral-history-labor/.

Perlstein, Anna. KECK-65. Interview by Debby Dane. *National Park Service*, October 24, 1985.

Peter Morton Coan. *Toward a Better Life: America’s New Immigrants in Their Own Words from Ellis Island to the Present*. Prometheus Books, 2011.

Plummer, Ken. *Documents of Life: An Introduction to the Problems and Literature of a Humanistic Method*. London: George Allen and Unwin, 1983.

Portelli, Alessandro. “Living Voices: The Oral History Interview as Dialogue and Experience.” *The Oral History Review* 45, no. 2 (2018): 239–48. https://doi.org/10.1093/ohr/ohy030.

———. “Uchronic Dreams: Working Class Memory and Possible Worlds.” *Oral History* 16, no. 2 (1988): 46–56. https://www.jstor.org/stable/40179011.

Oral History Association. “Principles and Best Practices,” October 2009. https://oralhistory.org/about/principles-and-practices-revised-2009/.

Rieping, Holly Anne. “Audio Segmenting and Natural Language Processing in Oral History Archiving.” dspace.mit.edu, February 1, 2022. https://dspace.mit.edu/handle/1721.1/143185.

Roberts, Molly, Brandon Stewart, and Dustin Tingley. “STM.” Structuraltopicmodel.com, 2018. https://www.structuraltopicmodel.com/.

Shang, Wenyi, and Ted Underwood. “Improving Measures of Text Reuse in English Poetry: A TF–IDF Based Method.” *Lecture Notes in Computer Science*, March 17, 2021, 469–77. https://doi.org/10.1007/978-3-030-71292-1_36.

Sheftel, Anna, and Stacey Zembrzycki. “Slowing down to Listen in the Digital Age: How New Technology Is Changing Oral History Practice.” *The Oral History Review* 44, no. 1 (2017): 94–112. https://doi.org/10.1093/ohr/ohx016.

Sigrist, Paul. EI-469. *National Park Service*, n.d.

Social Security Administration. “Top Names of the 1990s,” 2010. https://www.ssa.gov/oact/babynames/decades/names1990s.html.

Tang, Jian, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. “Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis.” PMLR, January 27, 2014. https://proceedings.mlr.press/v32/tang14.html.

The American Folklore Society. “Position Statement: Compensation.” Accessed April 30, 2023. https://americanfolkloresociety.org/our-work/position-statement-compensation/.

The Statue of Liberty—Ellis Island Foundation, Inc. “Ellis Island,” February 4, 2020. https://www.statueofliberty.org/ellis-island/.

———. “Mission and History,” March 16, 2020. https://www.statueofliberty.org/foundation/mission-history/.

———. “National Immigration Museum,” March 16, 2020. https://www.statueofliberty.org/ellis-island/national-immigration-museum/.

———. “Overview and History,” March 4, 2020. https://www.statueofliberty.org/ellis-island/overview-history/.

Varricchio, Mario. “Golden Door Voices: Towards a Critique of the Ellis Island Oral History Project.” *Oral History Forum d’Histoire Orale* 31 (January 1, 2011): 4.

W. Bruce Croft, and John Lafferty. *Language Modeling for Information Retrieval*. Springer Science & Business Media, 2013.

Wahi, Monika M., David V. Parks, Robert C. Skeate, and Steven B. Goldin. “Reducing Errors from the Electronic Transcription of Data Collected on Paper Forms: A Research Data Case Study.” *Journal of the American Medical Informatics Association* 15, no. 3 (May 1, 2008): 386–89. https://doi.org/10.1197/jamia.m2381.

Yale University Fortunoff Archive. “Let Them Speak.” lts.fortunoff.library.yale.edu. Accessed April 30, 2023. https://lts.fortunoff.library.yale.edu/.

Appendix A: Interview series summary

Adapted from "Description of Individual Interview Series that Comprise the Ellis Island Oral History Project Interview Collection" by Paul Sigrist and Janet Levine.

