LREC-COLING 2024

The First Workshop on Holocaust Testimonies as Language Resources HTRes@LREC-COLING-2024

Workshop Proceedings

Editors
Isuri Anuradha, Martin Wynne, Francesca Frontini, Alistair Plum

21 May, 2024
Torino, Italia

Proceedings of the First Workshop on Holocaust Testimonies as Language Resources: HTRes@LREC-COLING-2024

ISSN 978-2-493814-18-0
EAN 9782493814180

Acknowledgements: Organisation of the workshop is supported by European Holocaust Infrastructure (EHRI) and CLARIN ERIC.
https://www.clarin.eu/HTRes2024

Introduction

Holocaust testimonies serve as a bridge between people today and history's darkest chapters, providing a connection to profound experiences of survivors. Testimonies stand as the primary source of information that describe the Holocaust, offering first-hand accounts and personal narratives of those who experienced it. The majority of testimonies are captured in an oral format, as survivors vividly explain and share their personal experiences and observations from that time period.

The creation of accessible, comprehensive, and well-annotated Holocaust testimony collections is of paramount importance to our society. However, transforming Holocaust testimonies into a machine-processable digital format and publishing according to the standards and best practices of spoken and oral corpora can be a difficult task. At the same time these collections empower researchers and historians to validate the accuracy of socially and historically significant information, enabling them to share critical insights and trends derived from these data.

This workshop aimed to investigate how language technologies and, in particular, techniques and tools from natural language processing and corpus linguistics can greatly contribute to the exploration, analysis, dissemination and preservation of Holocaust testimonies.

The workshop was open to diverse communities and disciplines. The archivists who curate and digitize archives, producing digital editions that may also be enriched with various levels of annotation and structured knowledge; the NLP experts who process and in some case also make use of testimonies to develop specialised language models for speech and text processing; and finally historians who analyse the digital archives, via interfaces or using advanced text mining techniques, to discover new insights, or, for example, to retrace the biographical trajectories of witnesses.

The accepted papers touch upon all the aforementioned topics. Del Grosso et al and Beniere present two cases of publication of digital editions making use of the TEI standard, while Dermentzi et al show how the same manually curated editions can be used to train NLP models (in this case for NER). Flinz and Leonardi turn their attention in particular to the treatment of dates and places, while Liu and Mattingly (in a paper presented at the conference but not published in these proceedings) show how a typology of places can be extracted using NLP techniques. Draxler et al illustrate how the use of speech technologies can facilitate the curation of oral history materials, while Ifergan et al show how narratives can be analysed with the use of topic modelling. Vitali and Brazzo show how structured knowledge can be used to identify deportation trajectories, while, finally, Wagner et al show how these trajectories can also be extracted automatically in a structured format, and then visualised, interrogating a Large Language Model with a zero shot approach.

All these contributions illustrate well how techniques and best practices from the NLP and Language Resources and Technology domains are now mature enough to be relatively easily applied to Holocaust testimonies and readily applied to enlighten new aspects and dimensions of research. At the same time these testimonies can be language resources of great value, providing language technologists with a testbeds for domain adaptation of existing models and applications. The workshop was sparked by a collaboration between CLARIN, the Language Resources and Technology Iinfrastructure, and EHRI, the European Holocaust Research infrastructure. This is the third in an ongoing series of collaborative workshop, following on from earlier ones held at King's College London in 2023 (Making Holocaust Oral Testimonies More Usable as Research Data) and Charles University Prague in 2024 (Natural Language Processing Meets Holocaust Archives), and the first to take place attached to a conference devoted to NLP or language resources and tools.

Two keynote speakers were invited by the two infrastructures, Michal Frankl, coordinator of EHRI in the Czech Republic, highlighted in his talk the requirements of historians in terms of digital resources, while Silvia Calamai, member of the CLARIN-IT national consortium and one of the coordinators of the CLARIN funded Voices from Ravensbruck project showed how holocaust testimonies can be considered as important language resources also for language studies, in particular for their value as multilingual corpora documenting very interesting sociolinguistic and variety aspects.

Organising Committee

Martin Wynne, Oxford University, UK
Francesca Frontini,CNR-Istituto di Linguistica Computazionale "A. Zampolli", Italy & CLARIN ERIC
Isuri Anuradha, Lancaster University, UK
Alistair Plum, University of Luxembourg, Luxembourg
Paul Rayson, Lancaster University, UK
Ingo Frommholz, University of Wolverhampton, UK
Ruslan Mitkov, Lancaster University, UK

Programme Committee

Angelo Mario Del Grosso, CNR-Istituto di Linguistica Computazionale "A. Zampolli", Italy
Arjan van Hessen, Radboud University
Estelle Bunout, University of Luxembourg, Luxembourg
Eveline Wandl-Vogt, Austrian Academy of Sciences, Vienna
Federico Boschetti, CNR-Istituto di Linguistica Computazionale "A. Zampolli", Italy
Gabor Toth, University of Luxembourg, Luxembourg
Henk van den Heuvel, Radboud University & CLARIN ERIC
Ian Gregory, Lancaster University, UK
Ignatius Ezeani, Lancaster University, UK
Jan Svec, University of West Bohemia
Johannes-Dieter Steinert, University of Wolverhampton, UK
Le An Ha, University of Wolverhampton, UK
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, Poland
Maciej Piasecki, Wroclaw University of Science and Technology, Poland
Maria Dermentzi, King’s College London, UK
Martin Bulin, University of West Bohemia, Czech Republic
Patricia Murrieta-Flores, Lancaster University, UK
Rachel Pistol, King’s College London, UK
Stefanie Rauch, Wiener Holocaust Library, UK
Renana Keydar, The Hebrew University of Jerusalem, Israel
Tim Cole, University of Bristol, UK

Invited Speakers

Michal Frankl, Masaryk Institute and Archives of the Czech Academy of Sciences
Silvia Calamai, University of Siena, Italy

Table of Contents

Workshop Program

21 May 2024

9.30–9.45 **Welcome and Introduction**

9:45–10:30 **First Keynote Speech: Talking about Holocaust research (what type of digital resources they want? And why?) by Michal Frankl**

10:30–11:00 **Coffee Break**

11.00–12:00 **First session: Editions and their exploitation**

The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of Voci dall'Inferno Project
Angelo Mario Del Grosso, Marina Riccucci and Elvira Mercatanti

TEI Specifications for a Sustainable Management of Digitized Holocaust Testimonies
Sarah Bénière, Floriane Chiffoleau and Laurent Romary

Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools
Maria Dermentzi and Hugo Scheithauer

12.00–13.00 **Second paper session: Dates and Places**

Dates and places as points of attachment for memorial contents in the ISW corpus: 1938 as a turning point
Carolina Flinz and simona leonardi

Creating a Typology of Places to Annotate Holocaust Testimonies Through Machine Learning
Christine Liu and William J.B. Mattingly

21 May 2024 (continued)

13.00–14.00	**Lunch Break**
14.00–15.15	**Second Keynote Speech: Silvia Calamai, Università degli Studi di Siena, member of the CLARIN-IT consortium The Voices from Ravensbruck project: Bringing together what is dispersed**
15.15–16.00	**Third session: Testimonies and Narratives**
	Speech Technology Services for Oral History Research Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Pavel Ircing and Jan Lehečka
	Identifying Narrative Patterns and Outliers in Holocaust Testimonies Using Topic Modeling Maxim Ifergan, Omri Abend, Renana Keydar and Amit Pinchevski
16.00–16.30	**Coffee Break**
16.30–17.10	**Fourth session: Traces and networks**
	Tracing the deportation to define Holocaust geometries. The exploratory case of Milan Giovanni Pietro Vitali and Laura Brazzo
	Zero-shot Trajectory Mapping in Holocaust Testimonies Eitan Wagner, Renana Keydar and Omri Abend

21 May 2024 (continued)

17.10–17.50	**Pannel Discussion**
17.50–18.00	**Final Remark**

The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of Voci dall'Inferno Project

Angelo Mario Del Grosso, Marina Riccucci, Elvira Mercatanti
CNR-ILC, UNIPI,
Via Moruzzi, Pisa; Piazza Dante, Pisa
angelomario.delgrosso@cnr.it, marina.riccucci@unipi.it, e.mercatanti@studenti.unipi.it

Abstract

In Nazi concentration camps, approximately 20 million people perished. This included young and old, men and women, Jews, dissidents, and homosexuals. Only 10% of those deported survived. This paper introduces "Voci dall'Inferno" project, which aims to achieve two key objectives: a) Create a comprehensive digital archive: by encoding a corpus of non-literary testimonies including both written and oral sources. b) Analyze the use of Dante's language: by identifying the presence of Dante's lexicon and allusions. Currently, the project holds 47 testimonies, with 29 transcribed in full text and 18 encoded using the XML-TEI format. This project is propelled by a multidisciplinary and educational context with experts in humanities and computer science. The project's findings will be disseminated through a user-friendly web application built on an XML foundation. Though currently in its prototyping phase, the application boasts several features, including a search engine for testimonies, terms, or phrases within the corpus. Additionally, a browsing interface allows users to read and listen the original testimonies, while a visualization tool enables deeper exploration of the corpus's content. Adhering to the Text Encoding Initiative (TEI) guidelines, the project ensures a structured digital archive, aligned with the FAIR principles for data accessibility and reusability.

Keywords: XML-TEI, Holocaust Testimonies, Digital Archives

1. Introduction

The concentration camps we know are the camps of 1944 because almost none of the *Häftlinge*, namely the prisoners, survived to tell us about the other camps[1]. In the last years, many of the survivors have died. In addition, not all testimonies have told their story, but only a part, certainly not the majority. Many survivors have remained silent. Moreover, the *Camp* is *ineffable* (Levi, 2018, p. 688). Thus, for the testimonies, the problem was also to find the words to express the atrocity they suffered (Wiesel, 1995). To tell the story of the Camp, therefore, one must overcome the poverty of a vocabulary that does not have "*the words to say it*".

Saying that the *Lager* was *l'Inferno - the hell*[2] (Calderini and Riccucci, 2020) has allowed survivors to establish an immediate contact with their audience. This shared metaphor always recurs in holocaust testimonies.

Since 2016, professorship of Italian Literature at the University of Pisa has directed and coordinated, with the support of the CNR-ILC of Pisa, the research project called "Voci dall'Inferno". This scholarly initiative has two integrated and correlated objectives:

a) The digitization and encoding of a corpus, the largest possible, of testimonies.

b) The identification, quantification and evaluation of the presence of Dante's lexicon and imagery within those testimonies.

Early outcomes have been surprisingly important: given the vast scope of the project, the very high number of testimonies, most of which are unpublished, it has been necessary to involve computational methods and techniques[3] as well as to make the research activities a collaborative work, which has seen and continues to see the collaboration and contribution of many DH students.

2. Testimonies and Dante's Lexicon

The textual typologies through which the Lager has been reported to us *in words* are essentially two:

- Direct testimony - coeval and non-coeval - of those who experienced the extermination

[1] Encyclopedia of the Holocaust: https://www.ushmm.org/it

[2] It is worth noting that the Inferno here refers to Dante's literary work *Divina Commedia* (Alighieri, 2002), not the Christian concept of Hell

[3] Currently, the corpus is manually annotated and digital encoded. This initial, time-consuming effort will serve as a valuable ground truth for training and refining sophisticated machine learning and prediction tools for automated data extraction and encoding. Examples include leveraging current AI applications for tasks such as handwritten text recognition, automatic speech recognition, named entity recognition, topic extraction and modeling, semantic text alignment, and more. Once the final topics are defined, drawing upon the work done on similar collections, we also plan to incorporate formal ontologies and linked open data.

camp and reported it in forms that only rarely touch on literariness: the *modus dicendi* of this type of testimony is located in the space between the oral report (the interview) and the written one (the diary, the autobiographical/memorial story, the letter).

- Indirect testimony - coeval and non-coeval - of those who experienced the extermination camp and chose, to report it the form of narrative, therefore of literature. This kind of testimony is presented in the form of an organized story, thematically and stylistically structured.

The latter category encompasses works with manifested literary ambitions that would never have been written if the Holocaust and the deportation had not occurred. We are referring to that literary production like *Se questo è un uomo - If this is a man* (Levi, 1947), *Night* (Wiesel, 1960), *This Way for the Gas, Ladies and Gentlemen* (Borowski, 1976), *The Human Race* (Antelme et al., 1998), *L'Univers concentrationnaire* (Rousset, 1946), *Fatelessness* (Kertész, 1975). In the context of our research project we call this kind of testimonies "second level" ones.

We are faced with two distinct and differentiated forms of representation/restitution of the concentrationary universe. What has been progressively verified is not only that the lexicon of the Inferno of Dante breaks the silence and intervenes in those who are going to narrate the tragedy of their experience (Arquès, 2009; Pertile, 2010; Susteric, 2016; Taterka, 2002) providing the words "to say it", but also that the expressive faculties of all witnesses are influenced by the Dante's Inferno, even of those who have read Dante only on the school benches, or, even, of those who have received Dante as a heritage of oral culture.

By tacit agreement, all the *Häftlinge*, without exception, who have testified have chosen and adopted the term "*Inferno*" to convey their experience of the concentration camp to those who have never experienced it themselves.

There are at least two basic assumptions from which we cannot prescind:

1. The metaphorical nexus (Lager-Inferno) is a new nexus. It did not exist before, simply because before the 30s of the twentieth century, the Lager did not exist.

2. When the testimonies talk about the Lager as of the Inferno they do not refer to any inferno, or to an inferno and nothing else. They do so with the Dante's inferno in mind.

The data collected so far tell us that Dante breaks the silence, in the sense that survivors use Dante's lexicon to untie the knot of ineffability, to dilute the paralysis of mind and memory in the face of the emergence of the nefarious experience. This happens in all survivors (Riccucci and Riccotti, 2021).

Listening to and reading the testimonies of the Nazi camps, one realizes that to report on the concentrationary hell, the survivors of the extermination, people of every level of education, relied on the words of the Dante's vocabulary, that of the first canticle, for the most part, that of the Inferno. Of course, the testimonies are not all and not systematically or capillarily woven with Dante's verses, but these verses at a certain point burst from the lips of these men and women to express the inexpressible.

The testimonies can count on a lexical heritage and a collective imagery made up of Dante's words that have entered into common use, penetrated into the language of everyday life, into the speech of the family and of society, transmitted from generation to generation as an inheritance.

In the light of the above, computer science allows us to preserve and archive, to interrogate and find connections, to build maps, to intertwine stories[4].

3. Digital Archive and Corpus

The Voci dall'Inferno project has undergone three distinct phases, each contributing to its current state as a web platform for studying Holocaust testimonies[5]:

1. Development of a database for the management of the testimony records - the archive has been named *memoria-archivio* (Riccucci et al., 2021).

2. Creation of the corpus of the testimonies in XML-TEI format (Burnard, 2014).

3. Development of a web application for the presentation and interrogation of the data stored in the digital archive.

[4] Similar initiatives encompass projects like *Let Them Speak*, https://lts.fortunoff.library.yale.edu/, or *David Boder: from wire recordings*, https://ranke2.uni.lu/u/boder/, or archives like *CDEC*, https://digital-library.cdec.it/cdec-web/, as well as initiatives like *EHRI*, https://www.ehri-project.eu/. Other projects of a similar nature can be found at https://dhjewish.org/projects.

[5] The corpus - now encompassing about 500,000 tokens in Italian - is not yet exhaustive of all available primary sources. We plan to incorporate new oral and written testimonies during our research. The initial corpus is conceived as the first iteration of the project, designed to primarily establish the data model, the encoding schema, the workflow, and the scholarly framework. During this initial phase of data collection, resources were gathered from private and personal archives, public institutional archives, audio type interviews, and video interviews

The *memoria-archivio* database provides us with a web environment for the creation of the initial inventory of the testimonies. It preserves catalogographic and literary descriptions as well as, where present, also manages the transcriptions of the textual content of the testimony. In addition, the web environment allows for the comparison of the testimony lexicon with the text of the Dante's Comedy.

The application allows for the updating of the inventory, the records of the witnesses and the curators of the sources. Subsequently, memoria-archivio also integrated the management of documents in XML-TEI format. To date, the archive has 47 testimonies, of which 29 are full-text transcriptions and 18 documents has been also encoded in XML-TEI format (Fig. 1). These collection of testimonies includes known and lesser-known names of Italians who were deported, such as Samuele Modiano, Piero Terracina, Enrico Vanzini, Liliana Segre, Nedo Fiano, Shlomo Venezia, Primo Levi, Ida Marcheria, Goti Bauer, and many others.

Within the digital archive, the testimonies are divided into two macro-classes, which determine their representational, functional and exploitation aspects. On the one hand, there are the oral testimonies (Fig. 6) and on the other hand, the written testimonies (Fig. 2). While maintaining the specific differences, both classes follow the guidelines provided by the TEI consortium.

In particular, during the course of the project, a "One Document Does it All" (ODD)[6] was created - and gradually refined, which declares the modules, elements, attributes and possible values allowed for the encoding of the corpus. As for the encoding model, the written testimonies follow an image-based editing scheme of a diplomatic-interpretative type with a parallel-transcription approach to the representation of the text-document (Pierazzo, 2015). These kinds of digital documents can be published by means of the EVT Web application (Rosselli Del Turco et al., 2019).

To this end, the elements defined in module 11 (transcription) were used for the transcription of the primary source; module 13 (namesdates) for the representation of named entities; module 16 (linking) for the analysis of particular text structures; module 17 (analysis) for the semantic and linguistic-lexical annotation of textual units. As for the description of the primary source, the elements defined in module 10 (msDesc) of the TEI guidelines were adopted.

3.1. Metadating the Testimonies

Oral testimonies differ structurally from written ones due to the inherent characteristics of spoken language. Notably, oral testimonies are characterized by the temporal dimension of speech and the unique order in which utterances unfold.

This leads to the creation of specific "timeline" elements designed to synchronize the topics covered by the witness. This synchronization aims to align the timing of the speaker's utterances with their corresponding transcription.

```
<abstract><ab><list>
 <item synch="#TR1">
  <persName ref="#LS">
   <forename>Liliana</forename>
   <surname>Segre</surname>
  </persName>, rispondendo alla domanda
   postale da <persName ref="#AS">
   <forename>Anna</forename>
   <surname>Segre</surname>
  </persName>, parla di che cosa abbia
 </item>
</abstract>
```

```
<standOff>
 <timeline xml:id="TL1I"
 source="#reg_1B" unit="s">
 <!-- ... -->
 <when xml:id="TR1"
  absolute="00:00:00"/>
 <!-- ... -->
 <when xml:id="TR7"
  absolute="00:23:41"/>
 <!-- ... -->
 </timeline><!-- ... -->
</standOff>
```

In addition to the timeline, a section of reasoned summary, called *regesto*, has been introduced, which briefly illustrates the content of each division. Four different timelines have been defined in the encoding model.

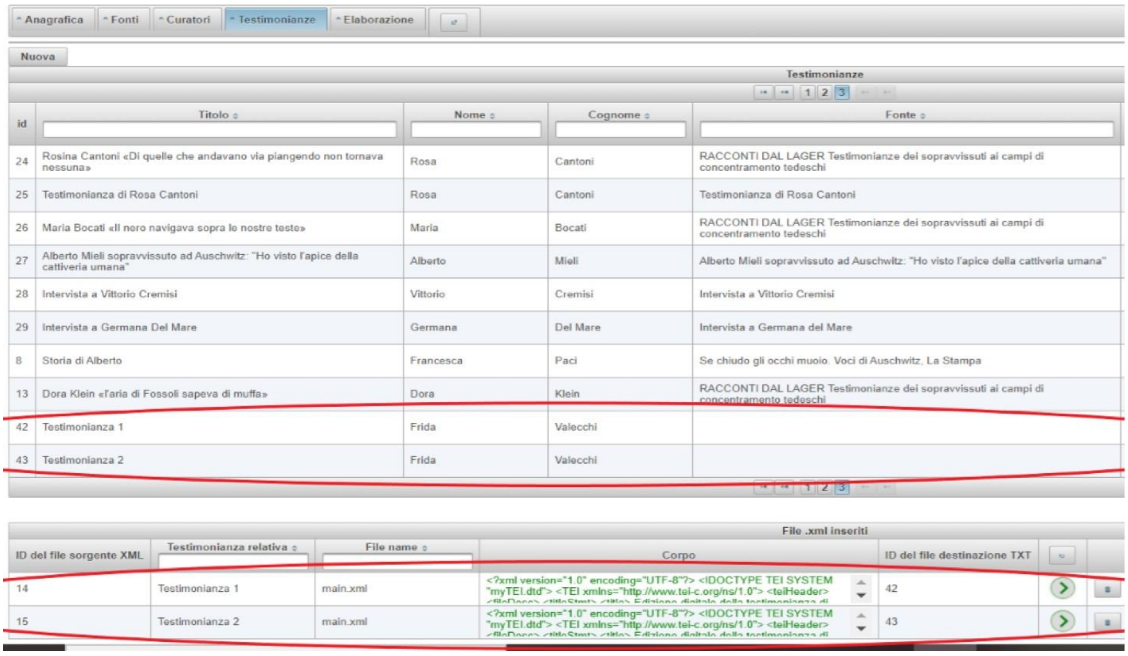
In the first timeline, the <when/> elements identify the moments when the various topics of the testimony are introduced, i.e. those summarized within the <item> elements present in the *regesto*, which, in turn, have a @synch attribute in order to connect the timing specified by the relative <when/> tag.

