Building a Corpus of Oral Testimonies of the Holocaust: a case study in corpus linguistics and oral history



Caitlin Wilson
Wadham College
University of Oxford

Dissertation submitted for the degree of
Master of Science in Digital Scholarship

Summer 2023

Abstract

This dissertation will give an overview of what goes into creating a digital cross-linguistic corpus. Using a case study, I will describe the elements, methodology, and possible research outcomes of building a digital corpus. The case study consists of a collection of oral testimonies of Holocaust survivors that have been transcribed into a collection of English, Polish, Russian, Czech, and Hungarian texts. Building on existing literature, this dissertation will describe how corpus linguistics can make research data more accessible to researchers of many different areas of scholarship. Having given an overview of how a corpus is created and digitally distributed, including the range of tools and platforms available for analysing corpora, I will explore two areas of possible research within corpus linguistics: geo-location using named entity recognition programmes, and syntactic analysis of hedging used in oral histories thanks to part of speech tagging. Through building a corpus of Holocaust Oral Testimonies, I hope to demonstrate how digital methods of corpus linguistics can help further interdisciplinary research within the digital humanities.

Notes to the Reader

The interviews used to create this corpus have no restrictions on access or use. All transcripts and original recordings are available to the public on the USHMM website (`https://www.ushmm.org`). The corpus created for this dissertation has been distributed via the Oxford CQPWeb server (`http://cqpweb.ling-phil.ox.ac.uk/CQPweb/`). Access can be granted via request to the system administrator, Martin Wynne. For examiners of this dissertation, a CQPWeb account has been created. The login credentials are:

- Username: examiner
- Password: ushmm_ht02

The corpus created for this dissertation project can be found in the category 'Holocaust Testimonies (experiment 2)' the category named 'Holocaust testimonies' is the first version of this project which was presented at the EHRI-CLARIN workshop in May. If you have any issues accessing the corpus or logging in to CQPWeb please contact Martin Wynne (martin.wynne@ling-phil.ox.ac.uk). All appendices are available in the adjoined zip file. These are:

Appendix A: Raw files for each subcorpus (tagged form) in files named 'HT02_language.vrt';

Appendix B: Metadata files for each subcorpus in files named 'HT02_metadata_language.tsv';

Appendix C: Database of GPE tagged items in file named 'HT02_migration.txt';

Appendix D: Database of first person singular present tense verbs in file named 'HT02_present_tense.txt';

Appendix E: Database of hedging techniques in file named 'HT02_hedging.csv.'

Contents

1 Introduction **1**

2 Background **2**
2.1 Oral History . . . 2
2.2 Corpus Linguistics . . . 3
2.3 Combining Oral History and Corpus Linguistics . . . 5
2.4 This project: an Annotated Multilingual Corpus of Oral Testimonies of the Holocaust . . . 7

3 Methodology **8**
3.1 Building the Corpus . . . 8
3.2 Cleaning the Files . . . 8
3.3 Translation . . . 10
3.4 Tagging . . . 10
3.4.1 Universal Dependency Tagging . . . 11
3.4.2 Named Entity Recognition . . . 12
3.4.3 Stanza . . . 13
3.5 Metadata . . . 14
3.6 Uploading and Aligning the Corpora . . . 15

4 Pilot Studies **16**
4.1 Mapping Locations and Movement in Oral Testimonies . . . 16
4.1.1 Description . . . 16
4.1.2 Methodology . . . 17
4.1.3 Results . . . 20
4.2 Hedging Techniques in Oral Testimonies . . . 21
4.2.1 Description . . . 21
4.2.2 Methodology . . . 22
4.2.3 Results . . . 24

5 Discussion **25**
5.1 Distant Reading of Oral History . . . 25
5.2 Ethical Considerations . . . 26
5.3 Areas of Improvement . . . 27

6 Conclusion **28**

References **30**

List of Tables

1 *Aggregated count of locations named in Holocaust Oral Testimonies by country.* 18
2 *Frequency breakdown table of 1.sg.pres verbs* 23

List of Figures

1 *Cleaned transcript from Tibor Baranski's oral testimony, in XML format.* 10
2 *Transcript of Margaret Beer's oral testimony parsed using NLTK, in VRT format.* 11
3 *Transcript of Margaret Beer's oral testimony tagged using Stanza, in VRT format.* 14
4 *Czech corpus metadata, in TSV format.* 15
5 *Choropleth map of countries named in the complete Holocaust Testimonies, created using Tableau.* 19
6 *Migration routes named in the complete Holocaust Testimonies, created using Tableau.* 20

1 Introduction

The preservation and analysis of historical narratives are crucial for understanding human experiences, particularly in the context of significant events like the Holocaust. Oral testimonies offer invaluable firsthand accounts, shedding light on diverse perspectives of those who lived through those harrowing times. To fully harness the potential of such testimonies, digital methods from corpus linguistics can help broaden our understanding of the lived experiences of of those who witnessed and survived the Holocaust.

Corpus linguistics tools and methods enable researchers to engage in 'distant reading' of textual sources, as proposed by Moretti (2000). This computational and quantitative approach is utilised in various disciplines, including sociology, linguistics, history, and cultural studies. While oral history studies are increasingly employing distant reading or corpus linguistics methods, dedicated literature on the topic is still limited. Ethical considerations, such as the fear of losing individual voices, are important when using a quantitative approach to oral history. However, distant reading should be seen as a *complementary* approach to close reading, as it allows a broader understanding of overarching themes, supporting informed close analysis of individual accounts.

This dissertation focuses on the creation of a digital cross-linguistic corpus of oral histories, using a case study of Holocaust survivor testimonies in English, Czech, Hungarian, Polish, and Russian. By building and studying this corpus, I explore the process, methodology, and potential research outcomes of digital corpus creation. The primary goal is to showcase how corpus linguistics enhances accessibility and enables interdisciplinary research by making data readily available to scholars from various fields. I first illustrate how this interdisciplinary approach fosters a deeper understanding of historical events, linguistic patterns, and the cultural significance embedded within the testimonies by exploring existing literature on oral history and corpus linguistics. I provide an overview of the process of creating and distributing the corpus using CQPWeb, a corpus linguistics tool. I then explore two areas of potential research: geo-location using named entity recognition and linguistic analysis of hedging exploiting part-of-speech tagging. These areas demonstrate the practical application of digital methods to the Holocaust Oral Testimonies corpus, facilitating interdisciplinary research within the digital humanities. Lastly, I provide a discussion of this project which emphasises how digital and interdisciplinary approaches to oral history can help our understanding of the Holocaust's impact on individual lives.

2 Background

2.1 Oral History

Oral history, or "the interviewing of eye-witness participants in the events of the past for the purposes of historical reconstruction" (Grele, 1996, 64), has drastically changed the way in which contemporary historical research has been practiced in the last century. This definition highlights the difference between the first-person eye-witness testimonies in oral history and the stories of oral traditions, which are transmitted across generations (Thomson, 2020). As Perks and Thomson (2015, 6) state, while interviews with members of social and political elites have long provided valuable insights to existing documentary sources, oral history's most significant impact lies in its ability to capture the experiences and perspectives of marginalised groups and ordinary people who might have remained "hidden from history." These narratives, which would otherwise have been confined to personal papers or fragments of autobiographical writing, are now being woven into the fabric of the historical record.

Historians have long been recording the evidence of historical witnesses in hand- or typewritten notes but these are often incomplete, not containing every detail, and can be shaped by the scribe's perspective. The creation of audio recording devices in the twentieth century allowed researchers to record interviews thereby creating primary sources that could be archived for future historical research. In everyday language, the term 'oral history' has expanded to encompass both audio and video recordings which can later be transcribed so as to be studied as textual material. While self-taped oral testimonies fall under the umbrella of oral history, oral histories are typically recorded in an interview. The interviewer therefore plays an important role in oral history, "guiding, triggering, probing, and recording" the interviewee's story (Thomson, 2020, 1).

As mentioned, oral history often captures the memories and perspectives of people and historical events that would otherwise be unlikely to be recorded in written sources. Indeed, as audio recording devices became more and more affordable, oral history emerged as a valuable resource of public history. Through its capacity to convey personal narratives, oral history contributes to engaging and enriching our understanding of the past by providing immersive experiences with those present at the time (Hamilton & Shopes, 2008). Notable oral history projects and collections include the Irish Life and Lore Project which preserves oral histories and folklore in Ireland (`https://www.irishlifeandlore.com`), the Ellis Island Oral History Project which captures the experiences of immigrants passing through Ellis Island in the early twentieth century (`https://www.statueofliberty.org/discover/stories-and-oral-histories/`), and the British Library Na-

tional Life Stories which is an ongoing project to preserve oral histories from people across the United Kingdom on subjects like community, work, and culture (https://www.bl.uk/national-life-stories).

