Voices of the Manhattan Project

Exploring the Cultural Memory of AHF Oral Histories about the Manhattan Project.

06 June 2023

Davide Romano
EPFL

Cindy Tang
EPFL

Junzhe Tang
EPFL

***Abstract*—Our project is an attempt to use computational tools to explore the memories contained in the Voices of the Manhattan Project, an archive containing more than 600 oral histories of people living during the Manhattan Project. We applied Topic modelling and Name-Entity Extraction Recognition (NER) to extract the salient topics and people that are remembered in this corpus. Furthermore, we connect our results to previous works in the field of memory studies to put them in the context of possible psychological and social constraints that affect the act of remembering itself. Alongside our results, we also engage in a discussion regarding potential future investigations that could stem from this project as a foundation.**

I. Introduction

A. The Manhattan Project

"I have become death, destroyer of worlds"
- Robert Oppenheimer, 16th July 1945

On that day, the first atomic bomb in history was detonated successfully, scientists were clapping, cheering and shaking hands. However, they did not take long to realize and think through the deadly potential of what they had just created. The first atomic bomb was the result of a United States research and development effort known as the Manhattan Project (MP) during World War II. This crucial period with remarkable scientific breakthroughs led to the emergence of new ethical, societal, and political issues, thus capturing the attention of numerous researchers. Notably, the Atomic Heritage Foundation has collected more than 600 interviews of the MP workers. It is within this rich corpus that our project finds its focus.

B. Oral History

Interviews are oral histories. In order to understand and contextualize our corpus, we first need to define the terms "communicative memory" and "cultural memory" developed by Assmann in his work about oral memory and history [1].

Communicative or "everyday" memory is characterized by the stories, experiences and conversations of people about a specific event. It can be disorganized, fragmented, inaccurate, dynamic and subject to psychological and social factors [3]. Its temporal existence is defined by the lifespan of people who directly experienced these events, which means around 80-100 years, according to Assmann. As we find ourselves in the 2020s, we are approaching the temporal limit for the events of the MP, mainly around 1941-1945.

When the communicative memory of a period gently disappears, society attempt to preserve it by building a cultural memory around it. Cultural memory encompasses the tangible elements of a society's culture, including written texts, rituals, art and monuments. These elements are intentionally created to preserve and evoke significant historical events in the collective consciousness. Assmann also employs the term "figures of memory" to refer to these memories objectified through different fixed representations.
He argues that cultural memory exists in two forms. It takes the state of potentiality when the past is preserved in archives, libraries, and museums and it transforms into actuality when these preserved representations are assigned contemporary significance within different social and historical contexts. The dataset we work with is in the state of potentiality, and how it is interpreted in our today's social context is the actual form of it.

II. Research Questions

The AHF oral histories corpus is one considerable "figure of memory" forming part of the cultural memory about the MP. It is the largest archive in terms of the number of interviews it includes. As memory and not history, it can potentially describe the event from a new and worthwhile point of view. How can we take part in this cultural memory? How can we identify what will be actually saved by these communicative memories? Then, how can we give meaning to them? Based on Assmann theories and the preceding motivations, we defined our two main research questions as:

1) What topics and individuals are retained in the communicative memories of the AHF oral histories?
2) How are these topics and individuals related to each other within the context of the AHF oral histories?

III. Secondary Literature

We mainly explored literature that can be grouped into four categories: history and events of the MP, prosopographical research, memory studies and psychological factors affecting remembering.

A. History and events of the Manhattan Project

To get informed about the context and the timeline of the main events of the MP we investigated three main sources: two different summaries of the main events of the MP made respectively by Alex Wellerstein[14] and by the U.S. Department of Energy and the book ”The Manhattan Project: the birth of the atomic bomb in the words of its creators, eyewitnesses, and historians”[10]. This book includes extracts from the interviews that are part of our dataset. This literature helped us to learn about the different locations where the project was developed [12], important sociodemographic information like the presence of women and minorities[15] and interpret the results we got from our computational models.

B. Memory studies

Our projects fall into the category of research in the field of memory studies, a discipline dedicated to examining the role of memory as a fundamental tool for recollecting and understanding the past. To understand the state of the art of this field we based on publications [9] discussing and comparing the vast amount of definitions created in this field. Futhermore, we read about Assmann’s work on communicative/cultural memory [1], as well as Brown’s paper on collective identities [3], which gave us insights in the possible different narratives and stories that get form in different social groups. This inspired us to investigate more the possible differences in the results grouped by occupational category.

C. Prosopographical research

As our interviews are mainly containing biographical information, we read an important study in the field of prosopographical research [7] that involves using a machine learning framework including NER and LDA techniques to perform quantitative analysis of people mentioned in newspapers. It finalizes with the creation of a ”gazette” used to identify the influential individuals and their common characteristics.