In the second timeline, on the other hand, all the segments in which the voices overlap have been collected. For this reason, the <when/> elements inside it are in pairs: one of them identifies the interval of the overlap.

The third timeline was instead created to group together the moments in which a change of speaker occurs. The fourth and last timeline allows to record the moments in which background noises overlap with the utterances.

The most significant XML-TEI elements used for the description of oral sources can be summarized as follows: the information relating to the medium

[6] See also at https://tei-c.org/guidelines/customization/



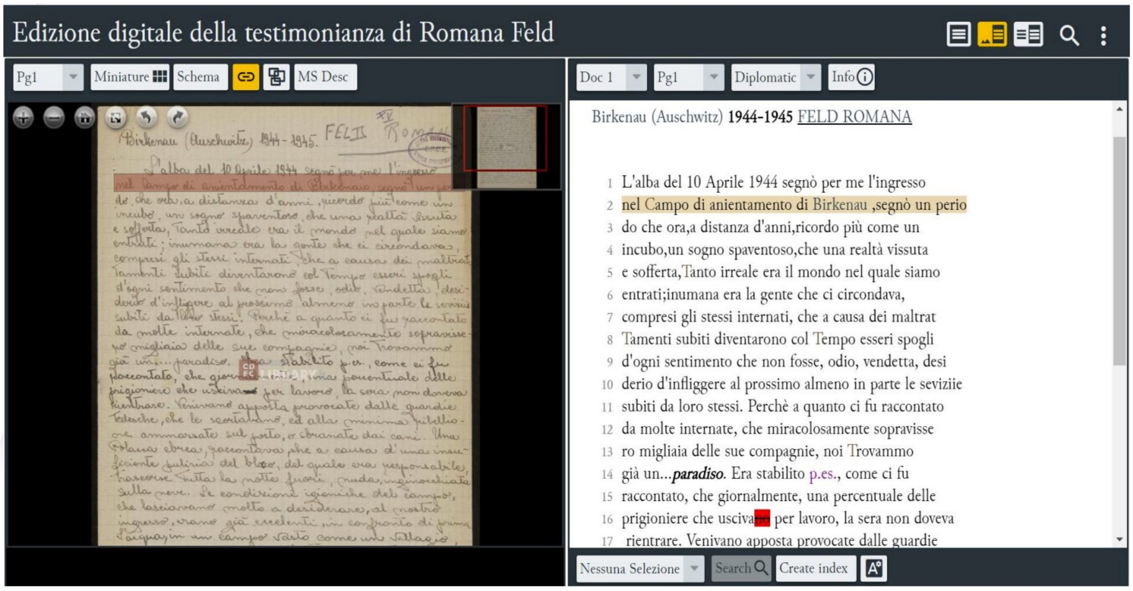
Testimonianze

id	Titolo	Nome	Cognome	Fonte
24	Rosina Cantoni «Di quelle che andavano via piangendo non tornava nessuna»	Rosa	Cantoni	RACCONTI DAL LAGER Testimonianze dei sopravvissuti ai campi di concentramento tedeschi
25	Testimonianza di Rosa Cantoni	Rosa	Cantoni	Testimonianza di Rosa Cantoni
26	Maria Bocati «Il nero navigava sopra le nostre teste»	Maria	Bocati	RACCONTI DAL LAGER Testimonianze dei sopravvissuti ai campi di concentramento tedeschi
27	Alberto Mieli sopravvissuto ad Auschwitz: "Ho visto l'apice della cattiveria umana"	Alberto	Mieli	Alberto Mieli sopravvissuto ad Auschwitz: "Ho visto l'apice della cattiveria umana"
28	Intervista a Vittorio Cremisi	Vittorio	Cremisi	Intervista a Vittorio Cremisi
29	Intervista a Germana Del Mare	Germana	Del Mare	Intervista a Germana del Mare
8	Storia di Alberto	Francesca	Paci	Se chiudo gli occhi muoio. Voci di Auschwitz, La Stampa
13	Dora Klein «l'aria di Fossoli sapeva di muffa»	Dora	Klein	RACCONTI DAL LAGER Testimonianze dei sopravvissuti ai campi di concentramento tedeschi
42	Testimonianza 1	Frida	Valecchi	
43	Testimonianza 2	Frida	Valecchi	

File .xml inseriti

ID del file sorgente XML	Testimonianza relativa	File name	Corpo	ID del file destinazione TXT
14	Testimonianza 1	main.xml	<?xml version="1.0" encoding="UTF-8"?> <!DOCTYPE TEI SYSTEM "myTEI.dtd"> <TEI xmlns="http://www.tei-c.org/ns/1.0"> <teiHeader> <fileDesc> <titleStmt> <title> Edizione digitale della testimonianza di	42
15	Testimonianza 2	main.xml	<?xml version="1.0" encoding="UTF-8"?> <!DOCTYPE TEI SYSTEM "myTEI.dtd"> <TEI xmlns="http://www.tei-c.org/ns/1.0"> <teiHeader> <fileDesc> <titleStmt> <title> Edizione digitale della testimonianza di	43

Figure 1: Memoria-Archivio Web Application



Birkenau (Auschwitz) **1944-1945** FELD ROMANA

L'alba del 10 Aprile 1944 segnò per me l'ingresso
nel Campo di anientamento di Birkenau ,segnò un perio
do che ora,a distanza d'anni,ricordo più come un
incubo,un sogno spaventoso,che una realtà vissuta
e sofferta,Tanto irreale era il mondo nel quale siamo
entrati;inumana era la gente che ci circondava,
compresi gli stessi internati, che a causa dei maltrat
Tamenti subiti diventarono col Tempo esseri spogli
d'ogni sentimento che non fosse, odio, vendetta, desi
derio d'infliggere al prossimo almeno in parte le seviziie
subiti da loro stessi. Perchè a quanto ci fu raccontato
da molte internate, che miracolosamente sopravisse
ro migliaia delle sue compagnie, noi Trovammo
già un...***paradiso***. Era stabilito p.es., come ci fu
raccontato, che giornalmente, una percentuale delle
prigioniere che uscivano per lavoro, la sera non doveva
rientrare. Venivano apposta provocate dalle guardie

Figure 2: EVT Web Application for Autographs' Primary Sources

was recorded using the <recordingStmt> element, contained in turn in the <sourceDesc> element, belonging to module 8 of the TEI guidelines (spoken). The <recording> element finally represents a single recording and contains all the information necessary to specify the context and responsibilities of the recording. Each <recording> element is connoted with the type attribute (@type) (audio or video) and a duration (@dur) for each single recording.

```
<sourceDesc>
 <recordingStmt>
<recording type="audio"
 dur="P30M41S" xml:id="reg_1B">
 <p>Registrazione 1 lato B</p>
 <date cert="low"
   when="2007-12-08">
   8 dicembre 2007</date>
</recording>
<recording type="audio"
 dur="P30M41S" xml:id="reg_2A">
 <p>Registrazione 2 lato A</p>
 <date cert="low"
   when="2007-12-08">
```

```
     8 dicembre 2007</date>
  </recording>
  <recording type="audio"
   dur="P24M32S" xml:id="reg_2B">
   <p>Registrazione 2 lato B</p>
   <date cert="low"
     when="2007-12-08">
     8 dicembre 2007</date>
  </recording>
 </recordingStmt>
</sourceDesc>
```

Finally, within the <profileDesc> block, the elements defined in module 15 (corpus) were used, in particular the <particDesc> element offers an accurate description of the people who took part in the conversation.

3.2. Transcribing the Testimonies

The two encoding models introduced in previous sections, namely the model for oral testimonies and the model for written testimonies, differ from each other both in terms of descriptive and analytical choices.

The logical structure of the written testimony often follows a predominantly epistolary grammar, but it can differ substantially due to authorial and editorial characteristics (authorial interventions on manuscripts or typewritten texts).

The primary source is represented using the TEI facsimile tagset to describe areas of interest through relevant attributes within the @zone element. These zones are then referenced by corresponding elements within the transcription section, adhering to the best practices outlined by the parallel-transcription method (TEI module 11)[7].

Within the transcription section (<text> block), various elements are used to represent the content of the primary source. These include:

- <subst> or <mod>: These elements mark the original text with authorial interventions.
- <choice> elements with their sub-elements: These elements record the original reading and any editorial interpretations made by scholars.
- <damage>, <unclear> and <supplied>: This element allows scholars to include missing or unclear information, providing any difficulties encountered due to material damage.

Named entities encoding follows the best practices suggested by the TEI guidelines, adopting the elements <person>, <org>, <place>, <event> and the respective <personName>, <orgName>, <placeName>. Dante's quotes and terminology have been annotated with the <cit> and <term> elements respectively.

As for oral testimonies, the transcription is divided into textual units called utterances using the <u> element. Each utterance is accompanied by the @who attribute which allows to associate the person who formulated it. In addition, the @xml:id and @synch elements are functional for a correct synchronization. The @trans attribute, on the other hand, specifies whether the participants' utterances follow one another or overlap.

During the testimony, numerous phenomena are recorded, such as pauses (element <pause> with attribute @type to indicate the length), non-lexical sounds (element <vocal>), prosodic events (element <kinesic>), background noises (element <incident>), inaudible or uncertain passages (elements <gap> and <unclear>), changes in paralinguistic features such as intonation, volume, rhythm, speed by means of the <shift /> element accompanied, as is appropriate, by the attributes @feature and @new.

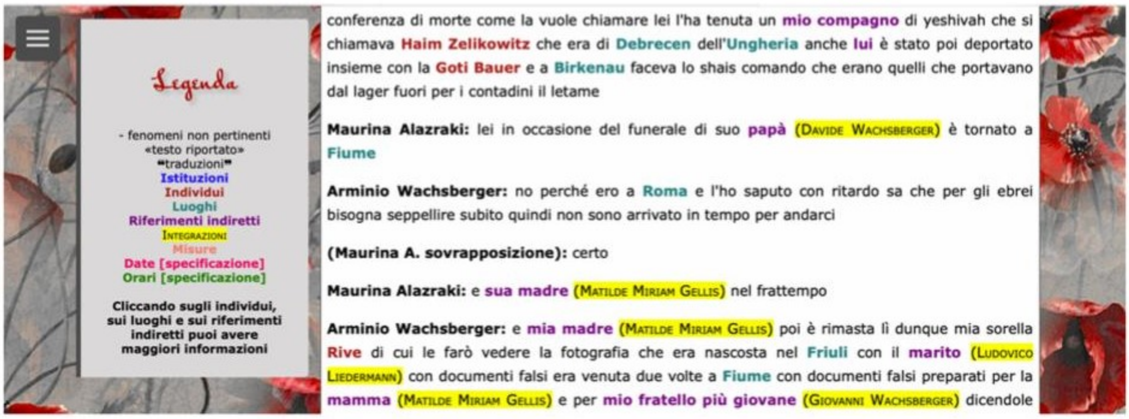
```
<u><!-- ... --!>
non hai un nome, perché
hai un numero,
<pause type="long"/>
ti chiamano per numero
<pause type="medium"/>
e quindi <pause type="short"/>
cercano di
degradarti <pause type="short"/>
con la fame <pause type="short"/>
<!-- ... ---!> </u>
```

```
<u who="#MARCHERIA" xml:id="m223"
  synch="#tlp457"> In questo
 <supplied>caos</supplied>
 sì, perché arrivavano i russi
 <pause type="short"/>
 e c'era il caos. Siamo
 <del type="repetition">siamo</del>
 <kinesic>
 <desc>
 Ida mostra la grandezza della piazza
 </desc>
</kinesic>
```

4. Voci Dall'Inferno Web Applet

The digital archive of testimonies would be less effective from a functional and scientific point of view without the presence of a software component dedicated to the extraction, manipulation, presentation and use of the data collected during the encoding phase.

[7] See also at https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PH-bov



Legenda

- fenomeni non pertinenti
«testo riportato»
"traduzioni"
Istituzioni
Individui
Luoghi
Riferimenti indiretti
INTEGRAZIONI
Misure
Date [specificazione]
Orari [specificazione]

Cliccando sugli individui, sui luoghi e sui riferimenti indiretti puoi avere maggiori informazioni

conferenza di morte come la vuole chiamare lei l'ha tenuta un **mio compagno** di yeshivah che si chiamava **Haim Zelikowitz** che era di **Debrecen** dell'**Ungheria** anche **lui** è stato poi deportato insieme con la **Goti Bauer** e a **Birkenau** faceva lo shais comando che erano quelli che portavano dal lager fuori per i contadini il letame

Maurina Alazraki: lei in occasione del funerale di suo **papà** (DAVIDE WACHSBERGER) è tornato a **Fiume**

Arminio Wachsberger: no perché ero a **Roma** e l'ho saputo con ritardo sa che per gli ebrei bisogna seppellire subito quindi non sono arrivato in tempo per andarci

(Maurina A. sovrapposizione): certo

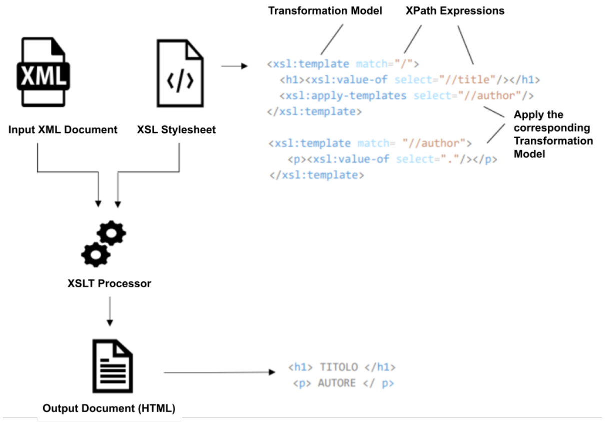
Maurina Alazraki: e **sua madre** (MATILDE MIRIAM GELLIS) nel frattempo

Arminio Wachsberger: e **mia madre** (MATILDE MIRIAM GELLIS) poi è rimasta lì dunque mia sorella **Rive** di cui le farò vedere la fotografia che era nascosta nel **Friuli** con il **marito** (LUDOVICO LIEDERMANN) con documenti falsi era venuta due volte a **Fiume** con documenti falsi preparati per la **mamma** (MATILDE MIRIAM GELLIS) e per **mio fratello più giovane** (GIOVANNI WACHSBERGER) dicendole

Figure 3: Voci dall'Inferno Web Application developed with Saxonjs2

Two different data restitution strategies were experimented, which make use of two different architectural approaches.

For the first approach, web applications were developed by leveraging the functionalities of a client-side library for processing XML documents, namely SaxonJS2[8] (Fig. 4).



```
<xsl:template match="/">
  <h1><xsl:value-of select="//title"/></h1>
  <xsl:apply-templates select="//author"/>
</xsl:template>

<xsl:template match= "//author">
   <p><xsl:value-of select="."/></p>
</xsl:template>
```

```
<h1> TITOLO </h1>
 <p> AUTORE </ p>
```

Figure 4: Saxon Processing Model

In relation to the second approach, the web applications were developed by means of the eXist-db[9] environment using the HTML templating module (server-side)[10].

Thanks to the use of the SaxonJS2 library, it is possible to integrate an efficient XSLT processor by delegating the browser's javascript engine to manipulate the DOM object of the HTML page.

The library exposes an effective API whose main methods are `SaxonJS.transform (options)` to execute the procedures defined by the transformation rules and `SaxonJS.XPath.evaluate(XPath)` to select appropriate sequences of XML nodes or process them according to the specifications of the XPath 3.1 standard.

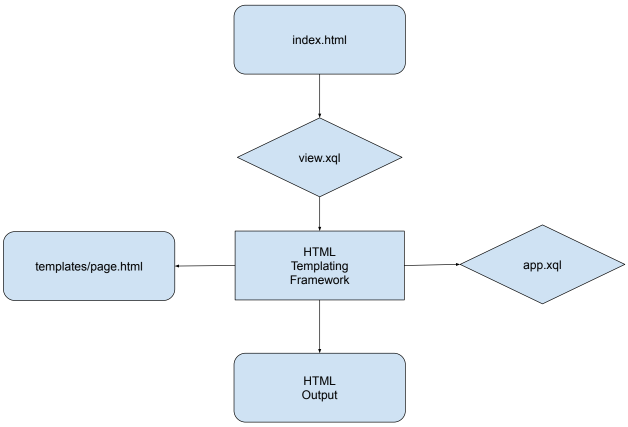


Figure 5: eXist-db Processing Model with HTML Templating

```
SaxonJS.getResource({
  location: "testimony.xml",
  type: "xml" })
  .then(doc => {
    const result =
     SaxonJS.XPath
     .evaluate(
      "//persName/text()", doc);
    const output =
     SaxonJS.serialize(result, {
    method: "xml",
    indent: true,
   "omit-xml-declaration":true});})
```

The image in Figure 3 shows a web page generated with the help of the SaxonJS2 library for

[8] See also at https://www.saxonica.com/saxon-js/index.xml

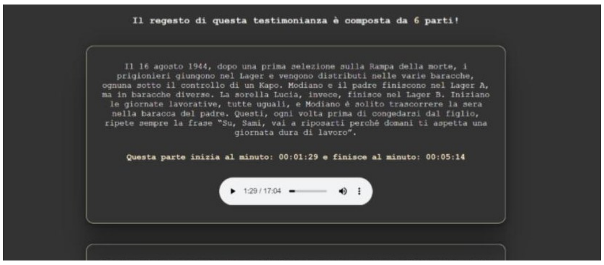
[9] See also at https://exist-db.org/exist/apps/homepage/index.html

[10] See also at https://github.com/eXist-db/templating

the visualization of the testimony of Arminio Wachsberger. It is possible to notice the transcribed text of the interview, the participants, the textual phenomena annotated and rendered graphically according to the styles indicated in the legend.

The second approach is based on the eXist-db technology. As introduced, the platform integrates a module dedicated to the dynamic generation of HTML pages starting from collections of documents in XML format and from procedures implemented using the XQuery instructions. The basic operation involves the use of HTML templates, in which appropriate directives and calls to XQuery functions are added. The functions implement the application logic to generate the HTML fragments useful to complete the actual HTML page.

A relevant feature of the eXist-db technology is the possibility to use the Apache Lucene library for indexing textual data and for the efficient interrogation of them. Figure 7 shows an example of querying and retrieving data related to the testimony of Ida Marcheria (partial word search "tren").



Il regesto di questa testimonianza è composta da 6 parti!

Il 16 agosto 1944, dopo una prima selezione sulla Rampa della morte, i prigionieri giungono nel Lager e vengono distributi nelle varie baracche, ognuna sotto il controllo di un Kapo. Modiano e il padre finiscono nel Lager A, ma in baracche diverse. La sorella Lucia, invece, finisce nel Lager B. Iniziano le giornate lavorative, tutte uguali, e Modiano è solito trascorrere la sera nella baracca del padre. Questi, ogni volta prima di congedarsi dal figlio, ripete sempre la frase "Su, Sami, vai a riposarti perché domani ti aspetta una giornata dura di lavoro".