Series	Count	Short description
Ellis Island Series 1990 - 2009	1400+	Conducted mainly by Paul Sigrist and Janet Levine, this collection embraced a "life narrative" format vs. earlier "immigration experience only" interviews. Introduction of the "Oral History Form" to conduct interviews.
Kate Moore Series 1993-1994	79	Interviews conducted by Kate Moore in the West and Midwest. Like the Ellis Island series, follows a "life narrative" format.
Dallett/Phillips Series 1989	60	Interviews conducted by Nancy Dallett and Andrew Phillips in the West and Midwest. Follow an "immigration experience only" format. *Note that Dallett was also an interviewer for the KECK series.
Treasures from Home Series 1989	9	Conducted for the exhibit design and installation for the Ellis Island Immigration Museum. Interviews about specific objects in the exhibits.
Guggenheim Productions Incorporated Series 1985-1987	27	Conducted by Charles Guggenheim for the purpose of extracting audio quotations to create voiceover for the film *Island of Hope, Island of Tears*. *Several of the interviewees had initially been interviewed for the KECK Series.
KECK Series 1985-1986	200	Produced to provide audio quotations for the Ellis Island Immigration Museum. Conducted by an urban planning firm, considered to be "immigration experience only" interviews. Interviewees were often chosen based on their extended detentions on the island and the depth of information they were able to relay in letters about being detained. *Several interviewees were later re-interviewed by Charles Guggenheim for his documentary film *Island of Hope, Island of Tears*
Morrison Zabusky Series 1974-1978	130	Conducted by Joan Morrison and Charlotte Zabusky to write the book "The American Mosaic". Not conducted by Ellis Island, but donated to Ellis Island in 1991. These interviews deal with the immigration experience and adaptation to life in America, but generally not life prior to arrival. Transcript quality varies and sometimes leave out entire segments of the interview.
All Other Series 1975 -	68	All other interviews that do not fall under the set above. These interviews are primarily homemade tapes donated to the Project. Only exist as audio recordings.
National Park Service Series 1973 - 1988	161	Interviews originating from Margo Nash, who was hired to head the Ellis Island Oral History Project in 1973. Includes interviews with those beyond who immigrated from Ellis Island. Mostly "immigration experience only" style interviews.