D. Constraints of conversational remembering

In this project, we are dealing with ”conversational remembering”[8]: through conversations a speaker is invited to remember past experiences. Conversational remembering can be viewed as a social practice that promotes the formation of collective memory. However, what is remembered often depends on both the audience and conversational dynamics, and it can end up in the interviewee intentionally censoring himself/herself. There are various factors that affect this practice:

1) Retrieval/Reexposure Effects & Retrieval-induced forgetting: What is remembered in a conversation can affect the subsequent memories of conversational participants not only by implanting new and misleading memories but also by reinforcing some memory and inducing forgetting in others.
2) Social contagion (of memory): the spread of memory from one person to others by means of verbal interaction.
3) Retelling instead of recalling: far often people simply will retell a story about the past without trying to be historically accurate.

IV. Data

The corpus comes from the *Voices of the Manhattan Project* [6], an online collection of about 600 audio and visual interviews with MP workers and their families. This public archive is collected by the *Atomic Heritage Foundation (AHF)* and the *Los Alamos Historical Society*. For our project, we use the transcripts of these oral interviews accompanied by metadata such as the name and the role of the interviewee in the MP, the date and location of the interview.

For our analysis, we need to take into consideration certain limitations. The MP had at its peak 150’000 workers while we are dealing only with 600 interviews, a limited subset, that may not fully capture the diversity of experiences. We are subject to voluntary or involuntary non-participation of some members and the ultimate selection of the AHF. Additionally, most interviews were conducted after the 2000s (Fig. 1), when the AHF was created. The interviewees were asked to remember events that happened more than 60 years ago. Some memories or perceptions would have faded away or changed [3]. In general, the AHF’s objective is to preserve personal testimonies and experiences through time, therefore most interviews are in a biographical and ”intimate” form [5]. Some interviews are conducted to get information for a specific exhibition or book, so the questions asked as well as the answers may be intentionally directed towards specific topics.

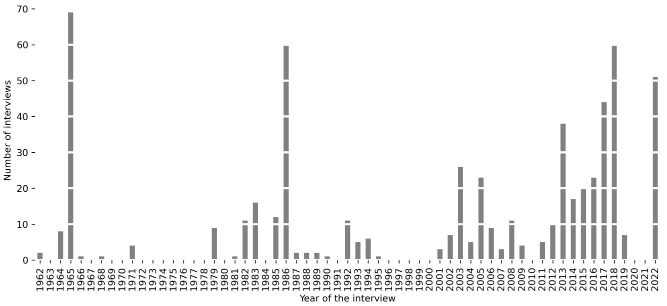

Fig. 1. Years in which the interviews were conducted.

Another important constraint is our cultural context. Starting from Assmann’s definitions of cultural memory, every communicative memory is then processed and edited based on the cultural values of the current time. This could cause biases in the questions of the interviews. When referring to ”bias,” we are specifically indicating the deliberate orientation of questions towards particular topics, events, or issues.

V. Methods

A. Data Scraping

We initiate by scraping interview transcripts from the AHF website[6], partitioning them into paragraphs, extracting additional biographical data from AHF’s profile page[4], and gathering metadata about each interview.

B. Pre-processing

We utilize a basic NLP pipeline from NLTK library for tokenization, stopword removal, casefolding, and lemmatization of the transcripts.

C. Topic Modelling

Applying a Latent Dirichlet Allocation (LDA) model from the Gensim library to our preprocessed transcript corpus is our initial step to get its topical content. We experiment with multiple hyper-parameters, assessing the topics extracted through qualitative and quantitative measures, specifically utilizing the coherence score [13] provided by Gensim to gauge semantic similarity among high-scoring words within a topic. After several iterations, the following hyper-parameters emerged as the optimal framework for our analysis.

```
LdaModel(num_topics=7,
         alpha=’auto’,
         eta=’auto’,
         passes=10,
         iterations=500,
         random_state=42)
```

The trained LDA model is then applied on each transcript paragraph to allocate the most likely topic.

D. Person Entity Extraction

We utilize *SpaCy* for Named Entity Recognition (NER) to detect individuals mentioned in the interviews.

1) Entity Disambiguation:
Post-NER, we identified various names, their frequency, and forms. However, the presence of partial names, like ”Robert,” ”Marshall,” ”Nichols,” ”Bob,” etc., creates ambiguities due to the overlap in the interviewees’ names. Hence, entity disambiguation is critical to select appropriate full names to replace these partial ones.

To this end, we use profile names as our knowledge base, each corresponding to a specific person’s full name. We then compute the similarity between extracted names and the knowledge base, resulting in one or more similar names for each extracted name. We apply methods such as Soundex[11] and string matching, with Soundex serving to index names based on their pronunciation.