Questa parte inizia al minuto: 00:01:29 e finisce al minuto: 00:05:14

Figure 6: Voci dall'Inferno Web Application for Audio and Regesto Features

The web application developed so far for Voci dall'Inferno (Fig. 8) has multiple functionalities implemented or under development such as: I) Catalog management and search (Fig. 1); II) Presentation and use of data in parallel with the primary source (Fig. 2); III) Search within the textual archive (Fig. 7); IV) Management of the regesto (Fig. 6); V) Management of speech conventions (Konrad, 2003) (Fig. 8); VI) Statistics of phenomena (Fig. 3); VII) Terminology management (Fig. 7); VIII) Management of quotations and allusions (Dante's ones in particular).

5. Conclusion

We have presented the *Voci dall'Inferno* project, a scholarly initiative, within an educational framework, building a digital corpus of Holocaust testimonies and a dedicated digital environment for searching and analyzing them. This project uniquely explores the presence of Dante's vocabulary and allusions within the testimonies. The digital collection adheres to the *Text Encoding Initiative* (TEI) schema, maximizing data accessibility, searchability, interoperability, and reusability in accordance with the *FAIR principles* (Wilkinson and al., 2016). Notably, the project is developing functionalities for data classification and extraction using machine learning techniques. These functionalities will enable automatic speech recognition and transcription, automatic search for literary tesserae, as well as automatic network analysis, further enriching the exploration and understanding of these testimonies.

6. Bibliographical References

Dante Alighieri. 2002. *La divina commedia. Inferno*. Armando Editore.

Robert Antelme, Jeffrey Haight, and Annie Mahler. 1998. *The Human Race*. Marlboro Press/Northwestern.

Rossend Arquès. 2009. Dante nell'inferno moderno:la letteratura dopo auschwitz. *Rassegna Europea di Letteratura Italiana*, 33.

Tadeusz Borowski. 1976. *This way for the gas, ladies and gentlemen*, 1st ed. edition. Penguin Books, New York.

Tadeusz Borowski. 2009. *Da questa parte, per il gas*. Un mondo a parte. L'Ancora del Mediterraneo.

Lou Burnard. 2014. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. OpenEdition Press, Marseille.

Sara Calderini and Marina Riccucci. 2020. Le parole per dire il lager. *Italianistica*, XLIX.

Imre Kertész. 1975. *Fatelessness*. Vintage.

Ehlich Konrad. 2003. Hiat: A transcription system for discourse data. *Talking Data: Transcription and Coding in Discourse Research*.

Primo Levi. 1947. *Se questo è un uomo*. De Silva, Torino.

Primo Levi. 2018. *Opere Complete*. Giulio Einaudi editore.

Lino Pertile. 2010. L'inferno, il lager, la poesia. *Dante*, 7.

Elena Pierazzo. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate, Surrey.

Marina Riccucci, Angelo Mario Del Grosso, Frida Valecchi, and Giulia Causarano. 2021. Testimoniare il lager: L'informatica al servizio della memoria. *Quaderni Di Umanistica Digitale*.

Marina Riccucci and Laura Riccotti. 2021. *Il dovere della parola. Le testimonianze di Liliana Segre e di Goti Herskovits Bauer*, volume XLIX. Pacini editore, Pisa.

Roberto Rosselli Del Turco, Chiara Martignano, Chiara Di Pietro, Giulia Cacioli, Angelo Mario Del Grosso, and Simone Zenzaro. 2019. DSE Visualisation with EVT: Simplicity is Complex. In *Complexities*.

David Rousset. 1946. *L'Univers concentrationnaire*. éditions du Pavois.

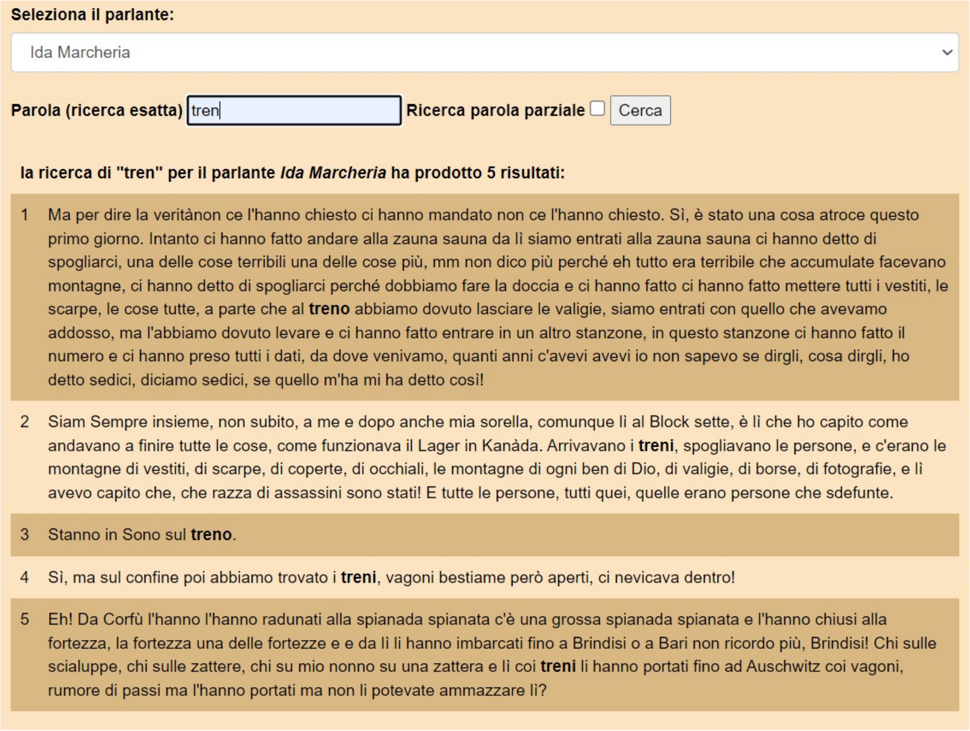
Federica Susteric. 2016. La dicibilità del male. la ricezione dantesca nelle testimonianze concentrazionarie. *Dante*, XII.

Thomas Taterka. 2002. Dante deutsch. *Studi sulla letteratura dei Lager*.

Elie Wiesel. 1960. *Night*. Hill 'n' Wang, New York.

Elie Wiesel. 1995. *La notte*. Giuntina editore.

Mark Wilkinson and al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.



la ricerca di "tren" per il parlante *Ida Marcheria* ha prodotto 5 risultati:

1 Ma per dire la veritànon ce l'hanno chiesto ci hanno mandato non ce l'hanno chiesto. Sì, è stato una cosa atroce questo primo giorno. Intanto ci hanno fatto andare alla zauna sauna da lì siamo entrati alla zauna sauna ci hanno detto di spogliarci, una delle cose terribili una delle cose più, mm non dico più perché eh tutto era terribile che accumulate facevano montagne, ci hanno detto di spogliarci perché dobbiamo fare la doccia e ci hanno fatto ci hanno fatto mettere tutti i vestiti, le scarpe, le cose tutte, a parte che al **treno** abbiamo dovuto lasciare le valigie, siamo entrati con quello che avevamo addosso, ma l'abbiamo dovuto levare e ci hanno fatto entrare in un altro stanzone, in questo stanzone ci hanno fatto il numero e ci hanno preso tutti i dati, da dove venivamo, quanti anni c'avevi avevi io non sapevo se dirgli, cosa dirgli, ho detto sedici, diciamo sedici, se quello m'ha mi ha detto così!

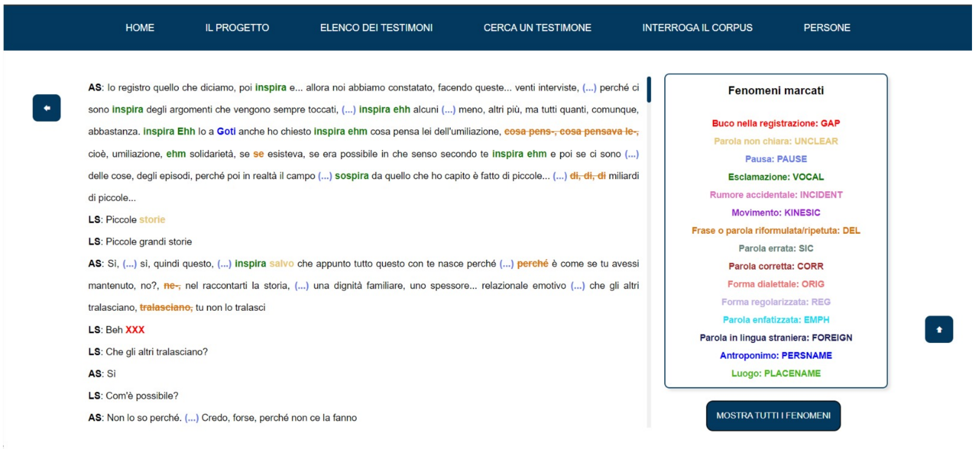
2 Siam Sempre insieme, non subito, a me e dopo anche mia sorella, comunque lì al Block sette, è lì che ho capito come andavano a finire tutte le cose, come funzionava il Lager in Kanàda. Arrivavano i **treni**, spogliavano le persone, e c'erano le montagne di vestiti, di scarpe, di coperte, di occhiali, le montagne di ogni ben di Dio, di valigie, di borse, di fotografie, e lì avevo capito che, che razza di assassini sono stati! E tutte le persone, tutti quei, quelle erano persone che sdefunte.

3 Stanno in Sono sul **treno**.

4 Sì, ma sul confine poi abbiamo trovato i **treni**, vagoni bestiame però aperti, ci nevicava dentro!

5 Eh! Da Corfù l'hanno l'hanno radunati alla spianada spianata c'è una grossa spianada spianata e l'hanno chiusi alla fortezza, la fortezza una delle fortezze e e da lì li hanno imbarcati fino a Brindisi o a Bari non ricordo più, Brindisi! Chi sulle scialuppe, chi sulle zattere, chi su mio nonno su una zattera e lì coi **treni** li hanno portati fino ad Auschwitz coi vagoni, rumore di passi ma l'hanno portati ma non li potevate ammazzare lì?

Figure 7: Voci dall'Inferno Web Application - Search Capabilities



AS: lo registro quello che diciamo, poi **inspira** e... allora noi abbiamo constatato, facendo queste... venti interviste, (...) perché ci sono **inspira** degli argomenti che vengono sempre toccati, (...) **inspira ehh** alcuni (...) meno, altri più, ma tutti quanti, comunque, abbastanza. **inspira Ehh** lo a **Goti** anche ho chiesto **inspira ehm** cosa pensa lei dell'umiliazione, ~~cosa pens-, cosa pensava le-,~~ cioè, umiliazione, **ehm** solidarietà, se **se** esisteva, se era possibile in che senso secondo te **inspira ehm** e poi se ci sono (...) delle cose, degli episodi, perché poi in realtà il campo (...) **sospira** da quello che ho capito è fatto di piccole... (...) ~~di, di, di~~ miliardi di piccole...

LS: Piccole storie

LS: Piccole grandi storie

AS: Sì, (...) sì, quindi questo, (...) **inspira** salvo che appunto tutto questo con te nasce perché (...) **perché** è come se tu avessi mantenuto, no?, ~~ne-,~~ nel raccontarti la storia, (...) una dignità familiare, uno spessore... relazionale emotivo (...) che gli altri tralasciano, ~~tralasciano,~~ tu non lo tralasci

LS: Beh **XXX**

LS: Che gli altri tralasciano?

AS: Sì

LS: Com'è possibile?

AS: Non lo so perché. (...) Credo, forse, perché non ce la fanno

Fenomeni marcati

Buco nella registrazione: GAP
Parola non chiara: UNCLEAR
Pausa: PAUSE
Esclamazione: VOCAL
Rumore accidentale: INCIDENT
Movimento: KINESIC
Frase o parola riformulata/ripetuta: DEL
Parola errata: SIC
Parola corretta: CORR
Forma dialettale: ORIG
Forma regolarizzata: REG
Parola enfatizzata: EMPH
Parola in lingua straniera: FOREIGN
Antroponimo: PERSNAME
Luogo: PLACENAME

Figure 8: Voci dall'Inferno Web Application - Current Implementation in eXist-db

TEI Specifications for a Sustainable Management of Digitized Holocaust Testimonies

Sarah Bénière[1], Floriane Chiffoleau[1,2], Laurent Romary[1,3]
[1]ALMAnaCH, Inria, Paris
[2]Le Mans Université, Le Mans
[3]Directorate for Scientific Information and Culture, Inria
{firstname.surname}@inria.fr

Abstract

Data modeling and standardization are central issues in the field of Digital Humanities, and all the more so when dealing with Holocaust testimonies, where stable preservation and long-term accessibility are key. The EHRI Online Editions are composed of documents of diverse nature (testimonies, letters, diplomatic reports, etc.), held by EHRI's partnering institutions, and selected, gathered thematically and encoded according to the TEI Guidelines by the editors within the EHRI Consortium. Standardization is essential in order to make sure that the editions are consistent with one another. The issue of consistency also encourages a broader reflection on the usage of standards when processing data, and on the standardization of digital scholarly editions of textual documents in general. In this paper, we present the normalization work we carried out on the EHRI Online Editions. It includes a customization of the TEI adapted to Holocaust-related documents, and a focus on the implementation of controlled vocabulary. We recommend the use of these encoding specifications as a tool for researchers and/or non-TEI experts to ensure their encoding is valid and consistent across editions, but also as a mechanism for integrating the edition work smoothly within a wider workflow leading from image digitization to publication.

Keywords: Standardization, TEI-XML, Digital Humanities, European Holocaust Research Infrastructure

1. Introduction

Research in the humanities has taken a turn with the advent of computational methods. The TEI—*Text Encoding Initiative*, or *Text Encoding for Interchange* (Unsworth, 2011; Holmes, 2016)—has been involved in the processing of textual data since 1988 (Schmidt, 2014) and has become a widely used standard in Digital Humanities for structuring textual documents at large (Burnard, 2014; Burnard, 2018). In 2018, the European Holocaust Research Infrastructure[1] (EHRI) published its first online edition of Holocaust testimonies: *BeGrenzte Flucht*, or "Bordered Escape", encoded according to the general TEI All schema, which we will discuss in Section 3.

Numerous digital scholarly editions of textual documents have been published, mainly of historical and literary texts (Schmidt, 2014), and have generally contributed to the advancement of Digital Humanities. Since the 1990s, the TEI has evolved and expanded greatly in a desire to meet the needs of the research community as much as possible (Bauman, 2011; Holmes, 2016; TEI Consortium, 2023). For example, the development of the Shelley-Godwin Archive project (Muñoz and Viglianti, 2015) coincided with the improvement of Chapter 11 of the TEI Guidelines "Representation of Primary Sources" (TEI Consortium, 2023), which proved incredibly useful to the community having to deal with legacy material.

The issue of standardizing encoding practices for specific purposes, such as the publication of Holocaust testimonies, remains to be addressed. Our corpus, the EHRI Online Editions, is a great test-bed for doing so. In the course of taking up the existing editions with the purpose of providing a stable publishing environment for them, we observed disparities and inconsistencies in the encoding from one edition to another due, in particular, to the improvement of the encoders' skills over time. As a result, the need for normalization within the EHRI Online Editions emerged, as well as a broader reflection on the standardization of the encoding of Holocaust-related documents.

This paper presents the TEI customization that we developed for the EHRI Online Editions, and how it can be extended to standardize the encoding of Holocaust-related textual documents. Section 2 presents the EHRI Online Editions, Section 3 deals with data structuration in TEI, and Section 4 focuses on the EHRI TEI customization[2]. Finally, Section 5 discusses the extension of the EHRI specifications to all encoding projects dealing with Holocaust-related documents.

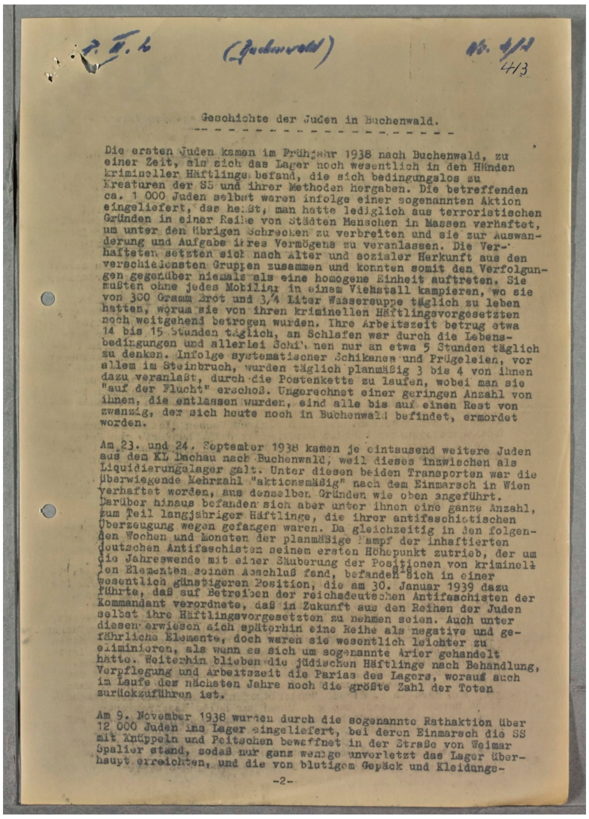
[1] https://www.ehri-project.eu/

[2] https://github.com/SarahBeniere/EHRI-Workflow/blob/main/ENCODING/Guidelines/ODD_EHRI.xml

2. The EHRI Online Editions

EHRI is a transnational consortium funded by the European Union (EU) with partnering institutions all across Europe, Israel, and the United States. It is coordinated by the NIOD Institute for War, Holocaust and Genocides Studies based in Amsterdam, Netherlands. EHRI is currently in its third phase (EHRI-3, 2020-2024), organized in twelve work packages (WP), among which the WP12 "New Approaches to Holocaust Research and Archiving".

Within the framework of WP12, EHRI has already published five online editions[3]. These digital editions are collections of archival documents held by EHRI's various partnering institutions, gathered together and processed by EHRI's editors and made available online[4].



P. III. h (Buchenwald)

413

Geschichte der Juden in Buchenwald.

Die ersten Juden kamen im Frühjahr 1938 nach Buchenwald, zu einer Zeit, als sich das Lager noch wesentlich in den Händen krimineller Häftlinge befand, die sich bedingungslos zu Kreaturen der SS und ihrer Methoden hergaben. Die betreffenden ca. 1 000 Juden selbst waren infolge einer sogenannten Aktion eingeliefert, das heißt, man hatte lediglich aus terroristischen Gründen in einer Reihe von Städten Menschen in Massen verhaftet, um unter den übrigen Schrecken zu verbreiten und sie zur Auswanderung und Aufgabe ihres Vermögens zu veranlassen. Die Verhafteten setzten sich nach Alter und sozialer Herkunft aus den verschiedensten Gruppen zusammen und konnten somit den Verfolgungen gegenüber niemals als eine homogene Einheit auftreten. Sie mußten ohne jedes Mobiliar in einem Viehstall kampieren, wo sie von 300 Gramm Brot und 3/4 Liter Wassersuppe täglich zu leben hatten, worum sie von ihren kriminellen Häftlingsvorgesetzten noch weitgehend betrogen wurden. Ihre Arbeitszeit betrug etwa 14 bis 15 Stunden täglich, an Schlafen war durch die Lebensbedingungen und allerlei Schikanen nur an etwa 5 Stunden täglich zu denken. Infolge systematischer Schikanen und Prügeleien, vor allem im Steinbruch, wurden täglich planmäßig 3 bis 4 von ihnen dazu veranlaßt, durch die Postenkette zu laufen, wobei man sie "auf der Flucht" erschoß. Ungerechnet einer geringen Anzahl von ihnen, die entlassen wurden, sind alle bis auf einen Rest von zwanzig, der sich heute noch in Buchenwald befindet, ermordet worden.