Oral history projects can also include the gathering of testimonies from witnesses in (quasi-)judicial frameworks, such as truth and reconciliation commissions or government inquiries. Such inquiries typically serve the purpose of comprehending historical injustices, acknowledging the experiences of victims and survivors, and potentially providing restitution or compensation (Thomson, 2020). An example of this type of oral history is the South African Truth and Reconciliation Commission which conducted public hearings in post-apartheid South Africa where both victims and perpetrators could give testimony in the hope of fostering an opportunity for healing, understanding, and reconciliation (https://www.justice.gov.za/trc/). Following the demise of Nazi Germany and the emancipation of captured Jewish and Slavic people, Roma communities, Soviet prisoners of war, and other persecuted populations, significant endeavours were undertaken to document the testimonies of both the victims and eyewitnesses of the atrocities perpetrated during the Third Reich. These oral testimonies exist among a vast collection of repositories in museums, archives, and universities across the world. One of the largest and most famous of these is the Fortunoff Video collection of Yale University (https://fortunoff.library.yale.edu). The archive comprises a vast collection of video recordings that aim to deepen visitors' understanding of the Holocaust through oral histories and personal narratives. Other notable collections of Holocaust testimonies include the USC Shoah Foundation Visual History Archive (https://vha.usc.edu), the Imperial War Museum (https://www.iwm.org.uk), the United States Holocaust Memorial Museum (https://www.ushmm.org), and the Bundesarchiv in Germany (https://www.bundesarchiv.de).

2.2 Corpus Linguistics

Modern corpus linguistics arguably began with Kučera and Francis's study of the Brown corpus in 1967. They describe their work as "a collection of lexical and statistical information obtained from analysis of [...] a computer-processible corpus of language texts" (Kučera & Francis, 1967, xvii). The Brown corpus consists of one million words of American English written text printed in 1961. It was the first of its kind, being the first 'large' computer-readable corpus. The collection of texts created to study language, however, far predates the twentieth century. We have evidence of work dating from the Middle Ages that saw the collection of words used in a given text and the contexts in which they appear (which we today call concordances) (Sinclair, 1991). Since the 1950s, the creation of corpora

has vastly expanded thanks to computerised, machine-readable text (McEnery & Hardie, 2013). Many of the first early modern corpora followed the same structure as the Brown Corpus, such as that Lancaster-Oslo-Bergen corpus of British English and the Kolhapur corpus of Indian English. As the turn of the century approached, language corpora got larger thanks to expansive collaboration projects such as the British National Corpus (BNC), published in 1995. This corpus, of which there have since been multiple new editions, consists of 100 million words from written and spoken samples. Alongside the BNC, many other national corpora have been created, providing insight into and allowing research of contemporary languages across the world. These include the Russian National Corpus (over 1 billion words), the KOTONOHA Japanese language collection of corpora (over 2 billion words), and the German Reference Corpus (4 billion words).

Nowadays, and largely thanks to improving computational methods, we have access to a large variety of corpora organised not only by language but also by text type, topic, and author. The Google Books Ngram corpus, for example, is a multilingual corpus charting the frequency of n-grams found in the printed sources collected and digitised as part of the Google Books project. Text corpora can span multiple languages, either in form of parallel translation corpora such as the ACTRES Parallel Corpus of Spanish and English, or as themed corpora that happen to contain different languages such as the aforementioned Google Books corpus or the ParlaMint corpus of European Parliamentary sessions. Nowadays, as computational methods and access to vast amounts of linguistic data become more widespread, most corpora are fully accessible online and indeed rely on data collected (or *scraped*) from the internet to build their collections of words and phrases e.g., WebCorp Live or the News on the Web corpus. As Joseph (2004, 382) wrote in *Language*: "we seem to be witnessing as well a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data".

As for the utilisation of linguistic data, corpus linguists make use of a variety of tools to analyse corpora. These tools are many and varied, ranging from those are have been purpose built for specific projects and corpora to those that support the study of any corpus. For example, the Hansard at Huddersfield tool (Hansard at Huddersfield Project, 2018) performs searches on the Hansard corpus of British Parliamentary debates. CLiC (Mahlberg, Stockwell, Wiegand, & Lentin, 2020) is a tool limited to the search of English literary texts, most notably the works of Charles Dickens. On the other hand, many corpus linguistics tools allow users to build and upload their own corpora and therefore are not limited to specific authors or topics. These include, but are by no means limited to, AntConc (Anthony, 2022), LancsBox (Brezina & Platt, 2023), and CQPWeb (Hardie, 2012). These tools are either accessible online or are designed to be downloaded as desktop tools for

users to use offline. All of these tools serve the purpose of allowing researchers to search, filter, and analyse a given corpus. All corpus linguistics tools, therefore, include at minimum the function to search, as Gries (2009, 1226) summarises:

- "frequencies of occurrence of linguistic elements, i.e. how often morphemes, words, grammatical patterns etc. occur in (part of) a corpus, etc.; this information is usually represented in so-called *frequency lists*";
- "frequencies of co-occurrence of these elements, i.e. how often morphemes occur with particular words, how often particular words occur in a grammatical construction, etc.; this information is mostly shown in so-called *concordances*."

Having determined the tools and research questions of corpus linguists, we may be inclined to say that we can define corpus linguistics. But this definition is still today debated in the literature of the field. Is it a method? a discipline? a theoretical paradigm? or none of the above, perhaps. Leech (1992) defines 'computational corpus linguistics', as he calls it "a new philosophical approach." Many other linguists consider it more of a method(ology) (Gries, 2009; McEnery, Xiao, & Tono, 2006). The answer to this debate (which C. Taylor (2008) summarises excellently) is not of great importance to this dissertation. Here, we are interested not in what corpus linguistics *is* but what it can *do*. The methodology of corpus linguistics covers a wide range of fields. To quote Léon (2005, 36) "what is called 'Corpus Linguistics' covers various heterogeneous fields ranging from lexicography, descriptive linguistics, applied linguistics – language teaching or Natural Language Processing – to domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies." I take the same view as Léon and use the term 'corpus linguistics' in this dissertation to refer to the wide range of tools and methods used to collect and analyse linguistic data from a large set of texts, i.e. a corpus.

2.3 Combining Oral History and Corpus Linguistics

Enhancing oral history research with corpus linguistics tools can be traced back to the early days of the British oral history movement when social historians and linguists collaborated closely, actively engaging in the exchange of ideas and methodologies (Roller, 2015). Dialectologists have, for example, long been contributing to the field of oral history by collecting personal stories when interviewing participants in accent and dialect surveys. Georg Wenker, for instance, researched and wrote the *Deutscher Sprachatlas* by collecting data on regional language variations and dialectal differences in Germany through questionnaires sent across the country (Chambers & Trudgill, 1980). In the 1940s we find another exam-

ple of accidental oral history collection from dialectologists in the Origins of New Zealand English (ONZE) project which analysed recordings of spoken language collected by the Mobile Recordings Unit. These recordings were originally collected in 1946-8 by the New Zealand Director of Broadcasting, James Shelly, to "[get] in touch with [musical] talent outside the cities, [...] and making known to New Zealanders in general how life is lived in their own country" (Folwer, *NZ Listener*, 6 February 1946, in Gordon et al. (2004)). These recordings were not initially aimed for linguistic research but have, in the early 21st century, been used to study "acoustic analysis, sociolinguistic variation and speech perception".

While oral history and corpus linguistics have been intertwined in the past, Perks and Robinson (2005) argue that in the past 50 years oral history has drifted away from linguistics and that only now are historians returning to these linguistic sources. One such project is Sealey's (2010) analysis of Millennibum Project (Dick, 2002) which gathered 144 oral history interviews of people living in Birmingham in 2001. Using corpus linguistics tools, Sealey analysed the corpus to identify discourse features of the interviews and to understand self-representation in oral histories. In another paper investigating the same corpus, she explores the issue of constraint on speakers' goals by analysing concordances of the expression *I couldn't* (Sealey, 2012). Another example of how corpus linguistics and oral history can pair together is Fitzgerald's (2022) book on the linguistic construction of clarity. In this work, Fitzgerald investigates the use of mental process verbs and expectation markers in oral history interviews taken from the Corpus of Irish Historical Narratives, illustrating how oral histories can be researched using methods from corpus linguistics. Fitzgerald's aim for this project was "to showcase the potential utility in collaboration between oral historians and linguists" and as so, addressing what Schiffrin (2003, 84) describes as a "dearth of linguistic analyses of oral histories."

There have also been projects of this nature that use Holocaust testimonies (and from which this dissertation draws inspiration). For example, Calamai et al.'s (2021) 'Voices of Ravensbrück' project saw the collection of oral testimonies from survivors of the Ravensbrück concentration camp for women. As listed in section 2.1, there exist many collections of oral histories of the Holocaust in either audio, video or text form. Some of these have been supplemented with annotations, for example, USHMM's Semantic Search (pilot) project contains Holocaust oral testimonies that have been tagged with semantic information (`http://wjbmattingly.com/flask-annoy/`). To the best of my knowledge, this dissertation project is the first of its kind. At present, there is exists no comparable multilingual collection of oral histories of the Holocaust annotated with both syntactic information and named entity data.