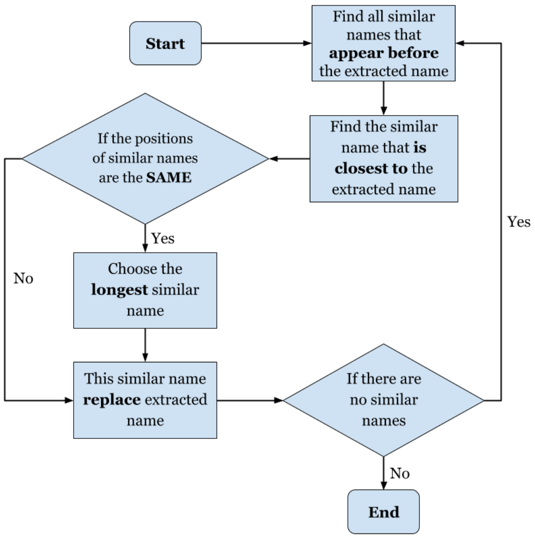
Start
Find all similar names that **appear before** the extracted name
Find the similar name that **is closest to** the extracted name
If the positions of similar names are the **SAME**
Yes
No
Choose the **longest** similar name
This similar name **replace** extracted name
If there are no similar names
Yes
No
End

Fig. 2. Flow chart of name linking

We then employ entity linking[2] to choose the most suitable similar name, using a method based on distance weights due to the nature of the interviews. Typically, when someone is initially mentioned in a conversation, their full name is used to clearly identify the person and provide sufficient context to the listener. Subsequently, they are usually referred to by their surname, first name, or nickname.

Therefore, for each name, we iterate as per Fig. 2: locate similar names that appear before the extracted name, find the closest one based on paragraph and sentence indices, and select the longest name in case of position ties. If no similar names are located, the loop ends. Finally, all partial names are replaced with the full names of individuals recently mentioned.

2) Manual Processing:
Despite our attempts, some data still requires manual processing to address issues like non-human personified names, company names mistaken as people, and unconventional nicknames not identifiable by Soundex. Effective handling of such cases may necessitate semantic analysis.

VI. Results & Interpretations

A. What are the AHF oral histories talking about?

The results of our LDA model, displayed in Table I, provide insights into the topics discussed in the AHF oral histories. We gave semantic names for each topic and identify the personalities present in the extracted words. To show in Fig. 3 the relative prominence of these topics within the interviews, we count the number of paragraphs in the transcripts that are most likely to belong to each topic. The two main topics revolve around Cold War & Politics and Family & Daily Life. The remaining topics focus on specific events or programs during the MP. It indicates that these memories enclose not only scientific and technical aspects, but also valuable insights into the interviewees’ daily lives, familial experiences, and the political context of this event.

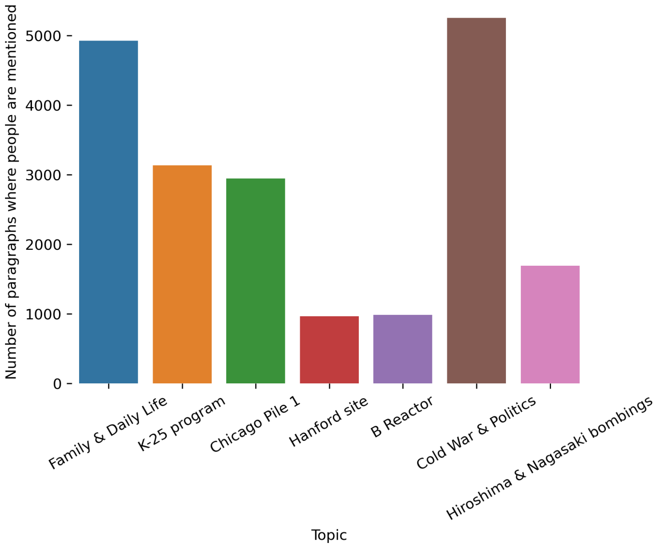

Fig. 3. Distribution of topics.

B. Who is remembered in AHF oral histories?

The analysis of the top 50 names (Fig. 5) confirmed the presence of well-known individuals associated with the Manhattan Project, as mentioned in the secondary literature. Notable figures include Robert Oppenheimer, the director of Los Alamos Laboratory, General Leslie Groves, the US Army Corps of Engineers officer overseeing the project, Edward Teller, the "father of the hydrogen bomb," and Enrico Fermi, creator of the Chicago Pile 1.

We observed a skewed distribution of mentioned individuals in the interviews, with 32% of mentions belonging to only 1% (50 names) of the recognized total. One possible hypothesis is related to the constraints of conversational remembering[8]. Over the course of time between the events and the interviews, individuals were likely susceptible to the phenomenon of retrieval-induced forgetting, wherein certain memories become reinforced through repeated discussions, and other memories are forgotten. As conversations and questions increasingly focused on the most remarkable events and individuals, it is possible that memories about lesser-known individuals and stories gradually diminished. This gradual fading may be attributed to the narrowing of their recalling as a result of the repeated emphasis on certain aspects of the past.

Furthermore, there is a discrepancy between the number of interviews and the paragraph count shown in Fig. 5. Some individuals have a strong presence within a few interviews, while others like E. Fermi or L. Groves may be mentioned in more interviews but rank lower in paragraph count. This is likely due to their status as prominent figures, leading to their mention in discussions about the project, even when there is no personal relationship with the interviewee.

Contrary to our initial hypothesis, R. Oppenheimer and L. Groves did not rank at the top. Instead, E. Teller had the highest paragraph count whose significant contributions came after World War II. This may be influenced by the recency of the interviews conducted and the attention given to the more powerful and contemporary hydrogen bomb. However, in terms of human death count, the fission bombs dropped on Hiroshima and Nagasaki had a larger impact on humanity. Despite this, the fusion bomb still attracts more attention. This observation raises questions about the relationship between scientific advancements and their impact on humanity. It suggests that scientific developments and current news tend to capture more attention than past events even if having a greater impact on human lives.