Am 23. und 24. September 1938 kamen je eintausend weitere Juden aus dem KL Dachau nach Buchenwald, weil dieses inzwischen als Liquidierungslager galt. Unter diesen beiden Transporten war die überwiegende Mehrzahl "aktionsmäßig" nach dem Einmarsch in Wien verhaftet worden, aus denselben Gründen wie oben angeführt. Darüber hinaus befanden sich aber unter ihnen eine ganze Anzahl, zum Teil langjähriger Häftlinge, die ihrer antifaschistischen Überzeugung wegen gefangen waren. Da gleichzeitig in den folgenden Wochen und Monaten der planmäßige Kampf der inhaftierten deutschen Antifaschisten seinem ersten Höhepunkt zutrieb, der um die Jahreswende mit einer Säuberung der Positionen von kriminellen Elementen seinen Abschluß fand, befanden sie sich in einer wesentlich günstigeren Position, die am 30. Januar 1939 dazu führte, daß auf Betreiben der reichsdeutschen Antifaschisten der Kommandant verordnete, daß in Zukunft aus den Reihen der Juden selbst ihre Häftlingsvorgesetzten zu nehmen seien. Auch unter diesen erwiesen sich späterhin eine Reihe als negative und gefährliche Elemente, doch waren sie wesentlich leichter zu eliminieren, als wenn es sich um sogenannte Arier gehandelt hätte. Weiterhin blieben die jüdischen Häftlinge nach Behandlung, Verpflegung und Arbeitszeit die Parias des Lagers, worauf auch im Laufe der nächsten Jahre noch die größte Zahl der Toten zurückzuführen ist.

Am 9. November 1938 wurden durch die sogenannte Rathaktion über 12 000 Juden ins Lager eingeliefert, bei deren Einmarsch die SS mit Knüppeln und Peitschen bewaffnet in der Straße von Weimar Spalier stand, sodaß nur ganz wenige unverletzt das Lager überhaupt erreichten, und die von blutigem Gepäck und Kleidungs-

-2-

Figure 1: Example of testimony

Bordered Escape contains testimonies on the forced emigration of the Jewish population of Austria after its annexation in March 1938. It focuses on the situation at the Czechoslovakian border, especially as Czechoslovakia's immigration policy became more and more restricted.

Early Holocaust Testimony is composed of written or transcribed oral testimonies on the persecution of the Jews in Nazi Germany. The testimonies span from 1933, when Adolf Hitler was appointed Chancellor, to the trial of Adolf Eichmann in 1961.

Diplomatic Reports gathers reports written by foreign diplomats stationed in Nazi Germany to their respective Ministry of Foreign Affairs.

From Vienna to Nowhere: The 1939 Nisko Deportations is a collection of testimonies and letters documenting the Nisko Plan, which aimed at creating a Jewish reservation, built by the Jews themselves, in Nisko and Lublin (Poland). The edition focuses on the deportation of approximately 1,600 Jewish men from Vienna to Nisko on the 20th and 26th October 1939 and what became of them.

Documentation Campaign is composed of testimonies of survivors collected in 1945 and 1946 during the "Documentation Campaign" in Prague (Czechoslovakia), which is one of the first postwar initiatives to document the events of the Holocaust.

3. Structuring Data in TEI

3.1. A Standard for Structuring Textual Documents

As briefly mentioned in the introduction, the TEI Guidelines are a widely adopted standard for structuring textual documents in, among other applications, digital scholarly edition projects. They are based on the W3C XML recommendation, and provide "a highly interoperable format" (Schmidt, 2014) with a set of recommended elements that come with a precise syntax and documentation. These recommendations are compiled in the TEI infrastructure as both a technical specification and extensive prose (TEI Consortium, 2023), thus ensuring a common knowledge on the encoding of textual data for research in the humanities. In the case of the EHRI Online Editions, choosing the TEI instead of developing their own arbitrary EHRI tagset has two main advantages:

1. Using the TEI gives relevance to the project, because it aligns with the values and practices of a wider community (Burnard, 2014; 2018) and thus facilitates the integration of the outputs within a wider corpus, as well as it increases the possibility to reuse existing editing, query, or publishing tools.

2. It also aligns with the pre-existing practices of EHRI as an infrastructure, given that their system already relies on XML technology, in

[3] https://www.ehri-project.eu/ehri-online-editions

[4] When unavailable on EHRI's website, the translations of the titles of the editions in English are our own.

particular on EAD-XML (Alexiev et al., 2019; Levy, 2019; Romary and Riondet, 2019).

According to Lou Burnard (2014; 2018), the success of the TEI in Digital Humanities projects lies in its three main characteristics:

1. Contrary to typical word processors like Microsoft Word or LibreOffice Writer—which tend to focus on the aesthetic rendering of the text—a TEI encoding is semantic. It is particularly useful for named entities disambiguation tasks. For example, the character string “Warsaw” could either refer to the city and capital of Poland Warsaw, or to the Warsaw Ghetto (Figure 2).

```
<placeName type="city">Warsaw</placeName>
<placeName type="ghetto">Warsaw</placeName>
```

Figure 2: Disambiguation of “Warsaw” in TEI

2. A TEI-XML file, like all XML files, is a succession of characters that both humans and machines can read and understand. As a result, the action of opening and reading the content of a TEI-encoded text is independent of any software, whereas a Microsoft Word document (.docx), for example, requires at the very least a word processor to open.

3. The TEI recommendations are sustained by the TEI Consortium and improved by the continuous involvement of the TEI community. In addition, because the Guidelines are available online and the community is active, it makes it an accessible technology for beginners.

3.2. Best Practices and Standardization

When encoding a text in XML, the encoder is free to use whatever tags they want and to give them a meaning of their own. In his article on TEI conformance, Lou Burnard (2018) gives the example of the <p> tag. Generally speaking, <p> is used to encode a paragraph, but we could decide that in the case of our encoding it means “potato”. This example highlights the relevance of a standard like the TEI. Nevertheless, criticism has been expressed toward the TEI as being too wide and too restrictive at the same time, or the choice of the tags being guided by human interpretation of the text, thus leading to an impediment of interoperability (Bauman, 2011; Schmidt, 2014).

While we agree with the fact that the encoders choosing which element they want to draw attention to makes interchange difficult *per se*, because it implies that everyone is aware of the purpose of said encoding, we argue that a solution could be the implementation of a schema and documentation by means of an ODD. The ODD—for *One Document Does-it-all*—is a TEI-XML file which contains both a customization of the TEI and its associated documentation. From the ODD file, we can derive a RelaxNG validation schema with the customized TEI specifications, but also the prose documentation for the human reader to understand the purpose and extent of the project’s encoding. In addition, an ODD established by an experienced TEI user can help a beginner to make sure their encoding is valid.

We previously alluded to a few inconsistencies in the TEI encoding of the EHRI Online Editions. This is due, on the one hand, to the improvement of the encoders’ skills over time, and on the other hand to the fact that the declared validation schema was “TEI All”. As the name suggests, the TEI All schema encompasses all elements and attributes from the TEI. However, no project would ever use them all, thus emphasizing the relevance of a TEI customization, which “expresses how a given project has chosen to interpret the general principles enumerated by the Guidelines, as well as formally specifying which particular component of the Guidelines it uses” (Burnard, 2018). In addition, this profusion of TEI elements can easily lead to confusion between several elements (typically <bibl>, <biblFull> and <biblStruct>), especially for encoders who might not yet be familiar with TEI-XML.

The TEI customization and specifications associated can help define a framework within which the encoders can work and apply best practices. For example, a good practice in TEI-XML consists in structuring the <body> of the <text> with at least one <div> (division) element (Figure 3). We decided to make this a mandatory rule in the EHRI specifications (Figure 4).

```
<body>
    <div type="transcription" xml:lang="de">
        <pb n="1"/>
        <p>[...]</p>
    </div>
</body>
```

Figure 3: Minimal template for the <body>

```
<schemaSpec ident="body" mode="change">
    <!-- div is mandatory in the body -->
    <content>
        <elementRef key="div" minOccurs="1"
    maxOccurs="unbounded"/>
    </content>
</schemaSpec>
```

Figure 4: Schema specification for <body>

This framework applies to both published and future editions. For editions that have already been published, we wrote a Python script to automatically apply the new RelaxNG schema to all the XML files[5]. For future editions, the schema should be applied instead of “TEI All” from the beginning.

As a final general good practice, we recommended using international norms like ISO to fill in the value for an attribute. The ISO norms we included in the EHRI specifications are:

1. ISO 639[6] codes for the representation of languages;
2. ISO 3166[7] codes for the representation of names of countries;
3. ISO 8601[8] standard for dates (`YYYY-MM-DD`).

4. TEI Customization for Holocaust Testimonies

4.1. Normalizing the EHRI Online Editions

Until now, the texts selected by the editors were transcribed and encoded manually (Frankl et al., 2018), which raised two main issues:

1. It is an extremely time-consuming and tedious task;
2. It is a source of encoding mistakes.

In order to write the ODD for the EHRI Online Editions, we needed to analyze the encoding practices of the encoders for the editions that had already been published: “Bordered Escape”, “Early Holocaust Testimony”, “Diplomatic Reports”, and “Nisko”. We noticed, for instance, recurring mistakes in the spelling of attribute values (i) or the usage of different languages (ii): (i) `@type="subeject"` or (ii) `@type="subjekt"` (German) instead of `@type="subject"`. Even though they may refer semantically to the same entity—a term (`<term>`) for example—the machine will consider them as different instances. This leads to an incorrect count of the occurrences and to referencing mistakes that are not easily detectable.

One of the normalizing aspects for the EHRI Online Editions which we considered important is the language chosen for encoding the metadata. In an edition gathering documents from different holding institutions, the metadata should be filled in thoroughly. In a spirit of data reuse, we thought that all metadata should appear at least in English. Some metadata can be translated, like the title of the document (Figure 5) or the name of its holding institution. For example, the original name for the Jewish Museum in Prague is “Židovské muzeum v Praze” (Czech), but we estimated that the most commonly understood language among EHRI partners would be English. Therefore, we established English as the main language for encoding the metadata.

```
<title xml:lang="en">List of Viennese Nisko
    deportees who died in Kamensk-Uralski</title>
<title xml:lang="de">Liste von Wiener Nisko-
    Deportierten, die in Kamensk-Uralski verstarben
    </title>
```

Figure 5: Encoding of the title of a document

Normalizing the EHRI Online Editions is the first step toward TEI specifications for the standardization of Holocaust-related documents in TEI-XML. Indeed, the ODD for the EHRI Online Editions serves three purposes:

1. Avoiding encoding mistakes as much as possible;
2. Setting up good encoding practices in general, especially in case any of the encoders is not yet familiar with TEI-XML;
3. Establishing a validation schema particularly suitable for Holocaust-related textual documents, derived from the ODD, insofar as simultaneously harmonizing the previously published EHRI digital editions and ensuring the consistency of the future ones.

4.2. Points of Interest in the EHRI TEI Specifications

Language Codes (ISO 639) Even though this is a mistake that was rapidly corrected in the second edition, we found some inconsistency in the codes chosen for the representation of languages as values for the `@xml:lang` attribute. It is naturally tempting to use a code that would be correct in one’s own native language, which can result in referencing mistakes like the misspelling of “subject” we mentioned in Section 4.1. A very common example of such bias is the representation of the German language: we could imagine either `"de"` for “Deutsch” (German), `"ger"` for “German” (English), and even `"all"` for “Allemand” (French). While all these codes are correct representations of the German language, they must not be used all at once within the same edition. As a result, we recommended that the encoders use the codes provided by the ISO 639 norm (Figure 6), available

[5] https://github.com/EHRI/ehri-online-editions

[6] https://www.iso.org/iso-639-language-code

[7] https://www.iso.org/iso-3166-country-codes.html

[8] https://www.iso.org/iso-8601-date-and-time-format.html

through the Iana Language Subtag Registry[9].

```
<valList mode="add" type="semi">
    <valItem ident="cs">
        <desc>Czech</desc>
    </valItem>
    <valItem ident="da">
        <desc>Danish</desc>
    </valItem>
    <valItem ident="de">
        <desc>Deutsch</desc>
    </valItem>
    <valItem ident="el">
        <desc>Modern Greek</desc>
    </valItem>
    <valItem ident="en">
        <desc>English</desc>
    </valItem>
    <valItem ident="es">
        <desc>Spanish</desc>
    </valItem>
    <valItem ident="fr">
        <desc>French</desc>
    </valItem>
    <valItem ident="he">
        <desc>Hebrew</desc>
    </valItem>
    <valItem ident="hu">
        <desc>Hungarian</desc>
    </valItem>
    <valItem ident="it">
        <desc>Italian</desc>
    </valItem>
    <valItem ident="ja">
        <desc>Japanese</desc>
    </valItem>
    <valItem ident="nl">
        <desc>Dutch</desc>
    </valItem>
    <valItem ident="pl">
        <desc>Polish</desc>
    </valItem>
    <valItem ident="ru">
        <desc>Russian</desc>
    </valItem>
    <valItem ident="sk">
        <desc>Slovak</desc>
    </valItem>
    <valItem ident="uk">
        <desc>Ukrainian</desc>
    </valItem>
    <valItem ident="yi">
        <desc>Yiddish</desc>
    </valItem>
</valList>
```

Figure 6: Language codes used by EHRI

Implementing Controlled Vocabulary The EHRI Portal[10] presents itself as one of the main resources about the Holocaust for it gathers information on archival sources from across the world. One of their primary achievements is the creation of controlled vocabulary. Among the EHRI terms, some are identified as linguistically distinct because they are vocabulary coined by the Nazis or specifically used in reference to the concentration and extermination camps. In the continuity of the encoding work performed by the EHRI encoders, we modified the specifications for the `<distinct>` element. As a result, we made the `@type` attribute mandatory and suggested a semi-open list of values containing `"camp_language"` and `"nazi_language"` (Figure 7). Hence, a dialog box with the list of possible values appears every time the `@type` attribute from the `<distinct>` element is filled in when encoding a text.

```
<elementSpec ident="distinct" mode="change">
    <attList>
        <!-- @type is mandatory and its value is
    either camp_language or nazi_language -->
        <attDef ident="type" mode="change" usage="
    req">
            <valList mode="add" type="semi">
                <valItem ident="camp_language"/>
                <valItem ident="nazi_language"/>
            </valList>
        </attDef>
    </attList>
</elementSpec>
```

Figure 7: Specifications for `<distinct>`

Including Translation(s) in a Single File The EHRI ODD is part of a broader workflow for processing Holocaust-related documents[11]. The last step of this workflow is the publication of the editions on a TEI Publisher[12] application dedicated to all the EHRI digital editions. In order to do so, we decided to include the documents in their original language as well as their translation(s) within a unique file bearing the EHRI identifier, for example `"EHRI-ET-WL16560413"` (Figure 1). This is done by ensuring the structuration of the `<body>` with first-level `<div>` (division) elements specified with the attributes `@type` (Figure 8) and `@xml:lang` (ISO 639 values).

Encoding Template for the `<teiHeader>` As we mentioned previously, particular attention must be given when encoding the documents' metadata. We created a template (Appendix A) to make sure that no available piece of information is missing. A good practice that needs to be implemented by the EHRI encoders is the use of the `<revisionDesc>` so as to follow all the modifications made within the file. The template also contains fields that are already filled in because their value is consistent for every single file: `<affiliation>` and `<authority>` will always be EHRI, and we share the documents according to the Creative Commons Attribution 4.0 International license (CC BY 4.0)[13].

[9] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

[10] https://portal.ehri-project.eu/

[11] https://github.com/SarahBeniere/EHRI-Workflow/tree/main

[12] https://teipublisher.com/exist/apps/tei-publisher-home/index.html

[13] https://creativecommons.org/licenses/by/4.0/

```
<elementSpec ident="div" mode="change">
    <constraintSpec scheme="schematron" ident="div
    -1">
        <constraint>
            <s:rule context="tei:TEI/text/body/div[
    @type]">
                <s:assert test="@type='
    transcription' or @type='translation'">Value
    for @type in first-level division is either
    transcription or translation</s:assert>
            </s:rule>
        </constraint>
    </constraintSpec>
    <attList>
        <!-- @type is mandatory and its value
    should either be transcription or translation
    -->
        <attDef ident="type" mode="change" usage="
    req">
            <valList mode="add" type="semi">
                <valItem ident="transcription"/>
                <valItem ident="translation"/>
            </valList>
        </attDef>
    </attList>
</elementSpec>
```

Figure 8: Specifications for first-level `<div>`

5. Discussion and Conclusion

This paper presents the TEI specifications developed in the context of the EHRI Online Editions. The implementation of the EHRI ODD is organized in two steps: the processing of editions that have already been published, and the processing of future digital editions. The texts of the previous editions must be validated against the RelaxNG schema derived from the EHRI ODD, and we have experimented with a Python script to automatically apply the new schema to the texts that were already encoded. As for future editions, the texts are to be encoded according to the TEI specifications defined in the EHRI ODD[14]. We present the EHRI ODD as a starting point for the standardization of encoding practices regarding Holocaust-related textual documents. Indeed, using semi-open lists for attribute values for example allows an extension to documents in more languages, and/or containing other types of specific vocabulary. As we are strong advocates of the open science approach, we make the EHRI ODD public and reusable according to the terms of the CC BY 4.0 license. Therefore, it can serve as a basis for the development of more complete encoding guidelines for Holocaust testimonies, following the “ODD chaining” tutorial by Lou Burnard (2016) for instance.

6. Acknowledgements

This work has been carried out in the context of the EHRI-3 project funded by the European Commission under the call H2020-INFRAIA-2018-2020, with grant agreement ID 871111 and DOI 10.3030/871111. We would like to warmly thank our colleagues at EHRI for their precious work on the digital editions, mainly Michael Bryant, Maria Dermentzi, Michal Frankl, Aneta Plzáková, Wolfgang Schellenbacher, and Magdalena Sedlická.

We would also like to thank our colleagues Lydia Nishimwe and Hugo Scheithauer for their thorough proofreading of the paper.

7. Bibliographical References

Vladimir Alexiev, Ivelina Nikolova, and Neli Hateva. 2019. Semantic Archive Integration for Holocaust Research. *Umanistica Digitale*, 1(4):131–175.

Syd Bauman. 2011. Interchange vs. Interoperability. In *Proceedings of Balisage: The Markup Conference 2011*.

Lou Burnard. 2014. *What is the Text Encoding Initiative?* OpenEdition Press.

Lou Burnard. 2016. *ODD chaining for Beginners*. GitHub.

Lou Burnard. 2018. What is TEI Conformance, and Why Should You Care? *Journal of the Text Encoding Initiative*, 1(12).

Michal Frankl, Michael Bryant, Jessica Green, Wolfgang Schellenbacher, and Magdalena Sedlická. 2018. Edition of Documents. Technical Report 654164 (H2020-INFRAIA-2014-2015), European Holocaust Research Infrastructure.

Martin Holmes. 2016. Whatever happened to interchange? *Digital Scholarship in the Humanities*, 32:i63–i68.

Michael Levy. 2019. Some Perspectives on the Practice of Sharing Collection Data. *Umanistica Digitale*, 1(4):21–32.

Trevor Muñoz and Raffaele Viglianti. 2015. Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive. *Journal of the Text Encoding Initiative*, 1(8).

Laurent Romary and Charles Riondet. 2019. Towards Multiscale Archival Digital Data. *Umanistica Digitale*, 1(4):89–99.

Desmond Schmidt. 2014. Towards an Interoperable Digital Scholarly Edition. *Journal of the Text Encoding Initiative*, 1(7).

[14] As of now, new EHRI editions have not been prepared yet.

TEI Consortium, editor. 2023. *TEI P5: Guidelines for Electronic Text Encoding and Interchange (ver. 4.7.0)*. TEI Consortium.

John Unsworth. 2011. Computational Work with Very Large Text Collections. *Journal of the Text Encoding Initiative*, 1(1).