2.4 This project: an Annotated Multilingual Corpus of Oral Testimonies of the Holocaust

The objective of this project is to compile a collection of transcribed interviews with individuals who survived or witnessed the atrocities of the Holocaust. These interviews, commonly referred to as oral testimonies or oral histories, have been gathered by various individuals and organisations over the past half-century to preserve firsthand accounts of life in the Third Reich. Each testimony provides a valuable perspective on the impact of Nazism across Europe, whether the individuals were directly persecuted or not. The aim of this project is to gather a significant number of these testimonies to which corpus linguistics methods can be applied, thus enhancing researchers' understanding of the collective experiences of Holocaust survivors. It is important to note that employing corpus linguistics methods does not diminish the significance of individual stories and accounts; rather, it allows for a broader comprehension by examining the commonalities and shared experiences among interviewees. By utilising corpus linguistics tools and methods, researchers can engage in 'distant reading' as proposed by Moretti (2000). Distant reading addresses the limitation of only being able to read a "canonical fraction" of the available literature, and it is a widely employed method in the study of data-rich literary history (Moretti, 2000, 54). Through corpus linguistics methods, researchers can conduct extensive searches across a wide range of testimonies, gaining deeper insights into the prevalent themes discussed. They can then zoom in on individual testimonies and conduct closer readings of specific survivors' accounts.

The corpus of oral testimonies created for this project contains around 50 interviews with Holocaust survivors and witnesses. The interviews were gathered from the United States Holocaust Memorial Museum (USHMM) collection of oral testimonies. The interviews themselves have been recorded either through funded initiatives or donated by individuals. One such initiative is the USHMM Perpetrators, Collaborators, and Witnesses: The Jeff and Toby Herr Testimony Initiative. This a multi-year project aimed to record the testimonies of non-Jewish witnesses to the Holocaust and to add these to USHMM's pre-existing collection of testimonies. The interviews in this collection were recorded anytime between the mid 20th century and the early 2010s. Another portion of the interviews in this corpus were collected by William Helmreich for his book *Against all odds: Holocaust survivors and the successful lives they made in America*. The interviews from this book were donated to the USHMM in 1992. The interviews in both of these collections were conducted in a range of languages. The corpus created for this dissertation brings together testimonies in English, Czech, Hungarian, Polish, and Russian, all of which have been manually transcribed by either the interviewer or

a third party.

3 Methodology

3.1 Building the Corpus

As discussed above, the oral testimonies used to create this corpus were collected by the United States Holocaust Memorial Museum. The USHMM website gives access to a repository of digital items from the museum's collection. Among these is a collection of over 80,000 oral testimonies recorded in the form of either interviews or self tapings with Holocaust survivors and witnesses. The testimonies form a collection of interviews partly recorded by the museum but also from a variety of other holding institutions, most notably the USC Shoah Foundation and the Fortunoff Video Archive. Most of these testimonies are saved in the form of audio mp3 recordings and around 14,000 of them include accompanying video of the interviews. The interviews have been recorded in a variety of languages, including English, Russian, Polish, Czech, Ukrainian, Yiddish, and more. Only a small portion (around 4,000) of the testimonies have been transcribed.

The testimonies collected for the purposes of this project all belong to USHMM's own collection. They have all been manually transcribed and can be downloaded as PDF files directly from the website. For this project, unformatted JSON copies of the transcripts were provided to me by USHMM. The files in this format, while much harder to read due to the lack of spacing or formatting of the metadata (which is included at the start of every transcript), are much easier to edit.

3.2 Cleaning the Files

The transcripts provided by USHMM required cleaning and editing before they could be passed through parsing programmes and uploaded to CQPWeb. The main task when it came to cleaning the files was removing all unnecessary information from the transcript. The original JSON files contained a large amount of metadata including file identifiers, language codes, source identifiers, record type, dates, and more. These metadata needed to be separated from the transcript so that only the interview itself remained. Moreover, when the JSON files were created, it appears that the headers and footers of the original PDF files were not removed and therefore appear at intervals in the body of the text. These too needed to be removed to allow for easier analysis of the interview.

The JSON files were cleaned using oxygenXML (Syncro Soft Ltd., 2023). A batch converted add-on was used to allow for multiple files to be processed at once. The

contents of the interview were extracted from the file by filtering the text to keep only what was included within the `'fnd_content_web'` element. Some manual editing was necessary as not all JSON files had been properly formatted and contained some text metadata in said element. The file identifier (text ID) and content of the transcription were then extracted from the files using the Saxon command-line XML processor. The same thing was done to extract the transcript metadata and save it in another file. Once the metadata was extracted, manual checks were done to ensure that all data was correct. For example, language metadata was sometimes incorrectly noted in the JSON file. Furthermore, certain transcripts did not include interviewer/interviewee names. These could be found by searching for the file names on the USHMM database (`https://collections.ushmm.org/search/`). Some more edits were made using global find and replace functions in a simple text editor to remove remaining extraneous text i.e., the headers and footers from the original PDF files.

The next step in cleaning the files involved speaker diarisation. Speaker turns were marked in the original transcripts with either bullet points, dashes, initials or the letters 'Q' and 'A'. These markers were replaced with XML elements tags <u> for each utterance with 'id' attributes also added, indicating the initials of the interviewer and interviewee (if these had been found). Utterances were then split into sentences using the same technique. These were marked with the element tag <s>. This process was semi-automatically run using a combination of an XML processor and global find and replace.

The output of the cleaning process was a collection of XML files organised into utterances and sentences. They included only the text id and the transcript of the interview (see Figure 1). File metadata was stored in separate XML files.

```
<text id="irn508677">
<u id="JF">
<s>
Mr. Baranski, can you tell us where and when you were born and a little bit about your background?
</s>
</u>
<u id="TB">
<s>
In the 1880's a Polish gentleman, after the third dismemberment of Poland, when Poland disappeared from the map of Europe and the
southern part of Poland was close to the Austro-Hungarian Empire, was ordered to spend his soldier time in Budapest.
</s>
<s>
This gentleman was my grandfather, Albert Baranski.
</s>
<s>
He fell in love with the Hungarian city and after his three years of soldier duty he married in Budapest; Rudolph (or in
Hungarian, Reszl) Baranski, my father, was born there.
</s>
<s>
And since Austria-Hungary was together for 400 years, it was quite natural that he fell in love with a lady who was born in
Wiener Neustadt, a suburb of Vienna, Maria Schelnader, and this lady became my mother.
</s>
<s>
These two persons were unfortunately were too short people in love because later in life my parents got divorced, and this I do
not want to comment on this day.
</s>
<s>
I was born on June 11, 1922, in Budapest, and according to superstitious people, on a very lucky time, Sunday morning, seven
o'clock.
</s>
<s>
I was raised in very, very deep and affectionate love, sometimes maybe a little bit soft love, but God was gracious enough to
give me a lot of illnesses which well balanced so, thanks God I did not become a tempered brat.
</s>
<s>
The greatest influence on my life was done first by my grandfather who was a highly educated person, a very multi-talented
person.
</s>
<s>
[unclear] who was absolutely superb in mathematics and engineering.
</s>
<s>
He was in the calculating office of the First Hungarian Aeronautics Office, only a little nucleus in the end of the 1910's and
beginning of 1920's, but except sports, this gentlemen knew almost every field.
</s>
<s>
He was not excellent, but a fairly good painter.
</s>
```

Figure 1: *Cleaned transcript from Tibor Baranski's oral testimony, in XML format.*

3.3 Translation

As an accompaniment to the original transcripts, non-English testimonies were translated into English. The translation was performed using DeepL (`https://www.deepl.com/translator`), a neural machine translation service. Cleaned versions of the Czech, Hungarian, Polish, and Russian transcripts were uploaded to DeepL to produce English translations. XML tags in the original files were ignored by the software meaning that the translated output contained the same sentence and utterance boundaries. This consistent segmentation of the files was vital to ensure easy alignment at a later stage.

The translated files were saved in XML format and separate metadata files were created to accompany them.