C. What role do they play in the Manhattan Project?

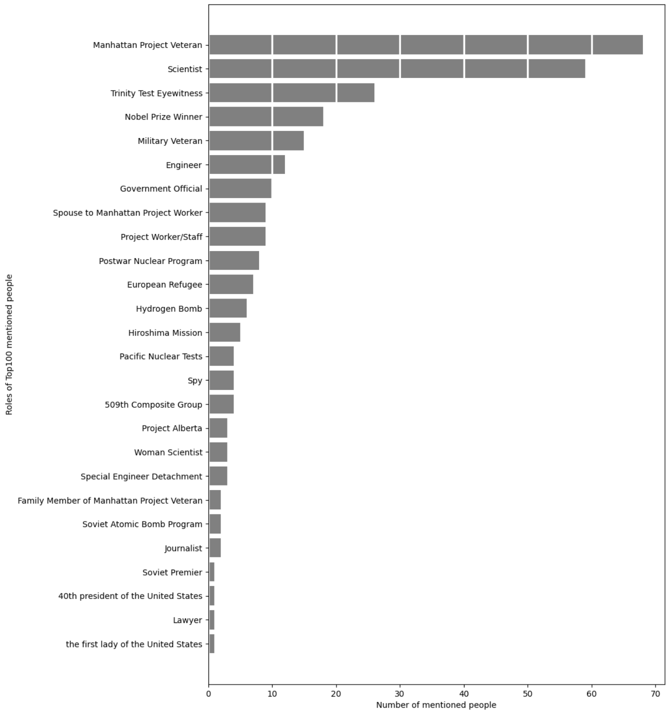

Fig. 4. Distribution of roles of top 100 mentioned people.

Fig. 4 illustrates the various roles played by the top 100 individuals mentioned in the interviews. The labels, created by the AHF, allow for multiple roles assigned to each person. The most prominent groups include scientists, Nobel Prize Winners, and Trinity Test Eyewitnesses, which aligns with expectations. However, there are interesting exceptions, such as military veterans, a dozen Project workers/staff, and a few women scientists, highlighting the diversity of roles represented in the interviews.

D. What is associated to these mentioned people?

The individuals most frequently mentioned in each of the seven topics extracted from the LDA analysis in Fig. 6 align

with the locations and events of the MP.
An intriguing point is the absence of R. Oppenheimer in the Hiroshima and Nagasaki bombings topic, which is rather linked with Harry Truman, George Marshall, and Franklin Roosevelt—namely, the US Presidents and the Chief of the US Army. This observation suggests that the bombings are more commonly associated with the decision-makers rather than the scientists and engineers.

Notably, Dorothy McKibbin stands out among the top 50 mentioned names (Fig. 5) as the sole woman. In the topic of family and daily life, she holds the highest position. McKibbin was considered as the ”Gatekeeper to Los Alamos”, responsible for welcoming new recruits to the MP. While she may not have made significant scientific contributions and is absent from project literature, she is frequently referenced in the interviews. This can be attributed to the workers' daily interactions with her at Los Alamos due to her role. It highlights the unique nature of our interview corpus, which captures memories rather than official historical records. McKibbin represents one of the ”hidden figures” who left a strong impression on MP workers, despite their limited recognition in historical archives.

To explore more in-depth these ”hidden figures” we created a ”gazette”[7], in Table II that associates for each entity we extracted from the interviews a topic, and a list of words.

E. Women talks about women?

When focusing on the subgroup of interviews with women scientists (Fig. 7), other women appeared among the top 20 mentioned names. This may be due to a potential bias in the interview questions. Given that women were a minority in the MP, there could be a higher probability of being asked about other women.

VII. Future improvements

Our work gave us inspiration of possible future directions and research investigations that would build upon our results.

A. Studying the dynamics of memories over time.

Comparing interviews conducted during different periods can provide valuable insights into the evolution of memories over time. As depicted in Fig. 1, our interviews are spread out over a significant interval, which adds an interesting dimension to our analysis. By considering the interview dates, we can investigate how the passage of time influences the recollection and portrayal of memories. It raises questions about the potential influence of previous interviews on those conducted later in time.

B. Exploring the ”hidden figures”

Dorothy McKibbin, as revealed in the interviews, represents a notable example of a ”hidden figure” within the MP. Conducting a more in-depth investigation into the experiences and contributions of individuals like her, who are less recognized in historical archives, can provide valuable insights into the hidden aspects of history through the lens of memory. It can shed light on the roles of individuals from diverse backgrounds, the dynamics of power and influence within the project, and the potential impact of these ”hidden figures” on the overall history.

C. Tool for further research

Our results, including topic modeling and gazette, can be utilized for historical and prosopographical research. They enable indexing of transcripts based on topics and individuals. One example of further research could be to explore the interplay between science and humanities in this context. Through our observations, the Hiroshima and Nagasaki bombings are more commonly associated with the decision-makers rather than the scientists and engineers. This can lead to a discussion about the social responsibilities of the huge consequence of these bombings. Such a study could provide valuable insights into how society perceives and prioritizes scientific progress in relation to its implications for humanity.