A. Template for the `<teiHeader>`

```xml
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title xml:lang="en"/>
            <title xml:lang=""/>
            <principal>
                <affiliation>
                    <orgName ref="https://www.ehri-project.eu">
                       European Holocaust Research Infrastructure
                    </orgName>
                </affiliation>
            </principal>
            <respStmt>
                <resp/>
                <persName/>
            </respStmt>
        </titleStmt>
        <publicationStmt>
            <authority>
                <ref target="https://www.ehri-project.eu">European Holocaust Research Infrastructure</ref>
            </authority>
            <availability>
                <licence target="http://creativecommons.org/licenses/by-sa/4.0">
                   Attribution-ShareAlike 4.0 International
                </licence>
            </availability>
        </publicationStmt>
        <seriesStmt>
            <title ref="{link to the online edition}"/>
        </seriesStmt>
        <sourceDesc>
            <msDesc>
                <msIdentifier>
                    <institution>
                        <orgName/>
                        <address>
                            <street>
                                <num/>
                            </street>
                            <postCode/>
                            <settlement/>
                            <country/>
                        </address>
                    </institution>
                    <collection/>
                    <idno/>
                </msIdentifier>
                <physDesc>
                    <p/>
                </physDesc>
            </msDesc>
            <bibl>
                <textLang/>
            </bibl>
        </sourceDesc>
    </fileDesc>
    <encodingDesc>
        <projectDesc>
            <p xml:lang="en"/>
        </projectDesc>
    </encodingDesc>
    <profileDesc>
        <creation>
            <origDate when=""/>
            <origPlace ref="{GeoNames link}"/>
            <persName ref="{EHRI entity}"/>
        </creation>
        <textClass>
            <catRef target="{link to EHRI portal}"/>
            <keywords>
                <term/>
            </keywords>
        </textClass>
        <langUsage>
            <language ident=""/>
        </langUsage>
        <abstract>
            <p xml:lang="en"/>
        </abstract>
    </profileDesc>
    <revisionDesc>
        <change when="" who="{}"/>
    </revisionDesc>
</teiHeader>
```

Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools

Maria Dermentzi*, **Hugo Scheithauer***
King's College London, Inria Paris, École pratique des hautes études
London, United Kingdom, Paris, France
name.1.surname@kcl.ac.uk, name.surname@inria.fr

Abstract

The European Holocaust Research Infrastructure (EHRI) aims to support Holocaust research by making information about dispersed Holocaust material accessible and interconnected through its services. Creating a tool capable of detecting named entities in texts such as Holocaust testimonies or archival descriptions would make it easier to link more material with relevant identifiers in domain-specific controlled vocabularies, semantically enriching it, and making it more discoverable. With this paper, we release EHRI-NER, a multilingual dataset (Czech, German, English, French, Hungarian, Dutch, Polish, Slovak, Yiddish) for Named Entity Recognition (NER) in Holocaust-related texts. EHRI-NER is built by aggregating all the annotated documents in the EHRI Online Editions and converting them to a format suitable for training NER models. We leverage this dataset to fine-tune the multilingual Transformer-based language model XLM-RoBERTa (XLM-R) to determine whether a single model can be trained to recognize entities across different document types and languages. The results of our experiments show that despite our relatively small dataset, in a multilingual experiment setup, the overall F1 score achieved by XLM-R fine-tuned on multilingual annotations is 81.5%. We argue that this score is sufficiently high to consider the next steps towards deploying this model.

Keywords: Holocaust Testimonies, Named Entity Recognition, Transformers, Multilingual, Transfer Learning, Digital Editions

1. Introduction

Launched in 2010, the European Holocaust Research Infrastructure (EHRI)[1] aims to support Holocaust research by making information about dispersed archival material held by institutions around the world more accessible and interconnected through the EHRI Portal[2] (Blanke et al., 2017). While the EHRI Portal is EHRI's flagship service, the EHRI Consortium is offering a series of additional resources, tools, and services that help researchers and archivists describe, analyze, enrich, and present Holocaust-related material using innovative methods (de Leeuw et al., 2018). Apart from the EHRI Portal, of particular relevance to this paper are the EHRI controlled vocabularies, the EHRI authority sets, and the EHRI Online Editions[3].

As an aggregator of multilingual Holocaust-related archival material from diverse institutions, the EHRI Portal is faced with a significant challenge relating to the fact that this material is often described not only in various languages but also using a variety of methodologies and in-house, language-specific controlled vocabularies that need to be normalized to a shared vocabulary to be smoothly ingested in the EHRI Portal (Erez et al., 2020). For this reason, EHRI has developed custom controlled vocabularies and authority sets mainly derived from already existing ones developed by institutions such as Yad Vashem (YV), the United States Holocaust Memorial Museum (USHMM), Arolsen Archives, etc. (Rodriguez et al., 2016; Erez et al., 2020), covering lists of concentration camps, ghettos, subject headings, personalities and corporate bodies[4]. These vocabularies are primarily used for indexing purposes in the EHRI Portal, allowing for semantic search (Colavizza et al., 2019) through keyword-based browsing and play a crucial role in achieving EHRI's goal of interlinking multilingual and heterogeneous Holocaust collections. They are also used to enhance the EHRI Online Editions and articles in the EHRI Document Blog[5] with more information and references to the EHRI Portal.

However, creating links between resources hosted across different EHRI services and the EHRI vocabularies is a resource-intensive process, usually done manually. Creating a tool capable of detecting named entities (NE) in texts such as Holocaust testimonies or the text in Holocaust-related archival descriptions would make it easier to link more material with relevant identifiers in the EHRI

[*] Both authors contributed equally to this work.

[1] EHRI project website. Accessed 2/27/2024.

[2] EHRI portal website. Accessed 2/27/2024.

[3] See EHRI controlled vocabularies, EHRI authority sets, and EHRI online editions. Accessed 2/27/2024.

[4] The aforementioned lists and sets are available online. See camps, ghettos, terms, personalities, and corporate bodies. Accessed 2/27/2024.

[5] See EHRI Document Blog. Accessed 2/27/2024.

vocabularies, semantically enriching it and making it more discoverable in the Portal and other EHRI services. The significance that reliable Named Entity Recognition (NER) and entity linking (EL) tools may have for EHRI has been highlighted in previous work (Rodriguez et al., 2012; de Leeuw et al., 2018). Having access to a good NER tool can help with building a reliable EL tool. EHRI partners have previously experimented with the development of such tools (Rodriguez et al., 2012; de Leeuw et al., 2018; Nikolova and Levy, 2018). However, since the publication of the most recent paper related to EHRI and NER (de Leeuw et al., 2018), EHRI's growth in resources and advances in Machine Learning (ML) promise better results compared to earlier experiments. In this paper, we report on recent work towards Holocaust-related NER.

Specifically, we treat the EHRI digital scholarly editions (i.e., EHRI Online Editions) as a dataset for training and evaluating ML-powered NER models. We have converted all available Extensible Markup Language (XML) files from the EHRI Online Editions into a trainable corpus in a format suitable for NER and have leveraged this dataset (See Table 2) to fine-tune a multilingual language model for NER. The resulting model can be used as part of a pipeline whereby, upon inputting some text into a tool that supports our models, potential named entities within the text will be automatically pre-annotated in a way that helps users detect them faster and link them to their associated controlled vocabulary entities. This has the potential to facilitate metadata enrichment of descriptions in the Portal and enhance their discoverability. It would also make it easier for EHRI to develop new Online Editions and unlock new ways for archivists and researchers within the EHRI network to organize, analyze, and present their materials and research data in ways that would otherwise require a lot of tedious work.

Our contributions are: the EHRI-NER dataset, a multilingual NER model for Holocaust-related texts, and experiments studying the multilingual learning and cross-lingual transfer capabilities of Deep Learning NER techniques. In what follows, we describe related work (Section 2) and provide detailed information on the source of our dataset, the EHRI Online Editions (Section 3). Subsequently, we detail how we put together the dataset (Section 4) and how we designed and carried out our fine-tuning experiments (Section 5). We conclude with a summary and future research pathways (Section 6).

2. Related Work

Previously, EHRI experimented with applying off-the-shelf NER tools to the Optical Character Recognition (OCR) output of type-written Holocaust survivor testimonies and newsletters for the crew of H.M.S. Kelly (Rodriguez et al., 2012). Due to the lack of an already available annotated corpus for domain-specific NER tools, Rodriguez et al. (2012) manually annotated the OCRed corpus compiled for their experiments. Given the lack of resources, their experiments remained limited and focused on comparing which of the then-existing NER tools yielded the best results. The maximum total F1 score achieved across all tools and datasets under consideration was 60% (Rodriguez et al., 2012). In 2018, de Leeuw et al. (2018) published another paper detailing EHRI's efforts to offer reliable NER services for the Holocaust domain. They reiterated the lack of suitable corpora and crafted their own gold corpus by crowd-sourcing annotations on transcripts of oral testimonies provided by the USHMM (de Leeuw et al., 2018; Nikolova and Levy, 2018). They used this corpus to develop person and location extraction services. Their methodology included fine-tuning and extending commercial software and they achieved an F1 score of 77% for person extraction. For location extraction, they adapted a proprietary service to tag and disambiguate locations in Holocaust testimonies. The details of these tools are not specified but the authors reported a resulting F1 score of 91% for the disambiguated place-related access points, although it is unclear how the first part of their pipeline (i.e. the tagger) performed. To our knowledge, neither the purpose-built NER datasets nor the EHRI-specific tools developed during earlier work are publicly available today or were formally deployed as EHRI services.

Apart from EHRI-related efforts, there is a broader interest in applying NER tools on Holocaust-related texts (Ezeani et al., 2023; Carter et al., 2022) as well as in developing domain-specific ones. Notable examples include Mattingly's (2021a; 2021b) lessons on Holocaust NER and Nanomi Arachchige et al.'s (2023) paper detailing their work on compiling and annotating an English corpus for Holocaust-related NER, which they used to train and evaluate rules-based and transformer-based (Vaswani et al., 2017) tools. Consistent with other publications (Luthra et al., 2023; Ehrmann et al., 2023), many of the Transformer-based models included in Nanomi Arachchige et al.'s (2023) experiments achieved high F1 scores across most of the entities considered, encouraging us to select a similar architecture for our experiments.

However, since the material processed by EHRI is diverse and multilingual, we wanted to work towards developing a single multilingual NER model that would leverage multilingual learning for cross-lingual transfer (Mueller et al., 2020; Ehrmann et al., 2023; Schweter et al., 2022; Wu et al., 2020). Multilingual NER in historical documents has seen a growing interest amongst the Digital Humanities (DH), Natural Language Processing (NLP), and cultural heritage communities (Ehrmann et al., 2023). In 2022, Ehrmann et al. (2022) introduced a shared task on NER and EL in multilingual historical documents, encouraging researchers to study approaches that can work well across different contexts and languages. Ehrmann et al. (2022) acknowledge that advances in AI thanks to the Transformer architecture and the increased availability of suitable resources create new opportunities for working towards such solutions. The same is true in the EHRI context, where since the work of Rodriguez et al. (2012) and Nikolova and Levy (2018), EHRI has produced a series of manually annotated digital scholarly editions. Although the original purpose of these editions was not to provide a dataset for training NER models, we argue that they nevertheless constitute a high-quality resource that is suitable to be used in this way. We therefore repurposed them to train multilingual Transformer-based NER models testing the hypothesis that we now have enough resources to develop a single domain-specific tool that can work reliably well across different languages and document types encountered in EHRI collections.

3. EHRI Online Editions

Since 2018, the EHRI Consortium has supported the development and publication of six Holocaust-related digital scholarly editions[6] (EHRI-Consortium, 2021; Frankl and Schellenbacher, 2018; Frankl et al., 2023; Frankl and Schellenbacher, 2023; Frankl et al., 2020; Garscha et al., 2022). Each edition enables digital access to facsimiles and transcripts of thematically related documents held by different EHRI partner institutions through a single web interface and unlocks new ways of presenting and browsing through historical sources using digital tools. Publishing a digital edition is a resource-intensive process. Notwithstanding the extensive archival research needed for selecting the documents, additional steps include transcribing and translating them and, most importantly, annotating words and phrases found within these texts and creating links with entities in controlled vocabularies provided by EHRI and third parties. Currently, this annotation is done manually by or under the supervision of subject matter experts, ensuring a high quality of annotations[7]. We repurposed these resources to convert them into a dataset suitable for training NER models, which we consider as a gold standard.

Each EHRI Online Edition consists of digitized documents originating from various archives that are selected, edited, and annotated by EHRI researchers using the Text Encoding Initiative (TEI) P5 standard (TEI Consortium, 2023), an XML schema, which supports their online publication. Editions enhance the edited documents by contextualizing the information contained within them and linking them to EHRI vocabularies and descriptions, and by visualizing georeferenced entities through interactive maps. Thanks to their encoding in TEI, they are fully searchable and can be filtered using facets such as spatial locations, topics, persons, organizations, and institutions. All documents within an edition have a transcript, either in their original language, a translation, or both, and have access to their facsimile. EHRI Editions are published without a regular schedule and it is possible to update them with new material or improve the already published documents. In the following paragraphs, we present each edition individually.

Begrenzte Flucht Edition The BeGrenzte Flucht (BF) edition (Frankl and Schellenbacher, 2018) gathers documents kept in various Czech and Austrian archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938, including official reports, correspondence, diplomatic notes, newspaper reports, and documents from Jewish aid organizations. The BF edition is in German and the vast majority of documents, if not originally in German, have been translated into German. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included.

Early Holocaust Testimonies Edition The Early Holocaust Testimony (EHT) edition (Frankl et al., 2020) contains selected and edited testimonies and reports kept in five different archives: the Wiener Holocaust Library in London, Yad Vashem in Jerusalem, the Jewish Historical Institute in Warsaw, the Hungarian Jewish Archives in Budapest, and the Jewish Museum in Prague. All of the documents have an English translation but transcripts of the original documents in Czech, German, Hungarian, Polish, Dutch, and Yiddish are provided.

Diplomatic Reports Edition The Diplomatic Reports (DR) edition (EHRI-Consortium, 2021) gathers documents created by the diplomatic staff of allied countries, opponents, and neutral countries. They all report on the German occupation. They include reports from the diplomatic staff of Denmark, Italy, Japan, Hungary, Slovakia, and the US. All of the documents have been translated into English, regardless of their original language.

[6] At the time of writing: 2/26/2024.

[7] More info about this process can be found on the website of each edition.

Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 Edition The Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 (ND) edition (Garscha et al., 2022) was created in cooperation with the Documentation Archive of the Austrian Resistance. It gathers documents on the history and the fate of the Viennese Jewish deportees to Nisko, Poland, in 1939. The source documents are from various archival institutions in different countries and are provided in German.

Documentation Campaign Edition The Documentation Campaign (DC) edition (Frankl et al., 2023) gathers documents held by the Jewish Museum in Prague and by Yad Vashem consisting of Holocaust survivor testimonies and photographs collected within the framework of the so-called "Documentation Campaign" in Prague, one of the earliest postwar projects to document the events of the Shoah, collecting evidence, documents, and witness testimonies. All of the documents have been translated into English but transcripts of the original documents in Czech and German are provided.

Uzavřít Hranice Edition Similar to the BF edition, the Uzavřít Hranice (UH) edition (Frankl and Schellenbacher, 2023) gathers documents kept in various Czech, Austrian, and other archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938. The UH edition is in Czech and the vast majority of documents, if not originally in Czech, have been translated into Czech. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included.

Since the EHRI Online Editions cover a variety of languages, document types, periods, and thematic and spatial areas of focus, training NER models on this dataset may lead to tools that can generalize better on different types of Holocaust-related documents, compared to training them only on testimony-based corpora like in previous work. This will hopefully make our models more robust and interoperable across different EHRI services.

4. The EHRI-NER Dataset

This section presents EHRI-NER, a multilingual NER dataset derived from the EHRI Online Editions. We fully released EHRI-NER on Hugging Face and GitHub[8].

4.1. Languages and Subsets

We sorted all TEI XML files available from the EHRI Online Editions by language. The resulting EHRI-NER dataset includes nine languages: Czech (cs), German (de), English (en), French (fr), Hungarian (hu), Dutch (nl), Polish (pl), Slovak (sk), and Yiddish (yi). We created a subset for each language since they are not represented in the same proportion.

As presented in Section 3, the dataset includes official reports, correspondences, diplomatic notes, newspaper reports, and testimonies. The creation dates of the documents span from 1936 to 2001.

4.2. From TEI XML to the IOB Format

To build the subsets, we created a Python script to parse the TEI XML documents and convert them to the CoNLL Inside-Outside-Beginning (IOB) format (Sang and De Meulder, 2003), which is typical for NER datasets (Ehrmann et al., 2016)[9].

The BF, UH, DC, and EHT editions all include translations of some of their original transcribed documents. To avoid contaminating our validation and test sets, we filtered them out. Additionally, both the BF and the UH editions contain some documents that overlap. We also filtered these out to avoid having duplicates in our dataset.

4.3. Entity Classes

Given that the primary purpose of this work is to enhance the services and facilitate the work of EHRI stakeholders, we used a custom typology of entity classes that corresponds better to how we envision deploying this tool in the EHRI environment, extending the CoNLL typology (Sang and De Meulder, 2003) to include classes such as camps and ghettos, which correspond to custom EHRI vocabularies used when annotating Holocaust materials to produce new EHRI Editions. However, our typology is coarser compared to more fine-grained typologies found in similar work (Nanomi Arachchige et al., 2023). We extracted all TEI elements `<persName>`, `<placeName>`, `<orgName>`, and `<date>` from the selected TEI XML files. The `<placeName>` element sometimes includes an attribute @type to indicate whether it is referencing a concentration camp or a ghetto. We distinguish between `<placeName>`, `<placeName type="camp">`, and `<placeName type="ghetto">` to include fine-grain camp and ghetto entities in addition to the coarse-grain location entity. The conversion table is presented in Table 1.

EHRI TEI XML files also contain the `<term>` entity, used for annotating various subjects related to

[8] See the EHRI-NER organization on Hugging Face to access the model and dataset and the EHRI-NER GitHub repository to access the dataset subsets per language.

[9] Our script is available on GitHub.

the Holocaust and for linking them with their associated entries in the EHRI vocabulary of terms[10]. However, we have chosen to consider these instances as non-entity tokens, as their broad coverage of themes, their variability, and lack of semantic regularity in how they are used in annotations make them unsuitable in a token classification context. Had we included them in our typology, we hypothesize that the NER models would tag a disproportionate number of tokens as terms, rendering the output noisy and confusing. Instead, EHRI is working on a different solution for extracting subject metadata, which is outside the scope of this paper.

The EHRI-NER dataset includes a total of 505758 tokens, with 5351 person entities, 9399 location entities, 1867 organization entities, 2237 date entities, 528 ghetto entities, and 1229 camp entities. The distribution of tokens and entity classes is detailed in Table 2.

TEI XML Element	Entity Class
`<persName>Helene Hirsch</persName>`	Person
`<placeName>Berlin</placeName>`	Location
`<orgName>Gestapo</orgName>`	Organization
`<date when="1937-10">Oct.  1937</date>`	Date
`<placeName type="camp">Auschwitz</placeName>`	Camp
`<placeName type="ghetto">getcie</placeName>`	Ghetto

Table 1: Conversion table for TEI XML Elements and Entity Classes.

4.4. Data Format and Preprocessing

We chose to convert TEI annotations and non-entity tokens into the CoNLL IOB format, as presented in Sang and De Meulder (2003) (see Table 3). The IOB format ensures that our dataset is interoperable with common NER tools. Each token and its annotation have been put on a separate line and there is an empty line after each sentence, as shown in the following example:

```
Von O
Gross B-CAMP
- I-CAMP
Rosen I-CAMP
Bahntransport O
nach O
Buchenwald B-CAMP
. O
```

Each language subset has been tokenized at the sentence and word levels. We used SpaCy (Honnibal et al., 2020) and its multi-language pipeline to process each subset[11].

5. Experimental Setup

We conducted two experiments to determine whether our dataset was sufficiently large for fine-tuning a reliable NER model that could be used in a real-life setting, e.g. speeding up named entity annotation when curating a new EHRI Online Edition. We also leveraged the multilingual aspect of our dataset to test XLM-RoBERTa (XLM-R) (Conneau et al., 2020) in a low-resource setting, as our dataset is significantly smaller than, for instance, the CONLL2003 NER dataset used for evaluating this model on a token classification task. In this section, we describe the model that we used for fine-tuning, the experiments we conducted on the dataset, and their results.