3.4 Tagging

The next step in the preparation of the corpus involved adding syntactic information to the text to allow for linguistic analysis. A first version of the corpus

constituted an effort to explore various methods of syntactic parsing and lemmatisation. A total of 4 parsers were tested during this trial: NLTK (Bird, Loper, & Klein, 2009), spaCy (Honnibal & Montani, 2015), RNNTagger (Schmid, 2019), and TreeTagger (Schmid, 1994). These parsers all function in similar ways, either being run via command-line code or as Python packages (Python Software Foundation, 2021). The output of each programme is a VRT file (where each word is displayed on a new line) with columns displaying the lemma and part-of-speech (POS) tag of every word. To ensure that XML tags were not confused for words by the parsers, an additional line of code was added to tell the programme to ignore all lines of the input file beginning with '<'. Figure 2 shows an example of this output. Here we see that the second column contains the part-of-speech tag for each word according to the Penn Treebank conventions (A. Taylor, Marcus, & Santorini, 2003) and the third gives its lemma (dictionary head-word).

```
<text id="irn508623">
<u>
<s>
I       PRP     I
was     VBD     be
born    VBN     bear
[       NN      [
in      IN      in
1911    CD      1911
]       NN      ]
in      IN      in
Sighet  NNP     Sighet
,       ,       ,
Hungary NNP     Hungary
[       VBD     [
near    IN      near
Hungarian-Romanian      JJ      Hungarian-Romanian
border  NN      border
]       NN      ]
and     CC      and
lived   VBD     live
there   RB      there
until   IN      until
Adolf   NNP     Adolf
Hitler  NNP     Hitler
,       ,       ,
the     DT      the
German  JJ      German
dictator        NN      dictator
,       ,       ,
invaded VBD     invade
Hungary NNP     Hungary
.       .       .
</s>
<s>
At      IN      At
that    DT      that
time    NN      time
part    NN      part
of      IN      of
my      PRP$    my
family  NN      family
```

Figure 2: *Transcript of Margaret Beer's oral testimony parsed using NLTK, in VRT format.*

3.4.1 Universal Dependency Tagging

During the development of the beta version of the corpus, a significant challenge emerged when each tagger produced different POS tags for each language. This

disparity arose due to syntactic parsers' reliance on language-specific models for tagging and lemmatising texts, resulting in diverse POS tagsets across languages (e.g., Czech Academic Corpus tags, magyarlanc Hungarian tags, British National Corpus tags, etc.). This variation in POS tags posed a considerable obstacle, as it necessitated users of the corpus to familiarise themselves with distinct tagging conventions for each language. Additionally, it hindered the ability to conduct cross-linguistic searches using linguistic annotations effectively. To address this issue, the decision was made to adopt universal POS tagging for the second iteration of the project, employing the Universal Dependency POS tagset.

The Universal Dependency (UD) guidelines (McDonald et al., 2013) were created to "to develop cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective." The UD annotation scheme is based on the 'Interset interlingua' for morphosyntactic tagsets (Zeman, 2008), Google universal part-of-speech tags (Petrov et al., 2012), and the universal Stanford dependencies (De Marneffe et al., 2014). Alongside the UD POS tagsets, the corpus was tagged with detailed UD features (e.g., gender, person, number, mood, etc.). This is an extensible tagset where individual features are concatenated to create a unique tag (e.g., Case=Nom|Number=Sing|Person=1|PronType=Prs). The UD tag scheme allows users to search the entire corpus at once.

A potential drawback of this approach is that applying a standardised tagset, originally designed for one language, to multiple languages may lead to an imposition of grammatical structures and linguistic descriptions from one language onto others, regardless of their suitability. Ultimately, the goal of using UD tags across the 5 languages in this corpus is consistency. The standardised tags ensure uniformity in the annotation scheme, making it easier to compare and contrast linguistic features across different languages in the corpus. Language specific tagsets were thus also added to the parsing to compliment the UD tags. English testimonies, for example, were tagged with the Penn Treebank tagset (A. Taylor et al., 2003) while Czech testimonies were tagged with the Czech Academic Corpus tagset (Hladká et al., 2008). The tagsets used for each sub-corpus are linked in the 'Corpus Info' sidebar on CQPWeb to allow users to reference the tag scheme.

3.4.2 Named Entity Recognition

Along with universal tagging, several more improvements were made to the corpus. Notably, enhancing the tagging with Named Entity Recognition (NER). NER is the process of identifying and categorising proper nouns or nouns denoting a named entity.

There are many NER programmes freely available online. For this project, the widely used OntoNotes (large) package was utilised as it includes the most tags and is generally considered to be highly accurate. Where a word is not considered to be a named entity it is marked with '0'. As with the POS tagsets, the OntoNotes tagset was linked in the 'Corpus Info' sidebar to allow users to refer to the tag scheme when conducting queries.

3.4.3 Stanza

The Stanza Python package allows for users to tokenise, tag, and lemmatise texts in over 70 languages (Qi, Zhang, Zhang, Bolton, & Manning, 2020). This programme was used for the second version of the Holocaust Testimonies corpus. Each text was tagged using UD POS tags, language specific POS tags, UD language features (a standardised annotation that gives information on case, gender, number, tense, mood, and more of each word), and NER tags. As with the files in the beta version, the transcripts were output in VRT format with each token on a new line. The columns in each file were organised in the following format: token, UD POS tag, lemma, language specific POS tag, UD feature, NER tag. Figure 3 gives an example of an output file in the new version of the corpus.

```
<text id="irn508623">
<u>
<s>
I       PRON    I       PRP     Case=Nom|Number=Sing|Person=1|PronType=Prs      O
was     AUX     be      VBD     Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   O
born    VERB    bear    VBN     Tense=Past|VerbForm=Part|Voice=Pass     O
[       PUNCT   [       -LRB-   None    O
in      ADP     in      IN      None    O
1911    NUM     1911    CD      NumForm=Digit|NumType=Card      S-DATE
]       PUNCT   ]       -RRB-   None    O
in      ADP     in      IN      None    O
Sighet  PROPN   Sighet  NNP     Number=Sing     S-GPE
,       PUNCT   ,       ,       None    O
Hungary PROPN   Hungary NNP     Number=Sing     S-GPE
[       PUNCT   [       -LRB-   None    O
near    ADP     near    IN      None    O
Hungarian       PROPN   Hungarian       NNP     Number=Sing     B-NORP
-       PUNCT   -       HYPH    None    I-NORP
Romanian        ADJ     Romanian        JJ      Degree=Pos      E-NORP
border  NOUN    border  NN      Number=Sing     O
]       PUNCT   ]       -RRB-   None    O
and     CCONJ   and     CC      None    O
lived   VERB    live    VBD     Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin   O
there   ADV     there   RB      PronType=Dem    O
until   SCONJ   until   IN      None    O
Adolf   PROPN   Adolf   NNP     Number=Sing     B-PERSON
Hitler  PROPN   Hitler  NNP     Number=Sing     E-PERSON
,       PUNCT   ,       ,       None    O
the     DET     the     DT      Definite=Def|PronType=Art       O
German  ADJ     German  JJ      Degree=Pos      S-NORP
dictator        NOUN    dictator        NN      Number=Sing     O
,       PUNCT   ,       ,       None    O
invaded VERB    invade  VBD     Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   O
Hungary PROPN   Hungary NNP     Number=Sing     S-GPE
.       PUNCT   .       .       None    O
</s>
<s>
At      ADP     at      IN      None    O
that    DET     that    DT      Number=Sing|PronType=Dem        O
time    NOUN    time    NN      Number=Sing     O
part    NOUN    part    NN      Number=Sing     O
of      ADP     of      IN      None    O
my      PRON    my      PRP$    Case=Gen|Number=Sing|Person=1|Poss=Yes|PronType=Prs     O
family  NOUN    family  NN      Number=Sing     O
```

Figure 3: *Transcript of Margaret Beer's oral testimony tagged using Stanza, in VRT format.*

Once all the transcripts had been tagged and saved in individual VRT files, they were combined into a single VRT file for each language. The English translations of the Czech, Hungarian, Polish, and Russian interviews were first stored in individual files according to original language (these would later be aligned with the originals to allow for parallel translations in CQPWeb). These were then combined with duplicate files of the English interviews (those originally recorded in English) to create a single corpus containing all 47 interviews in English which would allow for all testimonies to be searched at once[1].

3.5 Metadata

As previously discussed, the metadata for each testimony were recorded in separate files so as not to clutter the interview files. The original transcripts of the testimonies as provided by USHMM contained a vast array of metadata including but not limited to: names of interviewer and interviewee, dates of the interview

[1] These files are available in Appendix A

(if such data existed), dates of acquisition by the Museum, language(s), historical provenance, original format (e.g., cassette recordings), accession identifier, etc. Of these metadata, I decided to record the following:

- Index Reference Number (IRN);
- Link to audio recording on the USHMM website;
- Name of interviewee;
- Archival accession identifier;
- (Original) Language.

As the testimonies were organised by language, the metadata for each interview were stored in separate TSV files for each interview language, including one for all files translated into English[2]. Care was taken to ensure that the metadata were recorded in the same order as the interviews in their respective languages as this would allow CQPWeb to identify the correct metadata for each transcript.

irn50713	https://collections.ushmm.org/search/catalog/irn50713	Karel Vašiček	2011.439.1	Czech	
irn50714	https://collections.ushmm.org/search/catalog/irn50714	Jirí Reynek	2011.439.2	Czech	
irn50715	https://collections.ushmm.org/search/catalog/irn50715	Daniel Reynek	2011.439.3	Czech	
irn50718	https://collections.ushmm.org/search/catalog/irn50718	Vojtěch Peřina	2011.439.5	Czech	
irn50851	https://collections.ushmm.org/search/catalog/irn50851	Bohumil Vrátil	2011.439.6	Czech	
irn50852	https://collections.ushmm.org/search/catalog/irn50852	Zdena Volfová	2011.439.7	Czech	
irn50853	https://collections.ushmm.org/search/catalog/irn50853	Jarmila Špírková	2011.439.8	Czech	
irn50855	https://collections.ushmm.org/search/catalog/irn50855	Jaroslav Samek	2011.439.9	Czech	

Figure 4: *Czech corpus metadata, in TSV format.*

3.6 Uploading and Aligning the Corpora

With the corpus ready for analysis, it was uploaded to a corpus linguistics tool which would allow it to be searched and analysed. The chosen tool was the Oxford CQPWeb website which was set up for researchers and students at the University of Oxford to share their text corpora.