D. Exploring the possible bias in the question.

In order to gain a more comprehensive understanding of our results, it would be beneficial to examine the formulation and content of the questions. We could potentially gain insights into possible biases that may influence the responses obtained.

E. Expanded LDA Analysis.

To enhance the depth of the analysis of our results we could explore qualitatively the results of our LDA with an increased number of topics, such as 30 or even 90 topics.

VIII. Conclusion

In this study, we apply computational methods such as Latent Dirichlet Allocation (LDA) and Name-Entity Recognition (NER) to delve into the manifold aspects of the MP and the individuals involved. Our analysis reveals a diverse range of topics discussed in the interviews, encompassing not only scientific advancements but also themes related to family, daily life, and politics. Furthermore, we observe among the vast workforce of about 150,000 individuals, only around fifty of them are frequently mentioned, highlighting the prominence of certain personalities within the cultural memory of the MP.

By examining the association between individuals and topics, we gain insights into how people connect specific personalities with particular aspects of the MP. Notably, our research brings to light the presence of ”hidden figures” like Dorothy McKibbin, whose contributions may have been overlooked in history but are strongly present in the collective memory of those who interacted with her.

When interpreting our findings, we took into account the constraints arising from our sample size, the impact of contemporary society’s view on these historical events, and the implications of ”conversational remembering” on the memories of the interviewees. These factors play a significant role in shaping the narratives shared during the interviews.

Through our exploration, we only treat the surface of the immense potential this corpus holds. While we start to extract insights and provide directions for further research, there is still much more to be uncovered. Further and deeper investigations should be undertaken to fully extract significant meaning from this archive.

Digital humanities not only provides answers but, more importantly, brings new questions and opens up new avenues of exploration. We aspire to contribute a small part to the cultural memory surrounding this pivotal moment in history, leading to new perspectives of exploratory or probatory researches, such as studying the dynamics of memories over time, investigating the experiences of minority communities that may be underrepresented in historical accounts, and exploring the interplay between science and humanities within the context of the MP.

References

[1] Jan Assmann and John Czaplicka. “Collective memory and cultural identity”. In: *New german critique* 65 (1995), pp. 125–133.

[2] Indrajit Bhattacharya and Lise Getoor. “Collective Entity Resolution in Relational Data”. In: *ACM Transactions on Knowledge Discovery from Data* (2007).

[3] Andrew D Brown. “A narrative approach to collective identities”. In: *Journal of management Studies* 43.4 (2006), pp. 731–753.

[4] Atomic Heritage Foundation. *Profiles*. URL: https://ahf.nuclearmuseum.org/ahf/bios/.

[5] Atomic Heritage Foundation. *Transforming the Relationship between Science and Society: The Manhattan Project and Its Legacy*. 2013. URL: https://ahf.nuclearmuseum.org/wp-content/uploads/2014/06/FINAL%5C%20Atomic%5C%20Heritage%5C%20Foundation%5C%20Workshop%5C%20Report.pdf.

[6] Atomic Heritage Foundation. *Voices of the Manhattan Project*. URL: https://ahf.nuclearmuseum.org/voices/.

[7] Aayushee Gupta et al. “A machine learning approach to quantitative prosopography”. In: *arXiv preprint arXiv:1801.10080* (2018).

[8] William Hirst and Gerald Echterhoff. “Remembering in conversations: The social sharing and reshaping of memories”. In: *Annual review of psychology* 63 (2012), pp. 55–79.

[9] Wulf Kansteiner. “Finding meaning in memory: A methodological critique of collective memory studies”. In: *History and theory* 41.2 (2002), pp. 179–197.

[10] Cynthia C Kelly. *Manhattan Project: The birth of the atomic bomb in the words of its creators, eyewitnesses, and historians*. Black Dog & Leventhal, 2009.

[11] Donald E. Knuth. *The Art of Computer Programming, Vol. 3: Sorting and Searching*. 2nd. Addison Wesley Longman Publishing Co., Inc., 1998.

[12] Bruce Cameron Reed. *The history and science of the Manhattan Project*. Springer, 2014.

[13] Shaheen Syed and Marco Spruit. “Full-text or abstract? examining topic coherence scores using latent dirichlet allocation”. In: *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE. 2017, pp. 165–174.

[14] Alex Wellerstein. “Manhattan Project”. In: *Encyclopedia of the History of Science* (2019). URL: https://ethos.lps.library.cmu.edu/article/35/galley/48/view/.

[15] Alex Wellerstein. “Women, minorities, and the Manhattan Project”. In: *Restricted Data: The Nuclear Secrecy Blog* (2015). URL: https://blog.nuclearsecrecy.com/2015/11/27/women-minorities-and-the-manhattan-project/.