5.1. Model

We chose to experiment with the multilingual Transformer-based masked language model `XLM-RoBERTa-large` (Conneau et al., 2020) as it demonstrates high efficacy in multilingual settings and strong cross-lingual transfer capabilities, especially on token classification tasks, without sacrificing per-language performance[12]. According to Nanomi Arachchige et al. (2023), this model outperforms the multilingual hmBERT (Schweter et al., 2022) model which was pre-trained on German, French, Swedish, Finnish, and English historical newspapers (thus not pre-trained in all of the languages present in our dataset). It is important to note that XLM-R has seen all languages represented in the EHRI-NER dataset during its pre-training.

The same fine-tuning parameters were kept for all our experiments. The learning rate is set at $3e^{-5}$, the number of epochs for training at 3 to avoid overfitting, the weight decay at 0.01, and the train and evaluation batch size at 16.

5.2. Experiments

Experiment 1: We fine-tuned XLM-R on all subsets (cs, de, en, fr, hu, nl, pl, sk, yi) to evaluate the overall performance of the model on a multilingual level. Instead of relying on a simple shuffle, and to ensure that all languages are represented in the train, validation, and test set, we first split each subset into train (80%), validation (10%), and test (10%) sets, using a seed of 42 for result reproducibility[13]. Each split subset is then concatenated and the final dataset is used for fine-tuning. Our objective was to acquire a single fine-tuned model

[10] See the EHRI Terms database. Accessed 2/27/2024.

[11] See the multi-language pipelines available on SpaCy website. Accessed 2/27/2024.

[12] See `XLM-RoBERTa-large` model card on Hugging Face website. Accessed 02/27/2024.

[13] The mentioned seed used for splitting the subsets was used for all the experiments.

ISO code	Language	Tokens	PERS	LOC	ORG	DATE	GHETTO	CAMP
cs	Czech	106392	1415	2627	359	741	**212**	**502**
de	German	**218570**	**2516**	**3592**	**871**	**950**	202	396
en	English	58 405	363	1015	225	287	52	77
fr	French	2273	3	39	8	4	0	5
hu	Hungarian	24686	157	304	148	97	2	114
nl	Dutch	1991	17	25	33	7	0	2
pl	Polish	18385	221	328	54	126	17	51
sk	Slovak	3550	30	158	11	21	0	0
yi	Yiddish	71506	629	1311	158	4	43	82
/	All	505758	5351	9399	1867	2237	528	1229

Table 2: EHRI-NER Dataset: tokens and entity classes distribution.

Entity	Example	Annotation
Person	Kurt Lichtenstern	B-PERS, I-PERS
Location	Moravská Ostrava	B-LOC, I-LOC
Organization	Pártfogó iroda	B-ORG, I-ORG
Date	1941 roku	B-DATE, I-DATE
Ghetto	getta łódzkiego	B-GHETTO, I-GHETTO
Camp	Auschwitz camp	B-CAMP, I-CAMP

Table 3: Entity types illustrated with examples and IOB tagging.

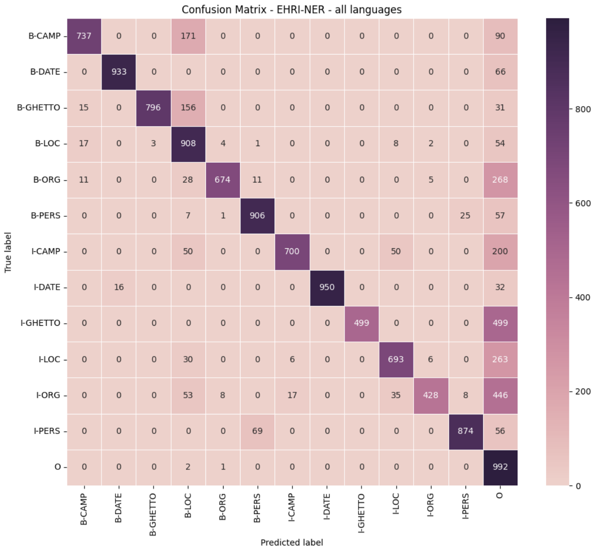
with reliably good performance across most if not all languages, suitable primarily as part of an editorial pipeline that streamlines the creation of new digital scholarly editions related to the Holocaust.

Experiment 2: To assess the cross-lingual capabilities of XLM-R in a low-resource setting, we fine-tuned it three more times—each time leaving out one language subset which was reserved for testing. Our chosen target languages were nl (experiment 2.1) and yi (experiment 2.2) as they represent some of the smallest subsets, while still containing enough examples for meaningful evaluation. For each target language, we fine-tuned XLM-R on every other subset split into train and validation (80% / 20%) and used the entire subset of the target language as the test set. This experiment sought to simulate a scenario where we would need to use our fine-tuned model to pre-annotate documents from a Holocaust domain but in a language not seen by our model during fine-tuning.

The fine-tuning processes were repeated three times for each experiment, we then computed the average of each of the three runs to obtain a reliable evaluation.

5.3. Evaluation

Experiment 1 yielded a consistent and satisfying overall performance across the validation and the test sets, with an overall F1 score of 81.3% for the former and 81.5% for the latter (Table 4), achieving higher scores compared to earlier EHRI NER work, surpassing Rodriguez et al.'s 2012 maximum total F1 score of 60% while additionally tagging domain-specific entities and exceeding the F1 score of 77% reported for the person tagger (de Leeuw et al., 2018). Domain-specific entities (Camp, Ghetto) are also consistently classified by the model. Only the Organization entity demonstrated poor F1 scores probably caused by the relatively low number of examples (1867 in total). This behavior has been previously observed by Rodriguez et al. (2012). The overall evaluations for cs, de, en, hu, pl, and yi test sets showed that the performance of the fine-tuned model is corollary to the number of examples in the training set. However, even though we see a



	B-CAMP	B-DATE	B-GHETTO	B-LOC	B-ORG	B-PERS	I-CAMP	I-DATE	I-GHETTO	I-LOC	I-ORG	I-PERS	O
B-CAMP	737	0	0	171	0	0	0	0	0	0	0	0	90
B-DATE	0	933	0	0	0	0	0	0	0	0	0	0	66
B-GHETTO	15	0	796	156	0	0	0	0	0	0	0	0	31
B-LOC	17	0	3	908	4	1	0	0	0	8	2	0	54
B-ORG	11	0	0	28	674	11	0	0	0	0	5	0	268
B-PERS	0	0	0	7	1	906	0	0	0	0	0	25	57
I-CAMP	0	0	0	50	0	0	700	0	0	50	0	0	200
I-DATE	0	16	0	0	0	0	0	950	0	0	0	0	32
I-GHETTO	0	0	0	0	0	0	0	0	499	0	0	0	499
I-LOC	0	0	0	30	0	0	6	0	0	693	6	0	263
I-ORG	0	0	0	53	8	0	17	0	0	35	428	8	446
I-PERS	0	0	0	0	0	69	0	0	0	0	0	874	56
O	0	0	0	2	1	0	0	0	0	0	0	0	992

Figure 1: Matrix confusion for predicted classes in the test set, when fine-tuning XLM-R on **all languages** (experiment 1, Section 5.2). The confusion matrix was normalized using a scaling factor of 1000.

decrease in the F1 scores depending on the size of the subset (minimum 73.2% overall F1 score for the hu test set), we still consider the performance of the fine-tuned model strong considering the relatively small size of some of the subsets.

The confusion matrix for predicted classes in the test set (Fig. 1) shows instances where the fine-tuned model occasionally misclassifies entities as non-entity tokens, I-GHETTO being the most confused entity. The fine-tuned model occasionally encounters challenges in extracting multi-tokens entities, such as I-CAMP, I-LOC, and I-ORG, which are sometimes confused with the beginning of an entity. Moreover, it tends to misclassify B-GHETTO and B-CAMP as B-LOC, which is not surprising given that they are semantically close and there are cases where even an expert would hesitate to pick a single label. Indeed, sometimes an entity such as the camp/ghetto "Theresienstadt" could be assigned any of these classes without introducing errors[14].

Overall, we argue that these scores are high enough to at least pre-annotate Holocaust-related textual documents when developing a new EHRI Online Edition or when wanting to enrich an archival description with access points that an archivist can verify. Additionally, as long as the new unseen texts to be fed into the model belong to a similar domain and period, we can assume that the scores will remain relatively consistent across all nine languages used for fine-tuning.

We released the fine-tuned XLM-R model on Hugging Face[15].

Experiment 2 revealed that we can leverage the cross-lingual capabilities of XLM-R depending to some extent on how much data it has seen about a specific language during its pre-training and on how many examples the training dataset has.

Experiment 2.1 showed unexpectedly high performance, about 94% overall F1 score, in one of the runs on the Dutch subset. However, it decreased in the second run to around 80% F1 score. After the third run, the overall F1 score of 84.6% proved that the fine-tuned model achieved satisfying performance, except for the classification of Organization entities (see Table 5), and despite not being evaluated on the Ghetto entity due to lack of examples. The confusion matrix shows that the fine-tuned model has trouble extracting multi-token entities, as noted in experiment 1 (Fig. 2).

Experiment 2.2 on Yiddish yielded poor performance, with an overall F1 score of 46.5% (Table 6). Only Person and Location entities showed an F1 score of above or equal to 50%. The Organization entity and the domain-specific entities Date, Camp, and Ghettos are all under 10% F1 score, the latter having an F1 score of 0. As depicted in Fig. 3, the model mainly misclassified entities as non-entity tokens, which is a common problem in NER (Luthra et al., 2023).

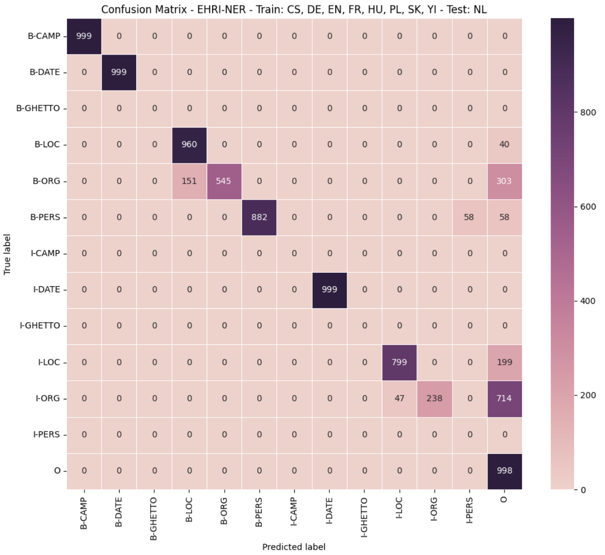


Figure 2: Evaluation of XLM-R on the **nl subset, when fine-tuned on all languages except nl**, by entity type (experiment 2.1, Section 5.2).

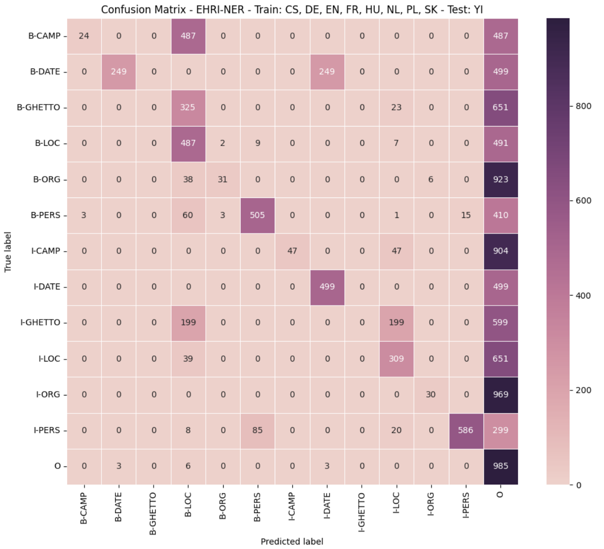


Figure 3: Matrix confusion for predicted classes in the **yi subset**, when fine-tuning XLM-R on **all languages except yi** (experiment 2.2, Section 5.2). The confusion matrix was normalized using a scaling factor of 1000.

The fluctuations in performance are probably related to the pre-training of XLM-R. As reported in Conneau et al. (2020), the model was pre-trained

[14] See more about the function of Theresienstadt here. Accessed 2/27/2024.

[15] See the EHRI-NER fine-tuned XLM-R model on Hugging Face.

	Validation Set				Test Set			
Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	**85**	90.3	87.5	/	83.8	**88.7**	**86.2**
Location	/	78.1	86	81.8	/	78.1	87.3	82.5
Organization	/	62.3	56.8	59.4	/	61.9	60.7	61.3
Date	/	81.5	**92.9**	**86.8**	/	81.1	90.3	85.4
Camp	/	76.4	68.7	72.3	/	73	72.7	72.8
Ghetto	/	75.2	75.2	75.2	/	**87.1**	80.7	83.7
Overall	98	78.9	83.9	81.3	98	78.6	84.7	81.5
Overall - CS test set	/	/	/	/	98.3	82.5	87.1	84.7
Overall - DE test set	/	/	/	/	98.6	78	86.6	82.1
Overall - EN test set	/	/	/	/	98	75.4	84.4	79.6
Overall - HU test set	/	/	/	/	98.5	71.9	74.6	73.2
Overall - PL test set	/	/	/	/	97.2	73.3	77.7	75.5
Overall - YI test set	/	/	/	/	98.5	75.6	78.8	77.2

Table 4: Evaluation of fine-tuned XLM-R on EHRI-NER on **all languages**, by entity type (experiment 1, Section 5.2), and specific overall evaluation on cs, de, en, hu, pl, and yi test sets. fr, nl, and sk test sets were omitted because of a lack of examples.

Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	100	96	97.9
Location	/	83.2	96	89
Organization	/	76.1	61.6	67.5
Date	/	100	100	100
Camp	/	100	100	100
Ghetto	/	/	/	/
Overall	98.7	86.4	82.9	84.6

Table 5: Evaluation of XLM-R on the **nl subset, when fine-tuned on all languages except nl**, by entity type (experiment 2.1, Section 5.2).

Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	68.9	53.1	59.9
Location	/	48.4	52.2	50
Organization	/	21.3	04.8	07.6
Date	/	00.7	41.6	01.3
Camp	/	28.4	02	03.7
Ghetto	/	0	0	0
Overall	96.6	47.2	46.2	46.5

Table 6: Evaluation of XLM-R on the **yi subset, when fine-tuned on all languages except yi**, by entity type (experiment 2.2, Section 5.2).

on 5025M tokens for Dutch, but merely 34M tokens for Yiddish. Therefore, we can hypothesize that the performance of XLM-R on the Yiddish subset is likely due to the limitations in its representation of this language after pre-training. This may have impacted the fine-tuning of the model and its cross-lingual capabilities for a token classification task on a small subset, such as the Yiddish subset, whereas the fine-tuning on the Dutch subset, despite being smaller, achieved a good performance. Other work on the zero-shot language transfer capabilities of multilingual Transformer models supports this hypothesis (Lauscher et al., 2020). Since the authors do not understand Yiddish, a comprehensive error analysis was not possible. However, it is worth noting that the challenges observed, as shown in experiment 1, can be mitigated when fine-tuning XLM-R on all subsets.

This experiment also confirms the hypothesis we made when discussing the lack of examples for the Organization entity and its consequence on the results in experiment 1.

6. Conclusion

In this work, we released EHRI-NER, a multilingual dataset for NER in Holocaust-related textual documents, built from the numerous TEI XML files made available across all EHRI Online Editions. We also evaluated the multilingual and cross-lingual capabilities of XLM-R by fine-tuning it on our dataset and proved that it can perform well when using relatively small domain-specific datasets. We also provided a baseline for future evaluations of NER systems on the dataset. Our future objective for the dataset is to include the translations mentioned in Subsection 4.2 while filtering them out from the train set. They indeed represent a sizable portion of data that would increase the number of examples in our dataset and could potentially lead to an increase in the fine-tuned model's performance.

For future work, we would like to experiment on multilingual named entity disambiguation, which would allow us to automatically link recognized entities with IDs in the EHRI vocabularies mentioned in the introduction (1). Another idea for future work could be to source similar annotated datasets and merge them with EHRI-NER. As a next step, we are planning to invite EHRI partners to evaluate our model qualitatively as part of their work and provide feedback. Based on that feedback, we can improve our model and deploy it as part of EHRI's cataloging and editorial pipelines. Another interesting course for future work would be to create a stable annotation typology for Holocaust documents with the help of experts. Finally, we hope to be able to provide a more complete baseline by experimenting with more multilingual Large Language Models (LLMs), including state-of-the-art LLMs for zero-shot and few-shot NER.

7. Acknowledgments

This research was conducted within the framework of the EHRI-3 project, which is funded by the European Commission under the call H2020-INFRAIA-2018–2020, with grant agreement ID 871111 and DOI 10.3030/871111.

We would like to acknowledge and thank Mike Bryant (King's College London), Lydia Nishimwe (Inria Paris, ALMAnaCH), and Armel Randy Zebaze (Inria Paris, ALMAnaCH) for their guidance and helpful discussions. The work described herein was made possible thanks to the previous work of the editors and contributors of the EHRI Online Editions, including the annotators, the people who produced digital facsimiles of the original archival material, and those who created the transcripts and translations. Finally, we would like to thank the open-source communities for generously sharing tools and know-how.

8. Bibliographical References

Tobias Blanke, Michael Bryant, Michal Frankl, Conny Kristel, Reto Speck, Veerle Vanden Daelen, and René Van Horik. 2017. The European Holocaust Research Infrastructure Portal. *Journal on Computing and Cultural Heritage*, 10(1):1–18.

Kirsten Strigel Carter, Abby Gondek, William Underwood, Teddy Randby, and Richard Marciano. 2022. Using AI and ML to optimize information discovery in under-utilized, Holocaust-related records. *AI & SOCIETY*.

Giovanni Colavizza, Maud Ehrmann, and Fabio Bortoluzzi. 2019. Index-Driven Digitization and Indexation of Historical Archives. *Frontiers in Digital Humanities*, 6.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ArXiv:1911.02116 [cs].

TEI Consortium. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Publisher: Zenodo.

Daan de Leeuw, Mike Bryant, Michal Frankl, Ivelina Nikolova, and Vladimir Alexiev. 2018. Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 58–66.

EHRI-Consortium. 2021. *Diplomatic Reports - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Maud Ehrmann, Damien Nouvel, and Sophie Rosset. 2016. Named Entity Resources - Overview and Outlook. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3349–3356, Portorož, Slovenia. European Language Resources Association (ELRA).

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022. Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 347–354, Cham. Springer International Publishing.

Sigal Arie Erez, Tobias Blanke, Mike Bryant, Kepa Rodriguez, Reto Speck, and Veerle Vanden Daelen. 2020. Record linking in the EHRI portal. *Records Management Journal*, 30(3):363–378. Num Pages: 16 Place: Bradford, United Kingdom Publisher: Emerald Group Publishing Limited.

Ignatius Ezeani, Paul Rayson, Ian Gregory, Erum Haris, Anthony Cohn, John Stell, Tim Cole, Joanna Taylor, David Bodenhamer, Neil Devadasan, Erik Steiner, Zephyr Frank, and Jackie Olson. 2023. Towards an Extensible Framework for Understanding Spatial Narratives. In *Proceedings of the 7th ACM SIGSPATIAL International*

Workshop on Geospatial Humanities, pages 1–10, Hamburg Germany. ACM.

Michal Frankl and Wolfgang Schellenbacher, editors. 2018. *BeGrenzte Flucht: Die österreichischen Flüchtlinge an der Grenze zur Tschechoslowakei im Krisenjahr 1938 - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Michal Frankl and Wolfgang Schellenbacher, editors. 2023. *Uzavřít hranice! - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Michal Frankl, Magdalena Sedlická, Hana Dauš, and Wolfgang Schellenbacher, editors. 2023. *Documentation Campaign - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI). Partner Institution: Masaryk Institute and Archives of the Czech Academy of Sciences.

Michal Frankl, Magdalena Sedlická, Wolfgang Schellenbacher, Daniela Bartáková, Michał Czajka, Jessica Green, Kat Hubschmann, Gábor Kádár, Yehudit Levin, Daphna Sehayek, Christine Schmidt, Zoltán Vagi, and Marta Wojas, editors. 2020. *Early Holocaust Testimony - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI), 2020.

Winfried Garscha, Claudia Kuretsidis-Haider, and Wolfgang Schellenbacher, editors. 2022. *VON WIEN INS NIRGENDWO: DIE NISKO-DEPORTATIONEN 1939*. EHRI Online Editions. Funded by: Nationalfonds der Republik Österreich für Opfer des Nationalsozialisten, Zukunftsfonds der Republik Österreich, Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, ahead-of-print(ahead-of-print).