Each of the sub-corpora and their respective metadata files were uploaded to CQP-Web via the admin control panel. For each corpus, s-attributes (XML tags) and p-attributes (word annotation information) is given. The p-attributes are those added by Stanza and are as follows:

- Universal Part of Speech tag;
- Lemma;
- Language Specific Part of Speech tag;

[2]These files are available in Appendix B

- Universal Language Features;
- Named Entity Recognition tag.

Documentation for each tagset was linked alongside the p-attributes to allow users to find a description of each tag.

Once all the sub-corpora were uploaded, individual transcripts could be aligned with their translated counterparts. Alignment was done using the CQP server command line to align the two VRT files according to the utterance boundaries (i.e. the XML element tag <u>). Once aligned, searches in either the original corpus or its translation will display the section of text in the other that matches the search.

4 Pilot Studies

In this section I will present two pilot studies conducted using data gathered from the corpus on CQPWeb. The hope is that these studies will give an idea of what kinds of research can be performed on a corpus of Holocaust Testimonies when using corpus linguistics tools. As these are intended as demonstrations of what can be done, I will only briefly discuss the methods used and the results obtained. Further discussion and analysis of these results would require more background on holocaust studies and discourse analysis than fits within the scope of this paper.

The first study uses location data gathered from Named Entities recorded in the complete English (including translations) corpus. The second study uses syntactic data gathered from Universal Dependency Features in the same corpus.

4.1 Mapping Locations and Movement in Oral Testimonies

4.1.1 Description

Using the CQPWeb Syntax query box we can search for Named Entity Recognition data. The NER tags were created by Stanza using the OntoNotes Large model. This model identifies named entities and sorts them into 18 categories including: dates, persons, languages, geopolitical entities, organisations, and more. This pilot study looks into the countries that are named by the Holocaust survivors and witnesses whose testimonies are collected in the 'Oral Testimonies of the Holocaust in English (incl. translations) (HT02)' corpus. Tagging the corpus with a NER programme provides us with additional information on words and allows us to undertake research on those specific word types such as looking up names of persons, locations, or dates mentioned, for example. These types of annotation

allow for the exploration of social networks, geographic locations, or key events mentioned by the interviewees.

Personal testimonies and oral histories are valuable sources of information on patterns of movement, migration, and displacement. Kohler (1901), for example, argues that personal narratives should be seen as interactive texts, aiding in recovering lost geographies and amplifying alternative voices in academic discourse. Moreover, oral histories add a layer of personal truth by adding the individual back into the wider narrative on migration. Witcomb (2019) for instance, explores how oral history enriches exhibitions about migration, adding a personal layer of interpretation to objects and prioritising individual stories as key narrative elements. In this study, I collect the names of countries mentioned in Holocaust oral testimonies in the hope of understanding whether they reflect patterns of movement or migration of survivors post war. For example, were there common escape routes or patterns of displacement? This initial investigation of the corpus will serve as a foundation for us to know which countries are the most and least mentioned by survivors and whether these refer to destinations chosen for emigration and resettlement post war. We can expect the mentions of countries to reflect the sample of interviewees present in the corpus. For example, many interviews were collected in Hungary and the U.S. so we can expect those countries to be highly represented.

4.1.2 Methodology

In order to collect location data from the corpus on CQPWeb, users can make use of the in-built CQP syntax function when creating a query. Information on how to use CQP syntax can be found in manual provided on the portal (`http://cqpweb.ling-phil.ox.ac.uk/CQPweb/doc/cqpweb-simple-syntax-help.pdf`). Named Entity information is stored in the 'ner' P-attribute for every corpus in this project. Following the OntoNotes conventions, countries, cities, towns, and villages are recorded as geopolitical entities and their NER tag is 'GPE'. Using CQP syntax we can thus search for all words that have been tagged with this NER tag in CQPWeb:

```
[ner=".*-GPE"]
```

This returns a list of 3,696 words that have been tagged as geo-political entities by the OntoNotes programme that was run on Stanza. This simple query organises tokens by text so that the first tokens listed are those in the first text in the corpus: Karel Vašiček's interview (originally in Czech). Tokens are further organised in the order in which they appear in the transcript. These tokens, or concordances, are given in context, with 12 tokens to both the left and right. Next, we select the 'Frequency Breakdown' option to view the list of tagged words and the number

of times each was used in the corpus (as well as this number represented as a percentage of all words tagged 'GPE'). This search gives us the following result: 752 different types and 3,696 tokens across 47 different texts . This frequency list was searched to find alternate names, such as 'U.S.' and 'America' or 'Hungaria' and 'Hungary' which were then aggregated. Lastly, the counts for cities and towns and the countries to which they belong were aggregated so that the, for example, 'Warsaw' and 'Auschwitz' were counted towards the count for 'Poland.' This cleaning resulted in a list of 40 countries and their occurrence counts (Table 1).

	Country	Count		Country	Count
1	Germany	470	21	Latvia	10
2	Hungary	447	22	Slovakia	10
3	United States	230	23	Spain	12
4	Russia	130	24	Portugal	8
5	Israel	120	25	Hong Kong	3
6	Romania	114	26	Moldova	3
7	Czechia	173	27	Finland	4
8	Poland	425	28	Belgium	4
9	Austria	60	29	Sweden	4
10	United Kingdom	53	30	Argentina	1
11	Netherlands	43	31	Serbia	1
12	Lithuania	45	32	Bolivia	1
13	France	47	33	Bermuda	1
14	Ukraine	66	34	Cyprus	1
15	Italy	36	35	Cuba	1
16	Switzerland	31	36	Denmark	1
17	Palestine	21	37	Ireland	1
18	China	67	38	Japan	1
19	Canada	16	39	Macedonia	1
20	Australia	4	40	Norway	1

Table 1: *Aggregated count of locations named in Holocaust Oral Testimonies by country.*

We can then use Tableau (`https://www.tableau.com`), a data visualisation software to create a choropleth map of the countries named. This map (Figure 5) indicates which locations were most referred to in the Oral Testimonies of the Holocaust corpus.

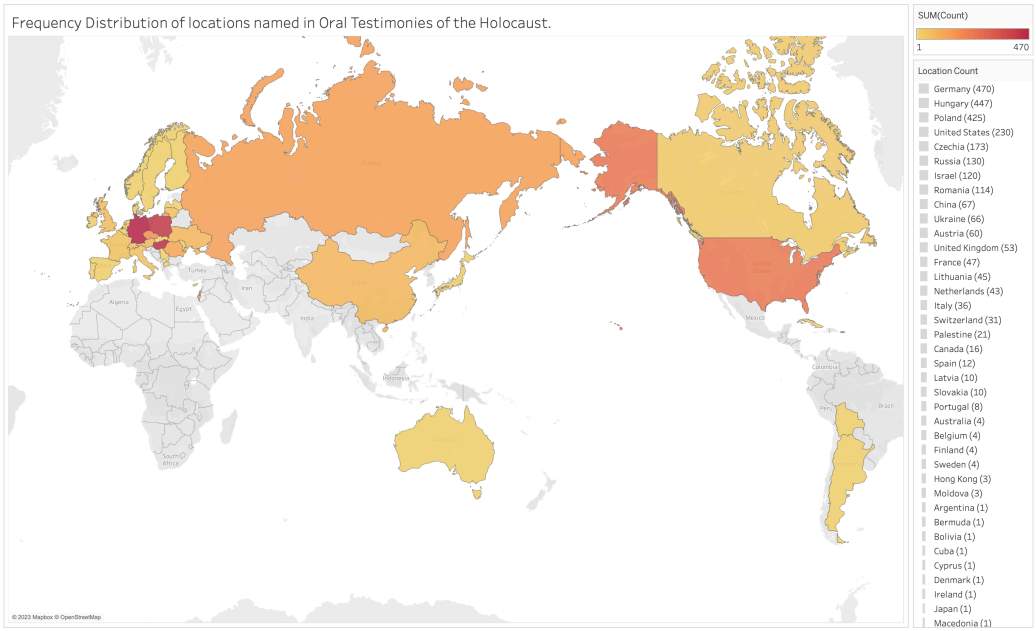


Figure 5: *Choropleth map of countries named in the complete Holocaust Testimonies, created using Tableau.*

Based on the data collected in Table 1, close reading of the testimonies was undertaken so as to gain a better understanding of the contexts in which these locations were talked about. The aim of this study, as previously mentioned, is to understand patterns of movement or migration of survivors post war. In order to achieve this, the corpus was organised by individual testimony and subsequently searched for GPE mentions (using the CQP syntax query tool). Each mention of a location was then analysed to determine whether the interviewee was discussing a destination of emigration or escape used by either themselves or people they knew from their home countries. This information was stored in a dataset (Appendix C) which set could then be visualised using Tableau (Figure 6). An interactive version of this map allows users to view individual migration routes between countries with tooltips revealing labels that indicate the number of times this migration route was mentioned in the corpus[3] .