Appendix

Topics	**Personalities**	**Words**	**Paragraph counts**
Cold War & Politics	Robert Oppenheimer, Edward Teller, Klaus Fuchs	oppenheimer, weapon, teller, robert, science, soviet, russian, student, hydrogen, term, edward, berkeley, committee, president, inaudible, communist, example, french, party, fuchs, ...	5’248
Family & Daily Life	-	father, mother, child, santa, married, fe, town, dad, girl, brother, kid, husband, bus, road, barrack, wonderful, camp, hall, sister, hill, …	4’853
Engineering of the K-25 program (Oak Ridge)	Leslie Groves, Kenneth Nichols	grove, barrier, engineering, equipment, diffusion, metal, operation, president, nichols, design, committee, decision, construction, york, dupont, keith, lawrence, chemical, columbia, colonel, …	3’138
Chicago Pile 1	Enrico Fermi, Leo Szilard, Eugene Wigner	fermi, szilard, reactor, chemistry, plutonium, student, wigner, experiment, pile, reaction, science, professor, neutron, chain, enrico, graduate, hanford, dupont, dr, chemist, …	2’933
Hiroshima & Nagasaki bombings	Leslie Groves, Harry Truman	japanese, japan, mission, plane, weapon, hiroshima, force, german, nagasaki, grove, dropped, crew, 000, b, airplane, ship, bombing, truman, island, germany,…	1’690
Hanford site	-	hanford, richland, river, dupont, waste, construction, 000, reactor, land, pasco, town, tank, columbia, 00, study, worker, report, cleanup, camp, law, …	964
Engineering of the B Reactor	-	reactor, plutonium, radiation, fuel, power, hanford, design, neutron, element, chemical, facility, dupont, system, tube, radioactive, operation, separation, level, engineering, rod, …	963

TABLE I
LDA results with 7 topics.

Personalities	**General Topic**	**Words from LDA with 7 topics**	**Words from LDA with 90 topics**
Robert Oppenheimer	Cold War & Politics	oppenheimer, weapon, teller, robert, science, soviet, russian, student, hydrogen, term, edward, berkeley, committee, president, inaudible, communist, example, french, party, fuchs, ...	oppenheimer, robert, kitty, party, frank, berkeley, communist, princeton, member, letter, activity, committee, died, peter, relationship, union, laughter, student, trouble, political,...
Edward Teller	Cold War & Politics	oppenheimer, weapon, teller, robert, science, soviet, russian, student, hydrogen, term, edward, berkeley, committee, president, inaudible, communist, example, french, party, fuchs, ...	oppenheimer, teller, weapon, hydrogen, russian, edward, committee, bethe, fission, implosion, thermonuclear, force, shot, ulam, radiation, calculation, design, clear, soviet, oppie,...
Dorothy McKibbin	Family & Daily Life	father, mother, child, santa, married, fe, town, dad, girl, brother, kid, husband, bus, road, barrack, wonderful, camp, hall, sister, hill, …	oppenheimer, dorothy, black, student, fe, santa, j, science, famous, east, graduate, town, someone, hill, mckibbin, palace, finished, white, class, eventually,...
Harry Truman	Hiroshima & Nagasaki bombings	japanese, japan, mission, plane, weapon, hiroshima, force, german, nagasaki, grove, dropped, crew, 000, b, airplane, ship, bombing, truman, island, germany,…	japanese, japan, hiroshima, surrender, truman, president, decision, emperor, weapon, camp, survivor, peace, nagasaki, bombing, 1945, grandfather, 000, ii, prisoner, thousand,...

TABLE II
”Gazette”. [7]