W.J.B. Mattingly. 2021a. Holocaust Named Entity Recognition.

W.J.B. Mattingly. 2021b. wjbmattingly/holocaust_ner_lessons. Original-date: 2021-01-04T18:25:11Z.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of Transfer in Multilingual Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing Named Entity Recognition for Holocaust Testimonies through Pseudo Labelling and Transformer-based Models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pages 85–90, San Jose CA USA. ACM.

Ivelina Nikolova and Michael Levy. 2018. Using Named Entity Recognition to Enhance Access to a Museum Catalog – Document Blog.

Kepa J Rodriguez, Vladimir Alexiev, Laura Brazzo, Charles Riondet, Yael Gherman, and Reto Speck. 2016. EHRI-2 - D.11.2 Road Map Domain Vocabularies. Deliverable GA no. 654164. Issue: GA no. 654164.

Kepa Joseba Rodriguez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. pages 410–414.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. ArXiv:cs/0306050.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical Multilingual Language Models for Named Entity Recognition. In *CEUR Workshop Proceedings*, Bologna, Italy. arXiv. ArXiv:2205.15575 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced Meta-Learning for Cross-Lingual Named Entity Recognition with Minimal

Resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9274–9281. Number: 05.

Dates and places as points of attachment for memorial contents in the ISW corpus: 1938 as a turning point

Carolina Flinz, Simona Leonardi
University of Milan, University of Genoa
Piazza S. Alessandro 1 ° I–20122 Milano, Piazza S. Sabina 2 ° I-16124 Genova
carolina.flinz@unimi.it, simona.leonardi@unige.it

Abstract

Aim of the paper is the identification and subsequent analysis of crisis years in the narrative biographical interviews with German speaking Jews from the corpus ISW (*Emigrantendeutsch in Israel: Wiener in Jerusalem/ Migrant German in Israel: Viennese in Jerusalem*); also the possible “chronological landmarks” within a year will be tackled, investigating how a certain year – 1938 – represents in the life story of the narrators a turning point, as it clusters most traumatic events linked to the Shoah. The transcripts were analysed using the tool *Sketch Engine*. An alternation of corpus-driven and corpus-based steps characterizes this study, which uses a quantitative-qualitative approach (see Lemnitzer and Zinsmeister, 2015) and integrates also approaches from narrative analysis. The research questions that guide our investigation are as follows: Are there any special dates that recur as chronological landmarks of crisis situations (Leonardi 2023a)? Which are they? Do they recur in connection with special places? Which ones?

Keywords: Israelcorpus, Corpus-driven, Corpus-based, Places, Landmarks, Chronology

1. Introduction[1]

We aim at investigating whether in the narrative biographical interviews from the corpus ISW (*Emigrantendeutsch in Israel: Wiener in Jerusalem/ Migrant German in Israel: Viennese in Jerusalem*), part of the so-called *Israelkorpus*, certain years stand out as crisis years, as the contents recalled to memory and verbalised concern political and social upheavals that had serious consequences on the lives of the persons interviewed and led to serious personal crises. Furthermore, we look for possible “chronological landmarks” for these crisis years (which events are recalled?) in order to examine the associated spatial constellations and their time-place interrelationships.

The transcripts were analysed using the tool *Sketch Engine*, in particular its *Concordance* tool for the analyses of the KWICs (Keywords-in-Context) and the text parts. An alternation of corpus-driven and corpus-based steps characterizes this study, which uses a quantitative-qualitative approach: the data were searched and extracted automatically, but also analysed and interpreted (see Lemnitzer and Zinsmeister, 2015).

The research questions that guide our analyses are as follows: Are there any special dates that recur as chronological landmarks of crisis situations (Leonardi 2023a)? Which are they? Do they recur in connection with special places? which ones?

After introducing the main aspects regarding the investigated corpus (2), visualizing the features which are most relevant for our analysis, the principal results of the analysis will be discussed (3). The paper concludes with a summary and an outlook for further research (4).

2. The so-called ‘Israelcorpus’ as an archive of life-stories and as an atlas

The so-called *Israelcorpus* is the result of long-term interview projects conducted by the German linguist Anne Betten and collaborators from 1989 until 2019 with German speaking Jews mainly in Israel (see Betten, 2023 for a survey on genesis of the project and archiving of the interviews and related materials). The whole *Israelcorpus* consists of three related corpora presently archived at the *Archiv für Gesprochenes Deutsch* (‘Archive for spoken German’[2]), which belongs to the program area *Oral Corpora* in the pragmatics department of the *Leibniz Institut für Deutsche Sprache* (‘Leibniz Institute for the German language’).

After Anne Betten changed from the University of Eichstätt, Germany, to the University of Salzburg, Austria, the recordings from the core corpus IS – *Emigrantendeutsch in Israel* ‘Emigrant German in Israel’ (188 recordings, cf. the corpus description in the DGD *Datenbank für Gesprochenes Deutsch / Database for Spoken German*[3]) were subsequently supplemented with further recordings with former Austrians, which were mostly collected by students and staff of the Institute of German Studies at Salzburg University during an excursion to Israel in December 1998. These recordings (28) make up the corpus ISW – *Emigrantendeutsch in Israel: Wiener in Jerusalem*[4], on which our analysis is based. A third corpus, ISZ – *Zweite Generation deutschsprachiger Migranten in Israel* ‘Second generation German-

[1] The two authors have been written the paper jointly. In particular, Carolina Flinz is responsible for §3, and Simona Leonardi for § 2. Introduction (§ 1) and Conclusions (§ 4) were written jointly.

[2] https://agd.ids-mannheim.de/index_en.shtml

[3] PID = <http://hdl.handle.net/10932/00-0332-C3A7-393A-8A01-3>

[4] PID = <http://hdl.handle.net/10932/00-0332-C42A-423C-2401-D>

speaking Migrants in Israel'[5] comprehends 100 interviews with second-generation individuals, mostly children of the interviewees from the corpora IS and ISW.

The three corpora can be accessed via the platform *Datenbank für Gesprochenes Deutsch*[6], after a free, one-time registration. Most of the interviews were recorded on audiotapes, which were subsequently digitalized by the IDS – only the most recent interviews were recorded digitally on minidisc and later on iPhone/iPad. The audio interviews, which are freely accessible online via the DGD-platform for research purposes, are currently stored as WAV-files. Whereas many of the interviews from the IS-corpus are not yet fully transcribed, literal transcripts of all the narrative interviews from the ISW-corpus are available via the DGD-platform, both as PDF files and as aligned text-to-speech FLN-files.

While the initial project aimed at investigating language maintenance/shift and sociolinguistic issues (cf. Betten 1995; Betten & Du-nour 2000), it was soon clear that the life stories collected in the course of the project could be analysed from various research approaches. As a matter of fact, the interviews from the three corpora have been so far analysed from several perspectives[7], which include for example, corpus-analytical and mixed-methods studies (cf. e.g. Ruppenhofer, Rehbein and Flinz, 2020; Flinz and Ruppenhofer, 2021; Flinz and Leonardi 2023a; Flinz and Leonardi 2023b; Pellegrino, 2023), conversation-analytical, and narratological studies on language, acculturation and identity (cf., e.g., the contributions collected in Leonardi, Thüne and Betten, 2016, and in Leonardi et al., 2023), as well as oral history studies (see especially the publications by the historian Patrick Farges, e.g. Farges, 2018 and 2020).

As narrative biographical interviews, the recordings from the *Israelcorpus* are life stories (Rosenthal, 1995); the narrative structuring of the life story stresses those events that mark discontinuities (both negative and positive) by lending these turning points retrospectively a special significance as essential constituents of the plot, which is made up of various narrative threads (cf. e.g. Polkinghorne, 1998). Although the narration of traumatic and extreme experiences was not originally at the centre of the Israel project, which, as mentioned above, focussed on questions of language maintenance/shift, the interviews of the *Israelcorpus* often address crucial experiences that caused identity breaks and reorientations and that often correspond to traumatic experiences. These were mostly based on the drastic political and social changes of the 1930s (in some cases even earlier, see Betten, 1995; Leonardi, 2016), which were associated with anti-Semitic measures and attacks. Perhaps even more traumatic were the further consequences of such early anti-Semitic incidents – at best flight and emigration, at worst deportation and imprisonment in concentration camps. The latter affected the interviewees themselves relatively seldom (but see e.g. Leonardi, 2016; Koesters Gensini, 2023; Schwitalla, 2023 for analysis of lager testimonies), as most of them emigrated before 1939, however, some of those who were still able to emigrate had also been imprisoned in camps, mostly in the wake of the pogrom of November, 9th-10th, 1938 (so-called "Reichskristallnacht").

Furthermore, in the process of storytelling, the speakers do not narrate their own life story detached from other life stories, but usually intertwine it into a broader, intergenerational family history that spans generations. As a result, the migration routes of previous generations (e.g. Betten and Leonardi, 2023; Pellegrino 2023a; Pellegrino 2023b) as well as the flight routes (cf. Haßlauer 2016; Schwitalla, 2016) or – as far as is known – the various stages of the expulsion or deportation and murder of other family members – including parents, grandparents, and siblings – who were unable to emigrate are often outlined (see e.g. Betten, 2008; Thüne 2016; Betten and Leonardi, 2023). In a nutshell, the so-called *Israel corpus* can be seen not only as an archive of life stories and of shoah testimonies, but also as an atlas of intergenerational Jewish migration routes.

2.1 The corpus ISW

The main features of the corpus ISW are summarized in the following table (1):

Recordings	28
Recordings (hrs)	51h 22m 26s 68
Genres	27 narrative interviews 1 reading
Speakers	24 13 men 11 women
Language	German
Transcriptions	27 (all the narrative interviews)
Collection period	1998-2011
Place of collection	Jerusalem, IL (25) Salzburg, A (2) Wien, A (1)
Interviewees' birth year	1915–1929

Table 1: Main features of the corpus ISW

Three recordings are interviews with couples (Mirjam and Aaron Alexander[8], Shaul and Hanna Baumann[9],

[5] PID = <http:// hdl.handle.net/10932/00-0332-C453-CEDC-B601-2>.

[6] <https://dgd.ids-mannheim.de/>.

[7] For a continually updated bibliography on the whole *Israelkorpus* cf. <https://www.zotero.org/groups/2219390/israelkorpus/library>.

[8] ISW_E_00001 (PID = http://hdl.handle.net/10932/00-0332-C42A-C81C-2701-A).

[9] ISW_E_00002 (PID = http://hdl.handle.net/10932/00-0332-C42C-B15C-2A01-B).

and Chava and Jeshajahu Karniel[10])[11]. Three speakers gave several interviews during the project: there are three recordings featuring Gerda Hoffer, the first is a narrative biographical interview like the other interviews of the corpus (1998)[12], the second a reading[13], also recorded in 1998, on the occasion of a public meeting during the excursion of the Institute of German Studies from the University of Salzburg, the third is an interview gathered by Michaela Metz in in 2010[14] in the course of a special project on childhood memories (on Metz' collection see Häußinger 2023). The journalist Ari Rath was interviewed five times, four times by Anne Betten between 1998 and 2000[15], and one time by Michaela Metz in 2010[16]; Jeanette Goldstein, finally, was interviewed during the 1998-excursion[17] and in 2010 by Michaela Metz[18].

The year of birth of the interviewees is shown in Figure 1, which reveals also that the corpus ISW is generationally quite homogeneous, since most of the speakers (17 out of 24) were born between 1920 and 1929.

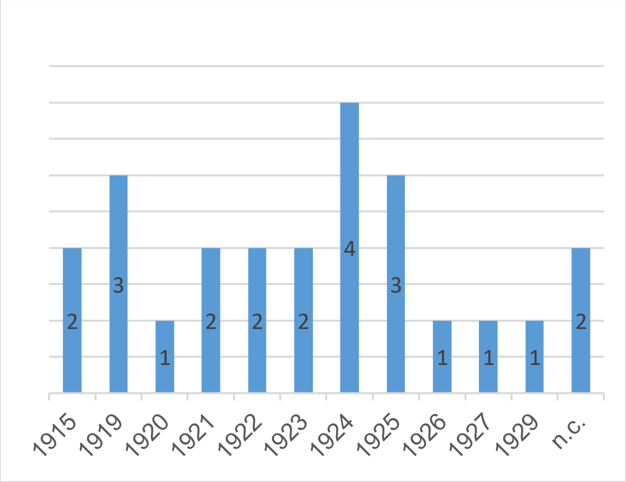


Figure 1: Interviewees' birth years

Figure (2) illustrates the year of the emigration from Vienna[19], and the year of immigration to Mandate Palestine / Israel – two datasets are shown, as the year of emigration in many cases does not coincide with the immigration year to Mandate Palestine / Israel; 4 persons immigrated to Mandate Palestine / Israel only after WW2. The year 1939 is both the year which features the highest number of emigrations (11) and the highest number of immigrations (12); this cluster lends the year 1939 a special relevance. The only other significant year is 1938, but only in relation to the value 'emigration' (8), whereas the number of immigrations stands at merely 2.

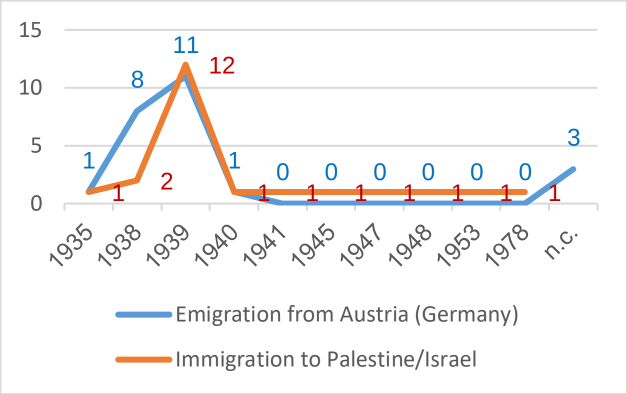


Figure 2: Emigration/Immigration year

Cross-referencing the data from Figure (1) with those from Figure (2) highlights that most speakers were minors at the time of emigration. Many of them could escape (or limit) Nazi persecutions thanks to special actions or programmes, as shown in Figure (3):

The Youth Aliyah (*Jugendalija*) was a Zionist aid organisation set up by Recha Freier and Eva Michaelis-Stern[20] from 1933; it organised the group emigration of unaccompanied Jewish minors to Palestine. A small number of secondary school students was awarded a 'student certificate' to Mandate Palestine (Löw, 2020).

[10] ISW_E_00016 (http://hdl.handle.net/10932/00-0332-C435-8C1C-5301-A)

[11] Other couples gave separate interviews, see Paul Rudolf Beer (ISW-_E_00003, http://hdl.handle.net/10932/00-0332-C42D-110C-2D01-A) and Shoshana Beer (ISW-_E_00004, http://hdl.handle.net/10932/00-0332-C42E-424C-3001-9), as well as Erich Goldstein (ISW-_E_00008, http://hdl.handle.net/10932/00-0332-C42F-EEBC-3C01-7) and Jeanette Goldstein (ISW-_E_00009, http://hdl.handle.net/10932/00-0332-C430-7A1C-3F01-A). In these cases, both interviewees were from Vienna, while in the couple-interviews only one person was Viennese (i.e. Mirjam Alexander, Shaul Baumann, and Jeshajahu Karniel).

[12] ISW_E_00011 (http://hdl.handle.net/10932/00-0332-C432-8B6C-4501-2).

[13] ISW_E_00012 (http://hdl.handle.net/10932/00-0332-C433-979C-4701-A).

[14] ISW_E_00027 (http://hdl.handle.net/10932/00-0332-C43A-3EAC-6E01-5).

[15] ISW_E_00019 (http://hdl.handle.net/10932/00-0332-C437-34DC-5B01-D); ISW_E_00020 (http://hdl.handle.net/10932/00-0332-C438-52EC-5D01-B), ISW_E_00021 (http://hdl.handle.net/10932/00-0332-C438-B62C-5F01-7), ISW_E_00022 (http://hdl.handle.net/10932/00-0332-C438-EACC-6101-4).

[16] ISW_E_00028 (http://hdl.handle.net/10932/00-0332-C43A-737C-7001-1).

[17] ISW-_E_00009 (http://hdl.handle.net/10932/00-0332-C430-7A1C-3F01-A).

[18] ISW-_E_00026 (http://hdl.handle.net/10932/00-0332-C43A-0A3C-6C01-7).

[19] Or from Germany, which applies to spouses in the couple-interviews.

[20] Eva Michaelis-Stern's interview with Anne Betten, where also the Youth Aliyah actions are thematized, is archived in the corpus IS (IS_E_00087); cf. Michaelis and Stern Michaelis, 1989.

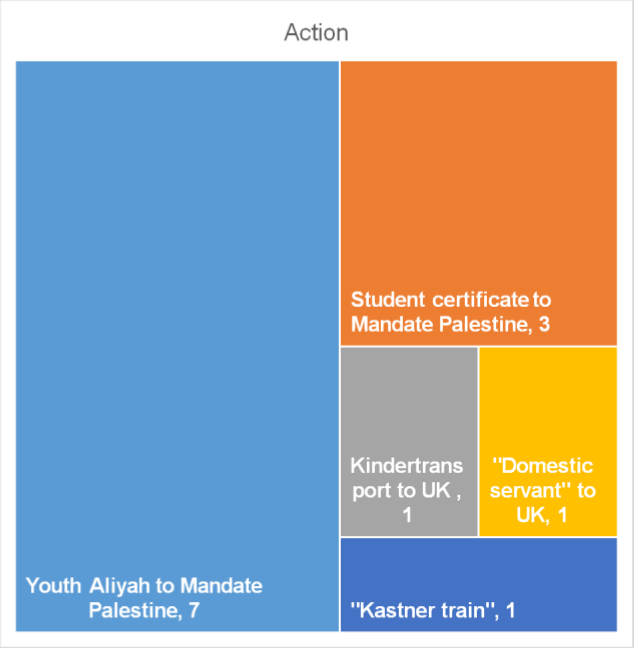


Figure 3: Rescue actions and programmes

While the Kindertransport was the most famous rescue action of persecuted children from Germany, Austria and other Nazi-annexed areas to the UK, older teens and young adults could emigrate to the UK on so-called ‘domestic permits’, in order to work as domestic servant[21]. “Kastner train” is the name given to an action which rescued over 1,600 Jews – 273 children – from Hungary during World War 2. The action was organized by the Hungarian-Jewish lawyer and journalist Rudolf Kastner (Rezső Kasztner), who negotiated with the SS officer Adolf Eichmann (Bauer 1994).

The corpus ISW gives thus insight to several actions which helped Jewish children and teens flee from Austria, which corresponds to different settings and migration routes; their narratives picture them.

3. Years in the corpus

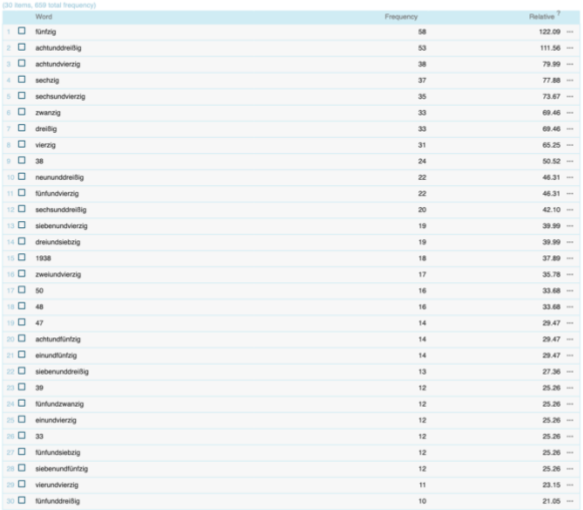
In order to identify the crisis years in the corpus, we used an approach characterised by a number of steps, both quantitative and qualitative. Firstly, the interviews from the ISW corpus (27) were uploaded into another tool, *Sketch Engine*. We thus rebuilt a corpus with the following features:

We subsequently extracted all the numbers[22] in the corpus (6,820, or 14,355.66 pmt). The relevant occurrences were then filtered by means of the feature *Frequency* according to the word forms of the numbers: 522 items were identified.