[3]Available at: `https://public.tableau.com/views/HT02_Migration/Sheet1?:language=en-GB&:display_count=n&:origin=viz_share_link`

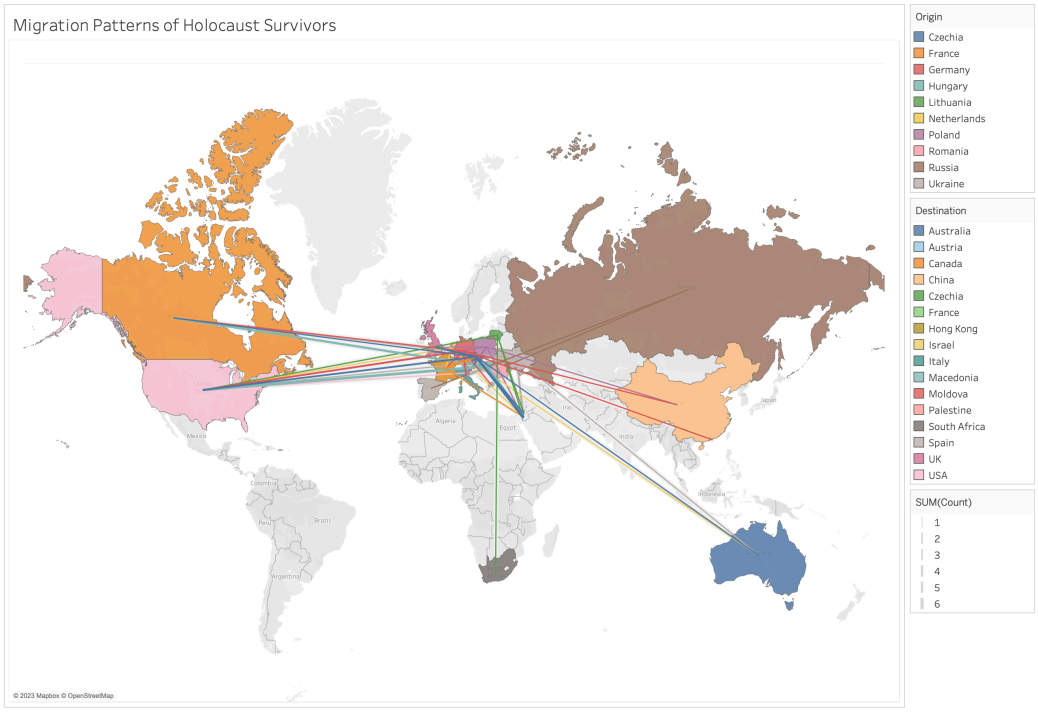


Figure 6: *Migration routes named in the complete Holocaust Testimonies, created using Tableau.*

4.1.3 Results

The choropleth map (Figure 5) of countries named by Holocaust survivors and witnesses in our corpus shows that the 5 countries the most talked about are Germany, Hungary, Poland, the US, and Czechia. Countries in Africa, Asia, and South America go unmentioned in this set of oral testimonies, save a few mentions of China and Cambodia. The close reading of locations named in our corpus of oral testimony reveals that survivors of the Holocaust emigrated to many different areas of the globe. Among the most reported destinations of relocation after the war are the United States, Isreal, and Canada. These trends are likely due, in part, to the fact that our corpus of testimonies was collected by the United States Holocaust Memorial Museum and therefore contains a large amount of interviews that were recorded in the US with people who had settled there. This map is by no means meant to be an accurate representation of migration routes taken by Holocaust survivors. Rather, my aim is to demonstrate how corpus linguistics tools can be used to extract large amounts of data from oral testimonies which can in turn be analysed and visualised with a range of computational tools.

4.2 Hedging Techniques in Oral Testimonies

4.2.1 Description

Roller (2015) argues that oral testimonies can provide valuable insights for linguistic research, encompassing various areas such as sociolinguistic studies of sociolects, language variation and change, narrative discourse analysis, and grammatical variation, among others. Spoken corpora, in particular, are an excellent of source of data on spontaneous speech in a variety of environments, dialects, and registers. This pilot study looks to compare the difference in speech between spontaneous conversation of adults and the language used in oral testimony interviews. Specifically, I am interest in the use of hedging. Hedging, as described by Biber (1991), refers to linguistic devices and strategies used to mitigate or soften the force or certainty of statements. It involves the use of language that allows for flexibility, caution, or ambiguity in expressing opinions, beliefs, or claims. Hedging has previously been studied using corpus-based methods (Aijmer, 1986; Kaltenböck, Mihatsch, & Schneider, 2012). Often, hedging is studied as part of general studies into spoken language. Notable examples include Biber (1991) and Precht (2003), both of whom use Multiple Factor Analysis (MFA) to identify groups of factors that tender to occur together in language. Other studies, such as Fitzgerald (2022) have studied specific hedging techniques in spoken corpora, in this case, mental process verbs. Fitzgerald's study highlights the high frequency of mental process verbs used in a corpus of oral histories. These verbs, such as *I think*, serve the dual purpose of expressing certainty by establishing authority and objectivity, while also conveying uncertainty through tentativeness ((Holmes, 1990, 187); (Aijmer, 1996, 21)).

This pilot study thus first examines how the language of Holocaust survivors and witnesses in their interviews compares to spontaneous speech with regards to hedging. In this study, I collect frequency data on a number of linguistic hedging techniques such as mental process verbs, specific adverbs, and clauses. This database is then compared to reference corpus in order to establish whether there is a difference in speech in oral testimony interviews with spontaneous speech. The reference corpus used is the London-Lund corpus of spoken English (LLC), a corpus of approximately 1 million words of spontaneous spoken English collected in the 1960s (Svartvik, 1990). The corpus can be searched on the CLARINO corpus management and analysis system, Corpuscle (`https://clarino.uib.no/korpuskel/home`).

As MFA requires far more data than would be feasible to collect for a pilot study of this size, simple Mann-Whitney U tests were run on R (R Core Team, 2018). With a larger sample size and by comparing more than 2 data sources (i.e., more

than the Holocaust Testimonies corpus and the LLC), MFA could be performed. This would allow for a more detailed understanding of which hedging techniques can be grouped together and whether there is a significant difference in the way people speak in oral history interviews and in spontaneous conversations.

4.2.2 Methodology

First, in order to verify Fitzgerald's claim that mental process verbs are the most common lexical verbs in oral testimonies, we can perform a simple CQP Syntax query of all verbs in the corpus. Using the syntactic information provided by Stanza, the Universal Dependency features provide information on case, gender, mood, number, aspect, tense, and more. These features are given as a bar-separated list such as:

Mood=Ind|Number=Plur|Person=2|Polarity=Pos|Tense=Pres|

To find all instances of present tense verbs used in the first person singular we can use the following query:

```
[word="I"][feats=".*Tense=Pres.*"]
```

This query returns 4,917 matches in 47 different texts, which equates to a frequency of 8,381.788 instances per million words. The results of this query can be viewed in a Frequency Breakdown table, the first 34 lines of which are shown in Table 2 (the full list can be found in Appendix D). Here, we can see that some of the most frequently used lexical verbs in the present tense first person singular are *I think, I know, I remember* and *I mean*, all of which are mental process verbs.

	Search result	Count	Percent		Search result	Count	Percent
1	I do	1601	32.56%	18	I give	20	0.41%
2	I think	420	8.54%	19	I come	19	0.39%
3	I'm	361	7.34%	20	I look	19	0.39%
4	I know	329	6.69%	21	I suppose	16	0.33%
5	I have	290	5.90%	22	I like	15	0.31%
6	I remember	284	5.78%	23	I live	15	0.31%
7	I mean	228	4.64%	24	I ask	14	0.28%
8	I am	227	4.62%	25	I get	13	0.26%
9	I say	176	3.58%	26	I hope	13	0.26%
10	I've	109	2.22%	27	I wonder	13	0.26%
11	I guess	100	2.03%	28	I going	12	0.24%
12	I want	89	1.81%	29	I wish	12	0.24%
13	I see	74	1.50%	30	I says	11	0.22%
14	I tell	73	1.48%	31	I thank	11	0.22%
15	I understand	51	1.04%	32	I believe	10	0.20%
16	I go	27	0.55%	33	I need	10	0.20%
17	I feel	24	0.49%	34	I recall	10	0.20%

Table 2: *Frequency breakdown table of 1.sg.pres verbs*

Next, I collect other occurrences of hedging techniques ranging from clause types and modals, to general hedging words and phrases such as *potentially* and *sort of*. The full list of hedges that I analysed are available in Appendix E. the frequency of hedging techniques is collected using either the CQPSyntax or simple query function on CQPWeb.