Appendix B: Topics and topic classifications

Topic	Highest probability words	Classification	Higher-order classification	Immigration-relation
1	thank_you, speaking, anything_else, glad, signing	Procedural - Thank you	Procedural	Other
2	national_park_service, giving, present, information, note	Procedural - Note	Procedural	Other
3	nice, trouble, lots, treated, big	Noise	Noise	Other
4	story, stories, telling, ladies, talk	Storytelling	Storytelling	Other
5	people, place, places, homes, poor	People & Places	People & Places	Other
6	mother, care, father, died, wanted	Mother	Family	Other
7	run, o'clock, ran, till, stop	Altercations	Altercations	Other
8	ellis_island, new_york, statue_of_liberty, coming, oral_history	Ellis Island	Ellis Island	Immigration-related
9	age, spell, born, maiden_name, birth_date	Biographical information	Biographical information	Other
10	ship, class, deck, voyage, cabin	Voyage - Ship	Voyage	Immigration-related
11	school, high, teacher, grade, class	School	School	Other
12	daughter, lives, sons, daughters, ago	Children 1	Family	Other
13	christmas, special, tree, holidays, holiday	Holidays	Ethnic Origin & Customs	Other
14	hours, union, trade, boss, labor	Unions	Labor	Other
15	italian, italians, words, boy, talk	Italian 2	Ethnic Origin & Customs	Other
16	armenian, turkish, means, saved, country	Armenian	Ethnic Origin & Customs	Other
17	bread, cooking, big, cooked, oven	Baking	Domestic Labor	Other
18	real, ways, swedish, guess, customs	Customs	Ethnic Origin & Customs	Other
19	read, write, yiddish, books, reading	Reading & Writing	Reading & Writing	Other
20	cold, south, north, hot, warm	Seasons	Seasons	Other
21	greek, restaurant, orthodox, community, stay	Greek 1	Ethnic Origin & Customs	Other
22	good, great, happy, feel, question	Affirmation	Procedural	Other
23	home, brooklyn, norway, carpenter, called	Home	House & Home	Other
24	england, london, british, called, eventually	English	Ethnic Origin & Customs	Other
25	day, days, saint, big, call	Saint Days	Ethnic Origin & Customs	Other
26	american, papers, citizen, paper, citizenship	Citizenship	Citizenship	Immigration-related
27	business, bought, buy, sell, sold	Commerce	Commerce	Other
28	cut, watch, wine, drink, pull	Making 1	Making & Craft	Other
29	children, life, feel, proud, names	Children 2	Family	Other
30	passport, wagon, border, warsaw, visa	Voyage	Voyage	Immigration-related
31	holland, dutch, folks, fairly, guess	Dutch	Ethnic Origin & Customs	Other
32	war, army, world_war, service, broke	Military	Military Service	Other
33	play, played, fun, playing, games	Leisure	Leisure	Other
34	beautiful, girl, woman, child, aunt	Women & Girls	Women & Girls	Other
35	number, call, telephone, phone, health	Telephone	Communication	Other
36	visit, guess, year, california, denmark	Procedural - Visit	Procedural	Other
37	germany, hitler, berlin, left, anymore	Hitler's Germany	War & Violence	Other
38	give, bad, god, sick, gave	Descriptive - Negative	Fear & Misfortune	Other
39	ireland, quiet, county, lovely, country	Irish	Ethnic Origin & Customs	Other
40	mine, coal, stove, heat, iron	Coal & Heating	Coal & Heating	Other
41	hungary, hungarian, big, year, managed	Hungarian	Ethnic Origin & Customs	Other
42	make, made, piece, making, bring	Making 2	Making & Craft	Other
43	men, women, killed, happened, government	Violence	War & Violence	Other
44	fish, market, fruit, air, fresh	Markets	Food	Other
45	italy, country, sicily, live, die	Italian 1	Ethnic Origin & Customs	Other
46	jewish, jews, hebrew, synagogue, jew	Jewish	Ethnic Origin & Customs	Other
47	family, cousins, brothers, families, uncles	Extended Family	Family	Other
48	sing, music, dance, singing, dancing	Song & Dance	Leisure	Other
49	talking, age, turned, turn, pennsylvania	Procedural - Place	Procedural	Other
50	milk, horse, horses, cows, eggs	Livestock	Agriculture	Other
51	clothes, shoes, made, dress, wear	Clothing	Clothing	Other