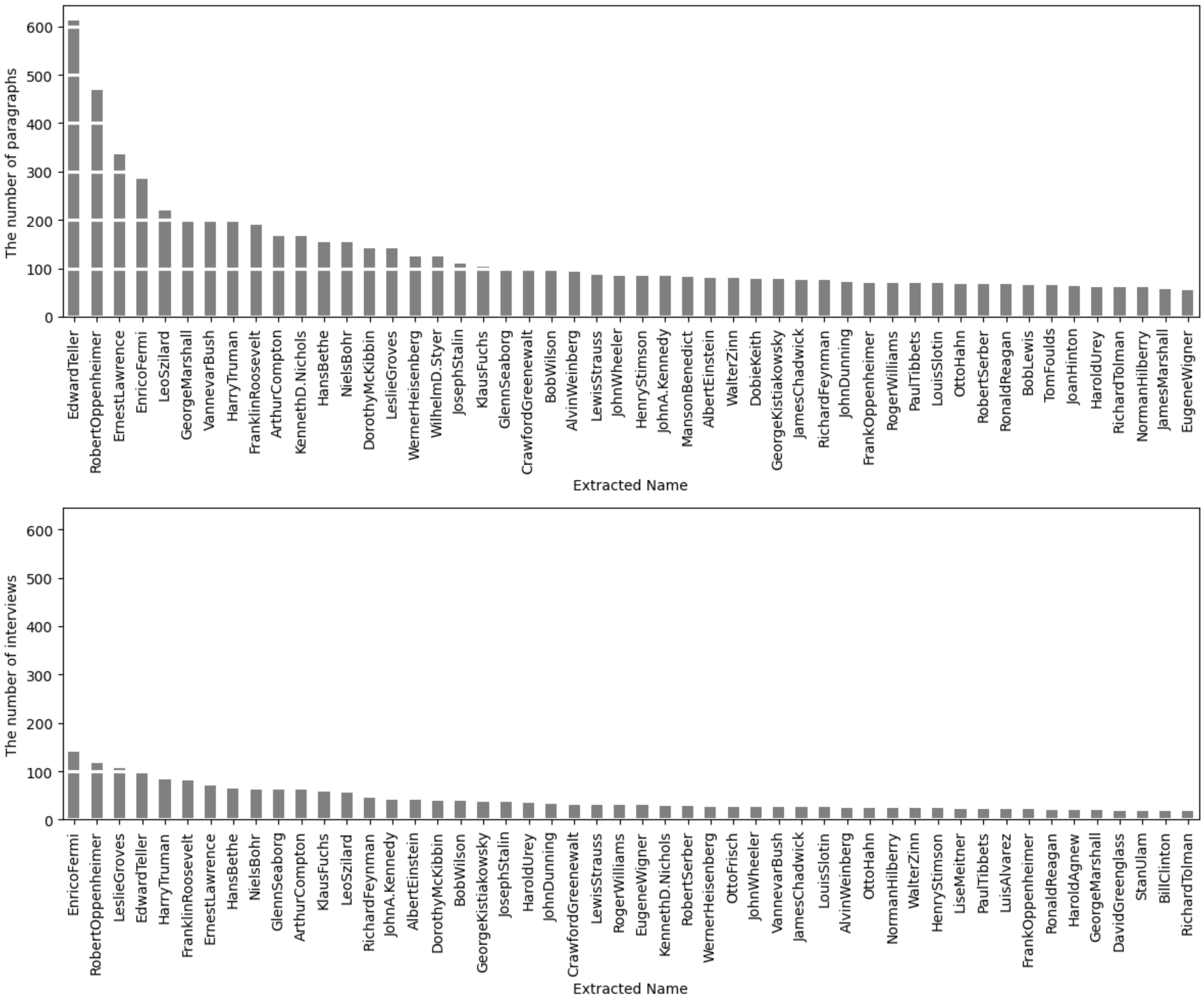

Fig. 5. Top 50 extracted names by the number of paragraphs vs by number of interviews