This data is then compared to identical data taken from a corpus of spoken English in order to determine whether there is any difference in the use of hedging techniques in regular speech compared to the language used in oral histories. As mentioned, the reference corpus used for this study was the London-Lund Corpus. The frequency (occurrences per million words) of each hedging technique is normalised using z-scores (data for which can also be found in Appendix E). Mann-Whitney U tests are then run on R for each category of hedging technique: metal process verbs, hedging adverbs, modals, whether and that clauses, and personal pronouns. I am employing the Mann-Whitney U test to compare the two corpora as it is a non-parametric statistical test that assesses whether there is a significant difference in the distributions of hedging techniques. This test allows for a robust analysis without relying on assumptions about data normality or equal variances.

4.2.3 Results

Examining the present tense verbs, we can use the frequency breakdown to identify the most commonly used ones. By excluding auxiliary verbs like *do, be,* and *have* ,CQPWeb reveals that *know* and *think* rank as the most frequently used verbs in the present tense. Close behind are the mental process verbs *remember, mean* and *guess*. Out of the 151 first singular present tense verbs recorded in the corpus, 62 fall under the category of mental process verbs. This means that approximately 41% of these verbs may indicate hedging. Mental process verbs often function as hedges since they allow speakers to present their ideas or beliefs as personal opinions or subjective experiences, rather than asserting them as absolute truths. This finding supports Fitzgerald's (2020) prediction that present tense verbs in oral testimony interviews frequently comprise mental process verbs used by interviewees to express their opinions, beliefs, or claims.

Looking next at the overall data on hedging, we find the following p-values of the Mann-Whitney U tests:

- Mental process verbs: 0.6961
- Hedging adverbs: 0.9015
- Modals: 0.9372
- Personal pronouns: 0.6857

The overall p-value (including data on clauses, which were excluded as there were only two categories) is: 0.8726. As the p-value is significantly smaller than the standard significance level (p=0.05), we can say that there is no significant difference in the distributions of hedging technique frequencies between the two corpora. There certainly appears to be differences in how people speak in spontaneous conversation and in oral history interviews when looking at individual statistics on features of hedging e.g., *I don't know* and *I don't remember* appear far more frequently in the corpus of oral testimonies than in the LLC (1738.748 and 288.087 occurrences per million words, respectively compared to 372.631 and 12.849 occurrences per million words in the LLC). The statistical tests, however, indicate that overall there is no significant difference in the use of language. This result may be due to the fact that only a small number of hedges were collected, and more importantly only two categories were compared (our Holocaust corpus and the LLC). More testing should be done to compare the language of oral testimony interviewees with that of spontaneous speech with a much larger sample size and with more techniques. A larger quantity of data would allow for MFA and studies emulating Biber (1991) and Precht (2003).

5 Discussion

5.1 Distant Reading of Oral History

Oral history and the collection of firsthand witness accounts have played an important role in Holocaust studies. In this dissertation, I argue that by using an interdisciplinary approach that combines oral history and corpus linguistics, researchers can broaden their understanding of the lived experience of those who suffered and witnessed Nazism under the Third Reich. Utilising existing collections of firsthand accounts and testimonies of the Holocaust, researchers can pair these rich data sets with tools and methods of analysis taken from corpus linguistics to perform distant reading of many histories at once. The term 'distant reading' (also known as text analysis or text mining), as coined by Moretti (2000), refers to a methodological approach in literary analysis that focuses on the analysis of large-scale literary data sets rather than close examination of individual texts. Methodological focus is placed on computational and quantitative methods to analyse patterns and trends across a corpus of texts. Moretti's (and other early proponents of distant reading) work focused on the comparative literature studies. He argued that this approach to literary study allows researchers to identify and analyse macro-level trends, such as genre evolution, narrative structures, and cultural shifts (Driscoll & Pierazzo, 2016). (Biber, 1991, 61) calls this type of research "macroscopic analysis" which he contrasts with "microscopic analysis [which] provides a detailed description of the communicative functions of particular linguistic features (e.g., clefts as markers of informational prominence, or first person pronouns as markers of personal involvement)". Distant reading or macroscopic analysis, thus provides insights into the historical, social, and cultural contexts in which literary works are produced which, in turn, help to deepen traditional close readings of individual texts.

Furthermore, the two pilot studies explored above demonstrate how augmenting testimonies with annotation can add a great value to the possible research outcomes. By incorporating these annotations into the testimonies, researchers gain access to a wealth of structured linguistic information that opens up new avenues for analysis and interpretation. The first pilot study on locations named in the Oral Testimonies of the Holocaust corpus shows that NER annotation adds a valuable layer of information to the testimonies. NER software identifies and tags named entities, such as names of people, places, organisations, or specific events, allowing researchers to uncover connections, affiliations, and contextual references within the testimonies. This annotation allows for the exploration of social networks, geographic locations, or key events mentioned by the interviewees. Researchers can then perform close reading of individual testimonies to examine how interviewees refer to and contextualise these named entities, offering insights into their personal

experiences, perceptions, and relationships. The syntactic annotation used in the second pilot study provides insights into the grammatical structure and relationships within the testimonies. It allows for the identification of syntactic patterns, sentence structures, and grammatical constructions used by the speakers. This information enables researchers to explore the syntactic variations and strategies employed by individuals when recounting their experiences. Such analysis can shed light on the narrative style, rhetorical devices, and linguistic choices used in oral testimonies, contributing to a deeper understanding of how individuals shape and frame their narratives. The augmentation of testimonies with annotations facilitates comparative analysis and cross-referencing across multiple testimonies. Researchers can identify shared syntactic structures, recurring named entities, or thematic patterns across different interviews. This comparative approach allows for the identification of common experiences, linguistic features, and sociocultural dynamics within the testimonies.

Lastly, I believe that oral histories remain an untapped mine of linguistic data and I agree with Schiffrin (2003) that there is still a distinct lack of linguistic analysis being conducted on oral histories. Researchers like Sealey (2010, 2012) and Fitzgerald (2022) are among the very few making use of the many vast collections of oral histories as linguistic corpora. Oral history interviews are an excellent source of (relatively) unprimed language spoken in natural contexts which can be used to study, for example, language variation and change and discourse analysis. Moreover, as oral histories are typically accompanied with some amount of metadata on the interviewee, linguists can garner a lot of information required for sociolinguistic research. This is not to mention the fact that oral histories can (sometimes unintentionally) function as preservation projects for endangered languages. Applying corpus linguistics methods to these many and varied language data sets can allow linguists to draw conclusions on how language is used by certain individuals, at given times, and on various topics.

5.2 Ethical Considerations

It is important to note that the application of distant reading methods to oral history research requires careful consideration of ethical considerations and the potential loss of nuanced contextual information.

A consideration to keep in my mind is the difference between using the audio or video recordings of testimonies versus their written transcripts. As Einwohner (2011, 422) points out, Holocaust testimonies include "a great deal of what could be considered sensitive material; respondents routinely cried while giving their testimonies, and it was obvious from reading the transcripts that giving testimony was a difficult experience for many." As corpus linguistics and text analysis are

based on the study of written data, using transcripts of the interviews was unavoidable. In order to address this ethical issue, I ensured that the testimonies' metadata on CQPWeb included links to the USHMM website on which the video or audio recordings were available. This means that users of the corpus could easily find and listen to the interviews, helping them not to lose track of the survivors and witnesses behind the words.

Another ethical consideration is one on which I have touched multiple times: losing sight of the individual testimonies when using distant reading methods on a large corpus. Distant reading and corpus linguistics serve as complementary research methods to more qualitative, in-depth analyses of individual testimonies. It can help identify recurrent themes and topics across multiple testimonies thereby prompting researchers to ask new questions and perform a deeper analysis of these topics. We must be wary of interpreting and presenting results from distant reading research without taking into account the lived experiences and narratives of the individuals behind the testimonies. We must avoid reductionism in our presentation of results and ensure that our findings are presented in a manner that is sensitive to the complexities and nuances of the oral testimonies.

As the data used for this project are publicly accessible interviews obtained from individuals who willingly provided their consent for their testimonies to be shared in this particular manner, no considerations into data protection, sharing, or anonymisation were necessary. Other corpus linguistics projects wishing to use oral histories should consider restrictions on access and use and contemporary oral history projects should consider informed consent and anonymity of participants. These considerations in mind, I believe that when used thoughtfully, distant reading techniques can provide valuable insights into large-scale oral history collections and complement more traditional qualitative close reading methods.

5.3 Areas of Improvement

The project presented in this dissertation is in its second version. The source material is unchanged from the first version presented at the EHRI-CLARIN workshop. Strides have however been made in improving the process to tokenise, lemmatise, and parse the testimonies. The first iteration made use of four different parsing programmes each with varying levels of success. Moving to a single programme not only helped to unify the output and ensure that all parsed files were identical in format and structure, but also helped cut down on preparation and programming time. If in future an updated parsing programme is used to create better parsing, my recommendation is that said programme is able to read, tokenise, lemmatise, and parse all languages in the corpus. If this is not the case, multiple parsers will need to be used to process all the various sub-corpora. Furthermore, the second

version of the corpus provides Universal Dependency tags alongside the language specific tags. This improvement is twofold: on the one hand, as UD is one of the most widely used tagsets and is relatively easy to understand, using UD improves the user experience by making it easy to perform searches with simple and logical tags, on the other hand, using the UD tagset ensures that all testimonies, no matter the language have the same tags. This means that users can search for a single syntactic tag (e.g. 'VERB') across all testimonies. Another improvement from the first version of the corpus is the addition of links to the interviews on the USHMM website. As discussed above, this upgrade allows users to access the full transcripts and listen to the interviews in their original language. This is particularly important in helping users put a face to the testimonies they are reading and ensuring that the survivors and witnesses are at the centre of the research.