52	understood, pause, gonna, wait, alright	Procedural - Understood	Procedural	Other
53	father, knew, belgium, found, told	Father	Family	Other
54	friends, friend, group, social, party	Community	Community	Other
55	morning, week, hour, late, evening	Time of Day 1	Time of Day	Other
56	germans, camp, soldiers, war, happened	Holocaust	War & Violence	Other
57	job, office, jobs, company, quit	Office Jobs	Labor	Other
58	russia, russian, revolution, russians, means	Russian	Ethnic Origin & Customs	Other
59	ship, travel, port, traveling, traveled	Voyage - Ship	Voyage	Immigration-related
60	uncle, cousin, wrote, aunt, letter	Relatives in America	Relatives in America	Other
61	religious, religion, taught, strict, prayer	Prayer	Ethnic Origin & Customs	Other
62	town, born, birth, originally, prepared	Procedural - Birthplace	Birth & Birthplace	Other
63	store, shop, stores, grocery, butcher	Stores	Commerce	Other
64	country, fact, world, experience, felt	Reflection on Immigration	Reflection on Immigration	Immigration-related
65	city, station, railroad, bus, subway	Transportation	Transportation	Other
66	house, room, lived, big, kitchen	House	House & Home	Other
67	poland, polish, lived, words, means	Polish	Ethnic Origin & Customs	Other
68	village, greece, mountains, mountain, call	Greek 2	Ethnic Origin & Customs	Other
69	room, looked, waiting, told, asked	Arrival & Examination	Ellis Island	Other
70	show, picture, pictures, movies, showed	Movies	Leisure	Other
71	brother, sister, older, sisters, brothers	Siblings	Family	Other
72	speak, learned, learn, language, spoke	Language	Language	Other
73	land, potatoes, garden, grow, ground	Farming 1	Agriculture	Other
74	church, catholic, priest, mass, churches	Church	Ethnic Origin & Customs	Other
75	left, to_america, leave, in_america, leaving	Leaving for America	Leaving for America	Immigration-related
76	night, saturday, friday, dog, times	Time of Day 2	Time of Day	Other
77	put, big, top, clean, hand	Lighting	Making & Craft	Other
78	walk, door, walked, open, front	Walking	Transportation	Other
79	years_old, today, half, ago, year	Age	Age	Other
80	back, stayed, go_back, stay, went_back	Going Back	Going Back	Other
81	wife, college, hospital, doctor, company	College	School	Other
82	french, book, foreign, written, student	French	Ethnic Origin & Customs	Other
83	sweden, compared, country, guess, lived	Swedish	Ethnic Origin & Customs	Other
84	boat, trip, sick, boats, big	Ship	Voyage	Immigration-related
85	turkey, established, country, living, heard	Turkish	Ethnic Origin & Customs	Other
86	water, wood, fire, running, big	Wood	House & Home	Other
87	street, lived, moved, apartment, avenue	City	City	Other
88	married, husband, met, meet, wife	Spouse	Family	Other
89	eat, table, meat, dinner, coffee	Eating	Food	Other
90	money, dollars, pay, paid, gave	Money	Money	Other
91	car, island, hotel, cars, drive	Cars	Transportation	Other
92	farm, farmer, farmers, farms, farming	Farming 2	Agriculture	Other
93	united_states, new_york, decided, immigration, states	United States	Immigration	Immigration-related
94	parents, europe, austria, czechoslovakia, river	Parents	Ethnic Origin & Customs	Other
95	food, ate, eat, kosher, fed	Food	Food	Other
96	work, worked, working, factory, hard	Blue Collar	Labor	Other
97	hair, new_jersey, short, eyes, dark	Person	Physical Description	Other
98	wonderful, close, interesting, loved, enjoyed	Procedural - Close	Procedural	Other
99	mother's, grandmother, father's, grandfather, lived	Family	Family	Other
100	gas, big, made, called, place	Gas	Transportation	Other

Appendix C: Oral History Form

Revised by Paul E. Sigrist, Jr., March 1993, as cited in Varricchio.

THE START AND THE OLD COUNTRY

Good morning/afternoon, this is _____ for the National Park Service. Today is _____, the _____, and I"m in _____, at the home of _____, who came from _____ in _____ when he/she was _____ years old. Why don"t you begin by giving me your full name and date of birth, please.

What is your maiden name? Spell it, please.

Where were you born? Spell it, please.

What size town? Describe what the town looked like? What was the major industry?

Father"s name? (spell it if unusual) Occupation? Describe what he looked like. Describe his personality and temperament. Is there a story about your father that you associate with your childhood?

Mother"s name? (spell it if unusual) What was her maiden name? (spell it if unusual) Occupation, if any? Describe what she looked like. Describe her personality and temperament. What were her chores around the house? Is there a story about your mother that you associate with your childhood?

Name all brothers and sisters. (spell if unusual)

Describe your house. What kind of dwelling did you live in? How large? How many rooms? What was it made out of? How was it heated? Was there a garden? What did you grow? What kind of furniture did you have? Was it in or out of town? Did you keep animals? Who else lived in the building?

Who did the cooking in the family? What was your favourite food? Did you help cook? Describe the kitchen. What was meal time like?

Were there other family members nearby, such as grandparents (spell names if unusual). Did you see them often? Were you especially close to someone in your family? Describe where they lived. Please tell any anecdotes about family members.

What was religious life like? What denomination? Was there a nearby house of worship? If so, please describe it. Describe how you practiced your religion in the home. Did you experience any religious

persecution or prejudice of any sort? Describe holiday celebrations (food, music, special activities, gifts, religious observations, etc.)?

Describe school life. Did you go to school? Where was the school? Was it crowded? Do you remember specific teachers or playmates? What was your favorite subject? Did you learn English prior to coming to America?