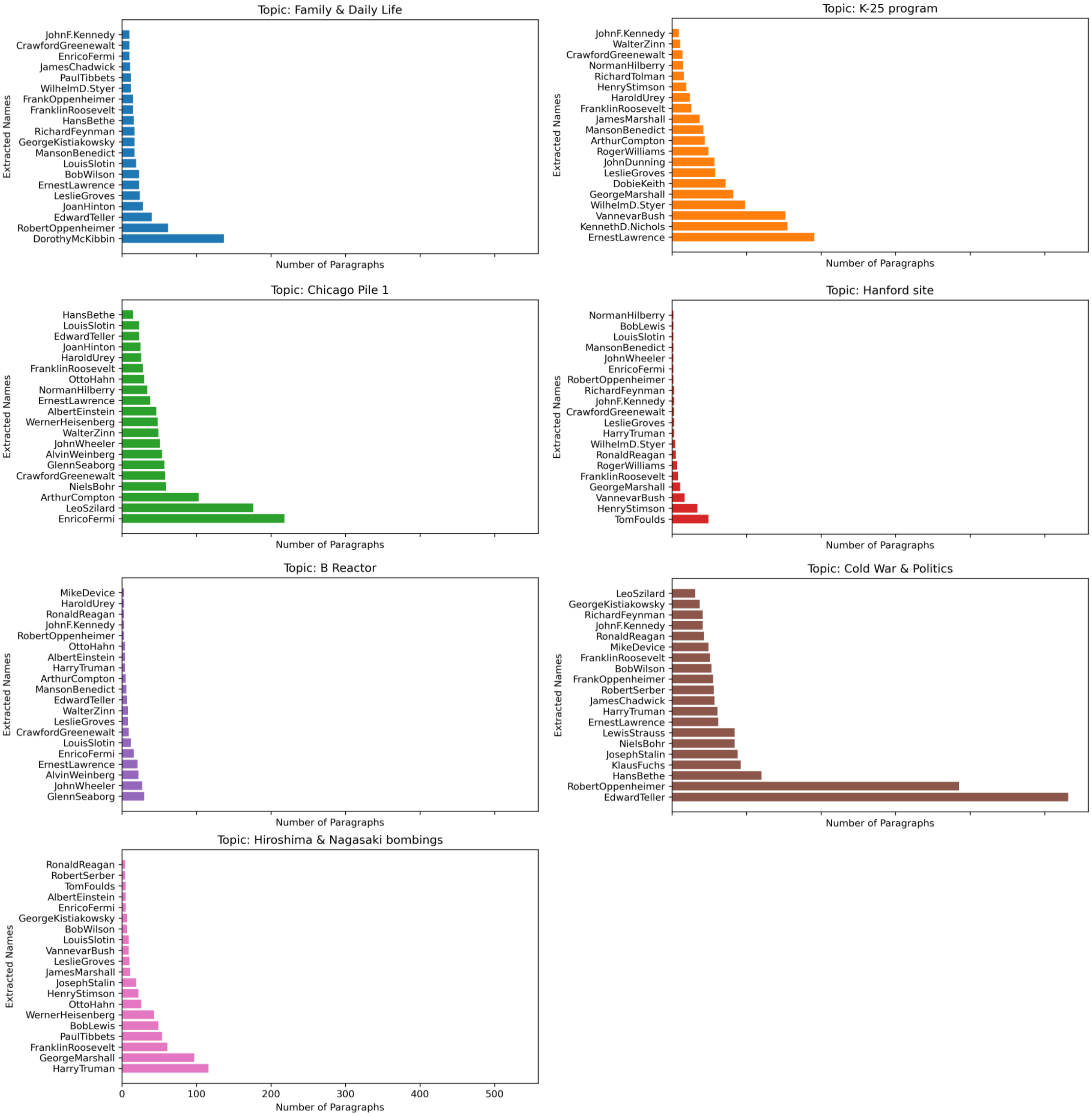

Fig. 6. Top 20 extracted names for each topic.

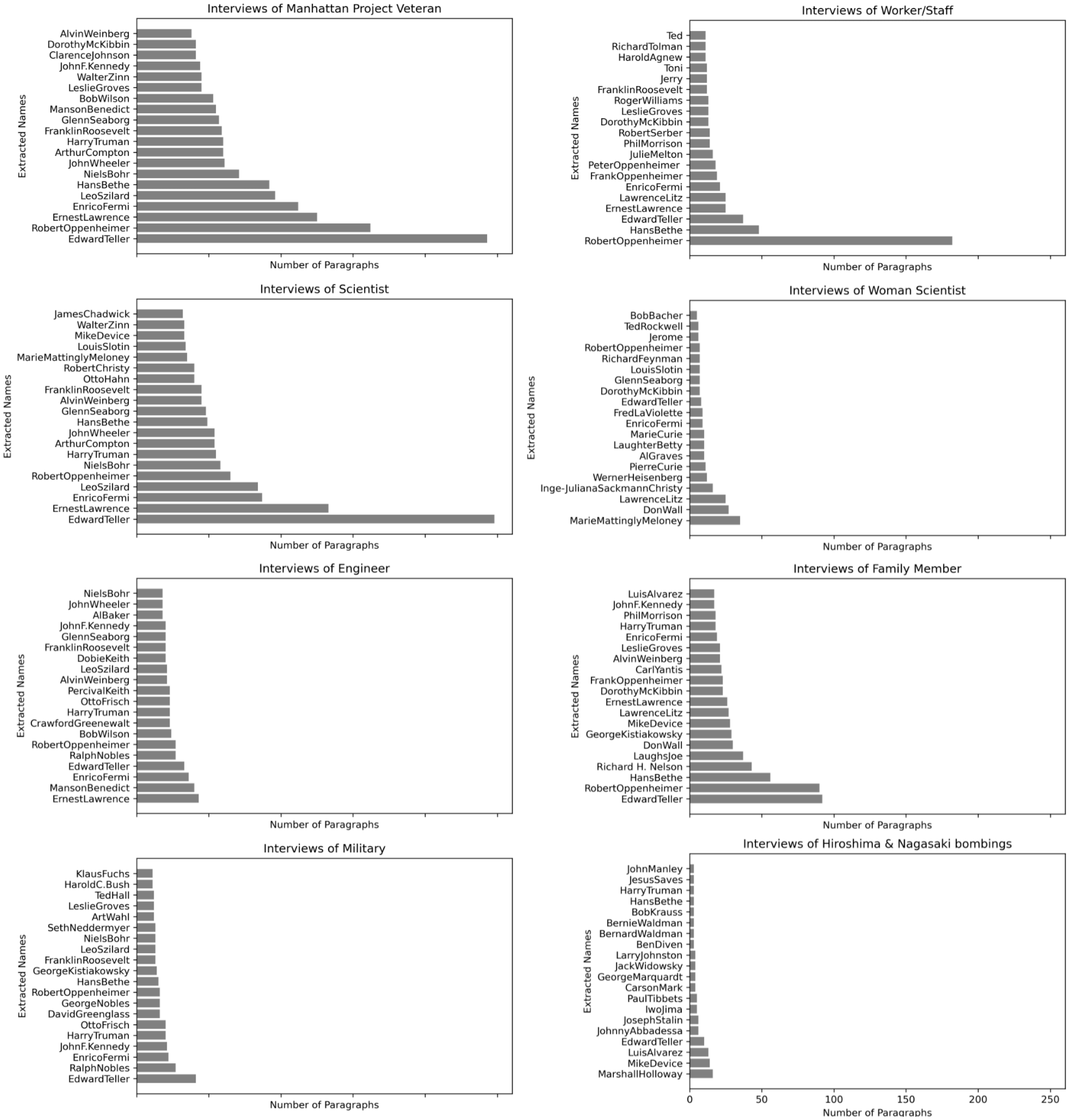

Fig. 7. Top 20 extracted names for each subgroups of interviewee's roles.