As for areas of the project that could benefit from further improvement, I think that the priority should be on adding more testimonies to the corpus. As it stands, the Oral Testimonies of the Holocaust in CQPWeb contains half a million words across 47 interviews. This number is enough to give users an idea of what life under the Third Reich was like but is by no means enough to draw conclusive findings. Should this project be recreated, my recommendation is that at least twice the number of testimonies be used. This would hopefully bring the number of words to a million which is more in line with the size of other small corpora. Work should also be done to improve the translation and alignment of testimonies. The scope of this dissertation project was limited by the number of testimonies provided by USHMM that had already been transcribed. Manual transcription of more testimonies was not within the scope of the project as was the use of speech to text services which are often expensive. Furthermore, this iteration of the project made use of the free version of DeepL to translate interviews conducted in Czech, Polish, Hungarian, and Russian. With more time and money, improved translation services could be used whether these be human or machine made. As for the alignment of original texts and their translated counterparts, the internal CQPWeb service was used. A few errors slipped through and users may find that certain queries return results without aligned translations. A more powerful alignment tool based on more than the XML tags would hopefully improve the user experience and ensure that all queries return both original and the aligned translated portions of the testimonies.

6 Conclusion

This dissertation has presented a comprehensive exploration of creating a digital cross-linguistic corpus of oral testimonies, highlighting its potential in advancing

research within the digital humanities. A case study was used to demonstrate this which involved the creation of an annotated multilingual corpus of oral testimonies of the Holocaust. The corpus is organised into five sub-corpora based on language, with the non-English testimonies translated into English for inclusion. By aggregating the translated testimonies with the original English testimonies, a single sub-corpus was formed, encompassing all 47 testimonies collected for this project. Annotations were created using a parsing programme implemented via Python that tokenised, lemmatised, and parsed each sub-corpus. Annotations were made using both Universal Dependency tagsets and language specific tagsets. Named Entity Recognition software was also used to tag named entities in the transcripts using the OntoNotes tagset. The entire corpus was uploaded to an online corpus linguistics tool, CQPWeb (Oxford). This tool allows users to perform queries on corpora to search for concordances and frequency list of tokens, lemmas, syntactic tags, or NER tags. This type of research allows for users to perform distant reading of multiple testimonies at once, thus allowing them to compare testimonies and find common themes that warrant further close reading.

In conclusion, this dissertation provides an in-depth investigation into the creation of a digital cross-linguistic corpus, showcasing how oral history research can be supplemented and improved with tools and methods taken from corpus linguistics. By using distant reading techniques and computational analysis, I hope that oral historians and corpus linguists alike can continue this line of interdisciplinary research to deepen our understanding of the lived experiences of those who survived and witnessed the Holocaust.

References

Aijmer, K. (1986). Discourse variation and hedging. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics ii* (pp. 1–18). Brill.

Aijmer, K. (1996). I think - an English modal particle. *Trends in Linguistics Studies and Monographs*, *99*, 1–48.

Anthony, L. (2022). *Antconc (version 4.2.0)*. Retrieved from `https://www.laurenceanthony.net/software`

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. Retrieved from `https://www.nltk.org`

Brezina, V., & Platt, W. (2023). *LancsBox X*. Retrieved from `http://lancsbox.lancs.ac.uk`

Calamai, S., Beeken, J., Heuvel, H., Broekhuizen, M., Hessen, A. v., Draxler, C., & Scagliola, S. (2021). Voices from Ravensbrück. Towards the creation of an oral and multi-lingual resource family. In *Clarin annual conference 2021*.

Chambers, J., & Trudgill, P. (1980). Dialectology cambridge university press. *ChambersDialectology1980*.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Lrec* (Vol. 14, pp. 4585–4592).

Dick, M. (2002). *Millennibrum: Bringing birmingham history to life cd-rom: Editor and main contributor to millennibrum: Bringing birmingham history to life*.

Driscoll, M. J., & Pierazzo, E. (2016). *Digital scholarly editing: Theories and practices* (Vol. 4). Open Book Publishers.

Einwohner, R. L. (2011). Ethical considerations on the use of archived testimonies in Holocaust research: Beyond the IRB exemption. *Qualitative Sociology*, *34*(3), 415.

Fitzgerald, C. (2022). *Investigating a Corpus of Historical Oral Testimonies: The Linguistic Construction of Certainty*. Taylor & Francis.

Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., & Trudgill, P. (2004). *New Zealand English: its origins and evolution*. Cambridge University Press.

Grele, R. J. (1996). Directions for oral history in the United States. *Oral history: An interdisciplinary anthology*, *2*, 62–84.

Gries, S. T. (2009). What is corpus linguistics? *Language and linguistics compass*, *3*(5), 1225–1241.

Hamilton, P., & Shopes, L. (2008). *Oral history and public memories* (Vol. 79). Temple University Press.

Hansard at Huddersfield Project. (2018). *Hansard at huddersfield.* Retrieved from `https://hansard.hud.ac.uk`

Hardie, A. (2012). Cqpweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), 380–409.

Helmreich, W. B. (1996). *Against all odds: Holocaust survivors and the successful lives they made in America*. Transaction publishers.

Hladká, B., Hajic, J., Hana, J., Hlavácová, J., Mírovskỳ, J., & Raab, J. (2008). The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, *89*, 41.

Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication*, *10*(3), 185–205.

Honnibal, M., & Montani, I. (2015). *spaCy: Industrial-strength Natural Language Processing in Python.* Retrieved from `https://spacy.io`

Joseph, B. D. (2004). The editor's department: On change in Language and change in language. *Language*, 381–383.

Kaltenböck, G., Mihatsch, W., & Schneider, S. (2012). *New approaches to hedging* (Vol. 9). Brill.

Kohler, M. J. (1901). The german-jewish migration to america. *Publications of the American Jewish Historical Society*(9), 87–105.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics*, *1992*, 105–122.

Léon, J. (2005). Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society for the History of Linguistic Ideas Bulletin*, *44*(1), 36–50.

Mahlberg, M., Stockwell, P., Wiegand, V., & Lentin, J. (2020). *CLiC 2.1. corpus linguistics in context.* Retrieved from `clic.bham.ac.uk`

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., … others (2013). Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 92–97).

McEnery, T., & Hardie, A. (2013). The History of Corpus Linguistics. In *The Oxford Handbook of the History of Linguistics.* Oxford University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Moretti, F. (2000). Conjectures on world literature. *New left review*, *1*, 54.

Perks, R., & Robinson, J. (2005). The Way We Speak: Web-Based Representations of Changing Communities in England. *Oral History*, 79–90.

Perks, R., & Thomson, A. (2015). *The oral history reader*. Routledge.

Petrov, S., Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., & Brockman, W. (2012). Syntactic annotations for the Google books ngram corpus. In *Proceedings of the acl 2012 system demonstrations* (pp. 169–174).

Precht, K. (2003). Stance moods in spoken english: Evidentiality and affect in british and american conversation. *Text & talk*, *23*(2), 239–257.

Python Software Foundation. (2021). Python language reference [Computer software manual]. Retrieved from `https://docs.python.org/3/reference/`

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Association for computational linguistics (acl) system demonstrations*.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Roller, K. (2015). Towards the 'oral' in oral history: using historical narratives in linguistics. *Oral History*, 73–84.

Schiffrin, D. (2003). Linguistics and history: Oral history as discourse. *Linguistics, language, and the real world: Discourse and beyond*, *84*, 113.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International conference on new methods in language processing* (pp. 44–49). Manchester, UK. Retrieved from `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Schmid, H. (2019). Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In *Proceedings of the 2019 Conference on Digital Access to Textual Cultural Heritage (DATeCH)*. Brussels, Belgium. Retrieved from `https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/`

Sealey, A. (2010). Probabilities and surprises: A realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics*, *31*(2), 215–235.

Sealey, A. (2012). 'I just couldn't do it': representations of constraint in an oral history corpus. *Critical Discourse Studies*, *9*(3), 195–210.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Svartvik, J. (1990). *The london–lund corpus of spoken english: Description and research* (Vol. 82). Lund University Press.

Syncro Soft Ltd. (2023). *Oxygen xml editor*. Retrieved from `https://www.oxygenxml.com/`

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. *Treebanks: Building and using parsed corpora*, 5–22.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal*, *32*, 179–200.

Thomson, A. (2020). Using Primary Sources: Oral History. *Bloomsbury History: Theory and Method Articles*.

Witcomb, A. (2019). Oral history and first-person narratives in migration exhibitions: Tracking relations between "us" and "them". In *Remembering migration: Oral histories and heritage in australia* (pp. 203–217). Springer.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Lrec* (Vol. 2008, pp. 28–30).