COMING TO AMERICA

Who decided to come? Did you know someone who was in America already? Was a family member sending money from America? Describe getting ready to go and getting the proper papers. Did you want to come to America? What did you know about America? How did your mother feel? How did your father feel? Did anyone give you a “good-bye” party?

How much luggage did you pack? What did you take? What did you leave behind? What kind of luggage did you have? Did you take food? Did you take special belongings? If so, what?

Who came to America with you?

THE VOYAGE

What port did you leave from?

How did you get from your home to this port?

Describe the journey to the port. Tell any stories about this process.

What was the name of the ship? (spell if unusual)

Did you have to wait for the ship once you got to the port? If so, where did you stay? With whom? How long? Describe the experience.

Did any family members see you off?

When did the ship depart? (month and year?)

What were the accommodations like on the ship? What class did you travel? Describe your accommodations? Describe the dining room. What was the food like? Were you allowed on deck?

Describe what you saw, heard and smelled. Was it rough or smooth? Did you or your travelling companions become ill? Tell any anecdotes about the voyage.

STATUE OF LIBERTY

Describe seeing land for the first time?

Describe seeing the Statue of Liberty for the first time? Did you know what it was? Describe other people"s reaction to this experience.

What were your first impressions of seeing New York City from the boat?

ELLIS ISLAND

How did you get from the ship to Ellis Island?

Describe your impressions of seeing the Ellis Island building for the first time? Describe your impressions of the inside of the building.

Were you frightened? Were you excited?

Do you remember what you and your travelling companions were wearing when you arrived at Ellis Island?

How did your travelling companions feel about being at Ellis Island?

Was Ellis Island crowded? Was it clean? How were you treated by the staff? Describe the medical examinations? Where did they do it? How did they do it? Did everyone have the same examination?

Describe some of the people or things you saw at Ellis Island.

Were you detained at Ellis Island? If so, why? How long? Where did you eat? Describe the experience. Where did you sleep? With whom? Describe the accommodations. Any stories you associate with staying at Ellis Island?

How were you entertained while you stayed at Ellis Island?

Who came to meet you? When? How did you leave Ellis Island?

LIFE IN AMERICA

What were your expectations of America?

Where did you go after you left Ellis Island? What address? What city? How did you get there? Who met you once you got there? Describe the trip to your destination (i.e. train trip, subway ride, taxi, boat, etc.)? Did you see anything you had never seen before?

Describe the apartment or house? How many rooms? How many people lived there? How was it furnished? How was it lit? How was it heated? Was there indoor plumbing? Describe the neighborhood? Who lived there? Did other family members live near by? Did you get along well with your neighbors?

What jobs in America did family members get? Who supported the family? Did you work when you first got here? Did anyone not work? Describe the various jobs?

Did you go to school? Describe the building and class. How did you feel about going to school? Were you treated well by your fellow students? Do you remember any of the teachers? If so, why do you remember him/her? Any stories or anecdotes?

How did you learn English? Describe how you learned English. How difficult was the process? How did your family members learn English? Any stories associated with learning English?

Did you experience bigotry or persecution in America? Any stories or anecdotes. What was religious life like in America? Did you live near a house of worship? If so, name it and describe it. Who was more religious, your mother or your father? Why?

When did you move from this address? Where did you move to? For how long? Describe what you did for entertainment.

Describe how your family members (i.e. mother, father, grandparents, etc.) adjusted to life in America? Did anyone return to live in their country of origin? If so, why? Was your family satisfied or dissatisfied with life in America? Describe the individual adjustments of your father and your mother.

Did any family tragedy occur during the years following your coming to America? If so, what? Describe the experience.

Briefly describe the course of your life (i.e. marriage, children, occupation, anecdotes about meeting your spouse, etc.).

CONCLUSION

Are you happy you came to America? Were your parents (or other pertinent family members) happy they chose to come to America?

(graciously) Well, that"s a good place to end this interview. I want to thank you very much for taking the time for us to come out and speak with you about your immigration experience (or some such gracious wrap-up statement, allow them to respond if they choose to).

This is _____ signing off with _____ on _____ the _____, for the

Ellis Island Oral History Project.