Tokens	Words
475.074	367.701

Table 2: Tokens and words in the corpus ISW

By means of a careful qualitative analysis, all irrelevant items (*zwei*, *5*, *CM* etc.) were removed until the number of items was reduced to 30 (659 occurrences).



(30 items, 659 total frequency)

		Word	Frequency	Relative ?
1	□	fünfzig	58	122.09 ···
2	□	achtunddreißig	53	111.56 ···
3	□	achtundvierzig	38	79.99 ···
4	□	sechzig	37	77.88 ···
5	□	sechsundvierzig	35	73.67 ···
6	□	zwanzig	33	69.46 ···
7	□	dreißig	33	69.46 ···
8	□	vierzig	31	65.25 ···
9	□	38	24	50.52 ···
10	□	neununddreißig	22	46.31 ···
11	□	fünfundvierzig	22	46.31 ···
12	□	sechsunddreißig	20	42.10 ···
13	□	siebenundvierzig	19	39.99 ···
14	□	dreiundsiebzig	19	39.99 ···
15	□	1938	18	37.89 ···
16	□	zweiundvierzig	17	35.78 ···
17	□	50	16	33.68 ···
18	□	48	16	33.68 ···
19	□	47	14	29.47 ···
20	□	achtundfünfzig	14	29.47 ···
21	□	einundfünfzig	14	29.47 ···
22	□	siebenunddreißig	13	27.36 ···
23	□	39	12	25.26 ···
24	□	fünfundzwanzig	12	25.26 ···
25	□	einundvierzig	12	25.26 ···
26	□	33	12	25.26 ···
27	□	fünfundsiebzig	12	25.26 ···
28	□	siebenundfünfzig	12	25.26 ···
29	□	vierundvierzig	11	23.15 ···
30	□	fünfunddreißig	10	21.05 ···

Figure 4: Items/occurrences (Screenshot from the automatic extraction with Sketch Engine)[23]

The occurrences were then downloaded and analysed in detail. The not relevant items were excluded.

In the following we present the main results of the analysis of the items (= the years) with a frequency > 10. There corresponds to 23 items (23 years), whose occurrences are represented in Figure 5:

The graphs in Figure 5 shows that the most frequently occurring year is *1938* (transcribed in the forms *achtundreißig*, *38*, *1938*) and it is also a very interesting year, as it clearly marks a contrast between the before and the after, so that “the break” is evident (1, 2)

[21] Cf. <https://www.cityoflondon.gov.uk/things-to-do/history-and-heritage/london-metropolitan-archives/collections/jewish-refugees-and-the-british-nursing-profession>.

[22] The extraction process was carried out using the *Sketch Engine Concordance* tool. By means of the CQL search option [tag="CARD"], all numbers in the corpus were identified.

[23] The screenshot from *Sketch Engine* shows that year dates in the transcriptions are written in various forms: in figures, both in full and abbreviated, and in words. For our analysis, all graphic variants were taken into account.

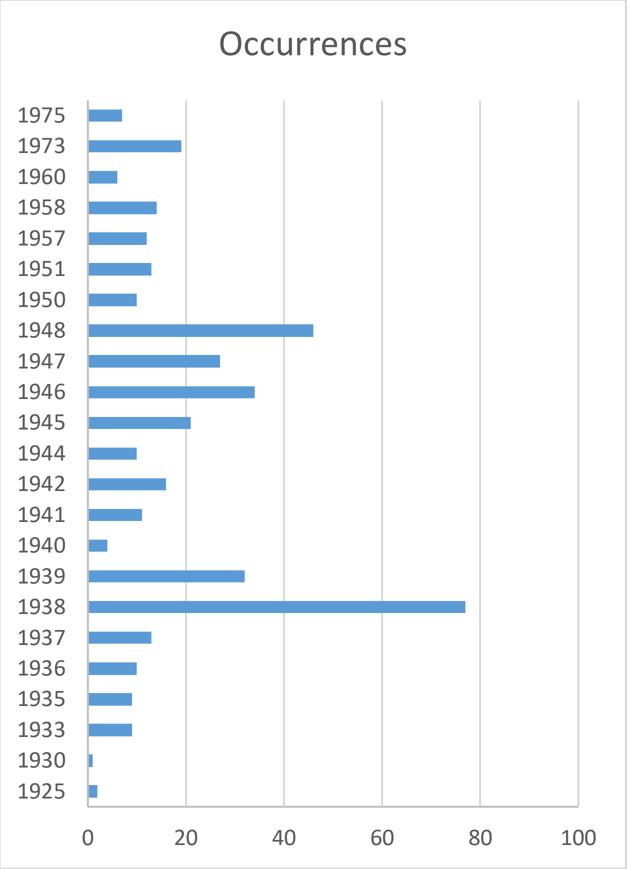


Figure 5: Years mentioned / occurrences

(1) SB: […] Der Hass war tief, tief drinnen und der kam dann raus mit + (h) + dem Anschluss.
IR: Das war achtunddreißig?
SB: Das war achtunddreißig. (ISW_00004) (The hatred was deep, deep inside and it came out with + (h) + the connection. IR: That was thirty-eight? SB: That was thirty-eight)

(2) […] aber das wirklich das war das erste Mal, + dass da son Schlag in die Familie kam.
VL: Das war achtunddreißig.
MH: Das war achtunddreißig (ISW_00010) (But that was really the first time + that there was a blow to the family. VL: That was thirty-eight. MH: That was thirty-eight)

1938 marks the turning point. The previous situation is recalled as almost "idyllic" compared to the aftermath (3), although acts of antisemitism did occur

(3) der Gedanke war nicht so, man hat immer in Österreich hat man / + Das war doch dreiunddreißig bis achtunddreißig, sozusagen fünf Jahre hat man ge/ gelebt in einem / in einer Idylle doch sozusagen. (ISW_00015) (The idea wasn't that, you always lived in Austria / + That was thirty-three to thirty-eight, five years, so to speak, you lived in an / idyll, so to speak)

(4) […] man wurde schon mal angepöbelt, ja ja, das das kam schon + kam schon vor. ++
IK: Schon vor achtunddreißig, oder?
MH: Vor achtunddreißig, ja, ja, + das passierte schon. Und man war die die die Eltern waren sehr bewusst, man soll nicht laut sein, man soll nicht auffallen. (ISW_00010) (you have been mobbed before, yes yes, that has happened + has happened before. ++ IK: Even before thirty-eight, right? MH: Before thirty-eight, yes, yes, + that did happen. And the parents were very aware that you shouldn't be loud, you shouldn't stand out.)

The critical events which took place in 1938 are often thematised, mentioning not only the name of the country (*Österreich*/Austria) and the capital (*Wien*/Vienna) – whence most of the speakers came from, but also places within it , such as the school, the classroom, streets, etc., i.e. key places where the interviewed people or their family members experienced antisemitic harassment, discrimination, and even violence, they were e.g. arrested and beaten. As a result, these places have taken on a symbolic value, because a linguistic and historical community endows them socially with networks of associations and meanings. These are then revealed in the verbalisation of memory content, as speakers embed the into the narration also the topological and chronological relationships between narrative figures on the one hand and places as well as times on the other. The latter thus become points of reference, or topological and chronological landmarks (s. Herman, 2001; Leonardi, 2023b).

As a matter of fact, connected to the mention of the year 1938 occur:

- 'places of time' (Brambilla, Flinz, and Luppi, 2023), such as Gestapo prisons, Gestapo headquarters, concentration camps;

- means of transport (Flinz, and Ruppenhofer, 2021), which enabled the rescue of the speakers, such as the train;

- organisations, such as the Youth Aliyah.

The fact that 1938 is a crisis year is also confirmed by the fact that it is a taboo year, which some speaker prefers not to talk about:

(5) Naja, 1938 erzähle ich lieber nicht, das war nicht so schön. Ich würde sagen, vielleicht sind es die zwei schrecklichsten Eindrücke ISW_00011). (Well, I'd rather not tell you about 1938, that wasn't so nice. I'd say they were perhaps the two most terrible impressions)

4. Conclusion and Further Research

The quantitative-qualitative analysis shows clearly that the year 1938 a (negative) turning point in the life stories collected in the corpus ISW is. It should be kept in mind that most of the speakers from the corpus were originally from Austria, and that in 1938, on March, 12th, Austria was annexed to Nazi-Germany (so-called *Anschluss*). This was a most traumatic event for most of the Austrian Jews. Furthermore, on November, 9th (in the night to November, 10th), 1938 Nazi Party's paramilitary forces carried out a pogrom against Jews and Jewish institutions all over the *Reich* (so-called *Reichskristallnacht*[24]). This was also a major trauma[25].

Further research is needed to investigate which places or spaces are associated with the year 1938 (which 'places of the time'? which 'transport mean'?) and which finer chronological landmarks can be identified (whether *'Anschluss'*, or *Reichskristallnacht*, or emigration, etc.). A similar research carried out with the interviews from the corpus IS, where most of the speakers were originally from Germany – and not from Austria as in the ISW – could reveal whether for German speakers the chronological landmarks are the same or whether differences exist. An additional point which requires further investigations regards finally how the various numbers were pronounced, which could be fruitfully explored thanks to the possibilities of modern speech and emotion recognition (SER).

[24] Both terms are currently dispreferred, as they are considered Nazi-biased – as a matter of fact the interviewees use them, as they didn't follow the discussion about them, which took place in Germany and in Austria since 1988.

[25] In the aftermath of the November pogrom the British government agreed to accept additional Jewish refugees – but only as unaccompanied minors – the *Kindertrasport* rescue action started within this framework.

5. Bibliographical References

Bauer, Y. (1994). Jews for Sale? *Nazi-Jewish Negotiations, 1933–1945*. Yale: Yale University Press.

Betten, A. (Ed.) (1995). *Sprachbewahrung nach der Emigration – Das Deutsch der 20er Jahre in Israel.* Vol. 1. *Transkripte und Tondokumente* Tübingen: Niemeyer.

Betten, A. and Du-nour, M. (Eds) (2000). *Sprachbewahrung nach der Emigration – Das Deutsch der 20er Jahre in Israel.* Vol. 2. *Analysen und Dokumente*. Tübingen: Niemeyer.

Betten, A. (2008). Schöne und schwere Gedanken an Lublin. In M. Stebler (Ed.), *Nicht nur ein Grund für Dankbarkeit. Festschrift für Jerzy Jeszke*, Lublin: Wydawnictwo werset, pp. 11–24.

Betten, A. and Leonardi, S. (2023). Das Interviewkorpus “Sprachbewahrung nach der Emigration / Emigrantendeutsch in Israel”: Ein sprach- und kulturwissenschaftliches Archiv des deutschsprachigen Judentums im 20. Jahrhundert. *Tsafon* Hors-série no 11 (Special issue *Archives de la Diaspora / Diaspora des Archives. Penser la mémoire de la dispersion à partir de l’espace germanophone*): 233–258.

Brambilla, M., Flinz, C., and Luppi, R. (2023). ‘Orte der Zeit’ im Korpus ISW. Eine linguistische Analyse des Zusammenspiels von Orten, Emotionen und Erinnerungen. *germanica;*, 33 (Special issue *Erzählte Chronotopoi: Orte und Erinnerung in Zeitzeugeninterviews und berichten zu erzwungener Migration im 20. Jahrhundert*): 253–278 <http://www.serena.unina.it/index.php/aiongerm/article/view/10745/11032>.

Farges, P. (2018). Pioneers, Losers, White Collars: Narratives of Masculinity Among German-Speaking Jews in Palestine/Israel. *Remembrance and Research. The journal of the Israel Oral History Association. ILOHA* 2: 33–50.

Farges, P. (2020). *Le Muscle et l’Esprit masculinités germano-juives dans la post-migration: le cas des yekkes en Palestine, Israël après 1933*. Bruxelles et al.: Peter Lang.

Flinz, C. and Leonardi S. (2023a). ‘Luoghi del tempo’, Ricordi ed emozioni nelle interviste del corpus Emigrantendeutsch in Israel. *Echo* 5/2023: 4–25. <https://ojs.cimedoc.uniba.it/index.php/eco/article/view/1874>.

Flinz, C. and Leonardi S. (2023b). Luoghi di transito, ricordi ed emozioni nel corpus Emigrantendeutsch in Israel: Wiener in Jerusalem (ISW). In: M. Castagneto/ M. Ravetto (a cura di), *La Comunicazione Parlata / Spoken Communication, Pubblicazioni del GSCP*, vol. 3, Roma, Aracne, pp. 589-615.

Flinz, C. and Ruppenhofer, J. (2021). Auf dem Weg zu einer Kartographie: automatische und manuelle Analysen am Beispiel des Korpus ISW. *Sprachreport* 1: 44–50.

Haßlauer, S. (2016). Fluchterlebnisse und ihr sprachlicher Ausdruck. Untersuchungen zu Agency, Emotionen und Perspektivierung in den Erzählungen zweier jüdischer Emigrantinnen. In S. Leonardi, E.-M. Thüne and A. Betten (Eds.), *Emotionsausdruck und Erzählstrategien in narrativen Interviews: Analysen zu Gesprächs-aufnahmen mit jüdischen Emigranten*, Würzburg: Königshausen & Neumann, pp. 201–230.

Herman, D. (2001). Spatial Reference in Narrative Domains. *Text*, 21(4): 515–541.

Koesters Gensini, S. (2023). “Nur ich bin im Lager [...] I’m... I’m a survivor”. Versprachlichte Erinnerungen an Lager im Israelkorpus. *Annali Sezione Germanica* 33 (Special issue *Erzählte Chronotopoi: Orte und Erinnerung in Zeitzeugeninterviews und berichten zu erzwungener Migration im 20. Jahrhundert*): 279–298.

Lemnitzer, L., and Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Leonardi, S. (2023a). Erinnerte Chronotopoi: Rekonstruktion von Krisensituationen in Erzählungen. *germanica;*, 33 (Special issue *Erzählte Chronotopoi: Orte und Erinnerung in Zeitzeugeninterviews und berichten zu erzwungener Migration im 20. Jahrhundert*): 121–150 <http://www.serena.unina.it/index.php/aiongerm/article/view/10739/11027>.

Leonardi, S. (2023b): Orte in der Versprachlichung von Gedächtnisinhalten. In S. Leonardi et al. (Eds.), *Orte und Erinnerung. Eine Kartografie des Israel-korpus*. Roma: Istituto Italiano di Studi Germanici, pp. 91–109.

Leonardi, S., Thüne, E.-M. and Betten, A. (Eds.) (2016). *Emotionsausdruck und Erzählstrategien in narrativen Interviews: Analysen zu Gesprächs-aufnahmen mit jüdischen Emigranten*. Würzburg: Königshausen & Neumann.

Leonardi, S. et al. (Eds.) (2023). *Orte und Erinnerung. Eine Kartografie des Israelkorpus*. Roma: Istituto Italiano di Studi Germanici.

Löw, A. (2020). German Reich and Protectorate of Bohemia and Moravia September 1939–September 1941. Berlin / Boston: De Gruyter Oldenbourg.

Michaelis, D. and Michaelis-Stern, E. (1989). *Emissaries in Wartime London 1938–45*. Jerusalem: Hamaatik Press.

Pellegrino, R. (2023a). Erinnerte Orte und Sprachen in den narrativen Interviews des sog. Israelkorpus: eine transgenerationale Perspektive auf die Erzählungen von Sprecher_innen der ersten Generation. *TRANS. Internet-Zeitschrift für Kulturwissenschaften | Internet journal for cultural studies | Revue électronique de recherches sur la culture* 27 <https://www.inst.at/trans/27/erinnerte-orte-und-sprachen/>.

Pellegrino, R. (2023b). Familienchronotopoi im Israelkorpus: Orte und Sprachen bei Sprecher_innen österreichischer Herkunft und ihren Familien. *Germanica;*, 33 (Special issue *Erzählte Chronotopoi: Orte und Erinnerung in Zeitzeugeninterviews und berichten zu erzwungener Migration im 20. Jahrhundert*): 177–208.

Polkinghorne, D. E. (1998). Narrative Psychologie und Geschichtsbewußtsein. Beziehungen und Perspektiven. In J. Straub (Ed.), *Die psychologische Konstruktion von Zeit und Geschichte. Erinnerung, Geschichte, Identität 1*, Frankfurt/M.: Suhrkamp, pp. 12–45.

Rosenthal, G. (1995). *Erlebte und erzählte Lebensgeschichte. Gestalt und Struktur* *biographischer Selbstbeschreibungen*. Frankfurt/M.: Campus.

Ruppenhofer, J., Rehbein, I., and Flinz, C. (2020). Fine-grained Named Entity Annotations for German Biographic Interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020),* Marseille, France,11–16 May 2020, European Language Resource Association (ELRA), pp. 4605–4614.

Schwitalla, J. (2016). Narrative Formen von Fluchterzählungen deutschsprachiger emigrierter Juden in der Nazizeit. In S. Leonardi, E.-M. Thüne, & A. Betten (Eds.), *Emotionsausdruck und Erzählstrategien in narrativen Interviews: Analysen zu Gesprächsaufnahmen mit jüdischen Emigranten*. Würzburg: Königshausen & Neumann, pp. 171–199.

6. Language Resource References

ISW = “Emigrantendeutsch in Israel: Wiener in Jerusalem”, AGD (Archiv für gesprochenes Deutsch), DGD, PID <http://hdl.handle.net/10932/00-0332-C42A-423C-2401-D>.

Creating a Typology of Places to Annotate Holocaust Testimonies Through Machine Learning

The paper was not included in the proceedings at the request of the authors, but they presented it at the workshop.

Speech Technology Services for Oral History Research

Christoph Draxler[1], Henk van den Heuvel[2], Arjan van Hessen[3], Pavel Ircing[4], Jan Lehečka[4]
[1]BAS / Ludwig Maximilian Universität, München, [2]Radboud University,
[3]University of Twente, [4]University of West Bohemia, Pilsen
draxler@phonetik.uni-muenchen.de, henk.vandenheuvel@ru.nl, a.j.vanhessen@utwente.nl,
{ircing,lehecka}@kky.zcu.cz

Abstract

Oral history is about oral sources of witnesses and commentors on historical events. Speech technology is an important instrument to process such recordings in order to obtain transcription and further enhancements to structure the oral account In this contribution we address the transcription portal and the webservices associated with speech processing at BAS, speech solutions developed at LINDAT, how to do it yourself with Whisper, remaining challenges, and future developments.

Keywords: Speech technology, workflows, automatic transcription, NLP

1. Introduction

Oral history testimonies rely first of all on the audio and/or video capture of the recorded material, typically in the form of interviews. Here, the first challenge is converting the audio signal into a readable text adequately reflecting the spoken word. Automatic Speech Recognition (ASR) has been employed since around three decades to obtain initial transcriptions of oral history interviews. The output text typically calls for extensive manual correction often equalling or exceeding (!) an effort equivalent to starting with manual transcription from scratch (Gref, 2022), especially if recordings are characterized by overlapping and/or dialectal speakers, background noises or mediocre recording quality.

The impressive performance of large AI-based speech models, especially their robustness to noise, the large range of supported languages, and the option to adapt them to additional languages with relatively little extra training, is these days greatly facilitating the generation of adequate transcripts in research areas such as oral history where spoken language is a major source of information.

However, depending on the researcher's needs there remain (other) challenges such as appropriate speaker attribution, output of more fine-grained speech events such as hesitation sounds (*uh*), stutters, word truncations, etc.

We have a longstanding track record in speech technological support for oral history research (see also Scagliola et al., 2020) in CLARIN ERIC[1] where we have initiated a resource family page for oral history corpora[2] and a transcription portal for oral history recordings[3] and speech processing facilities at LINDAT[4] (the Czech node of CLARIN ERIC). In this contribution we will address a number of speech technology tools and solutions in these contexts, remaining challenges and future developments.

More specifically, we will address the transcription portal and the webservices associated with speech processing at BAS (section 2), speech solutions and beyond developed at LINDAT (section 3), do it yourself with Whisper (section 4), remaining challenges (section 5) and future developments (section 6).

2. Webservices at BAS

The Bavarian Archive for Speech Signals (BAS) provides a large number of multilingual speech processing web services for academic users:

https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

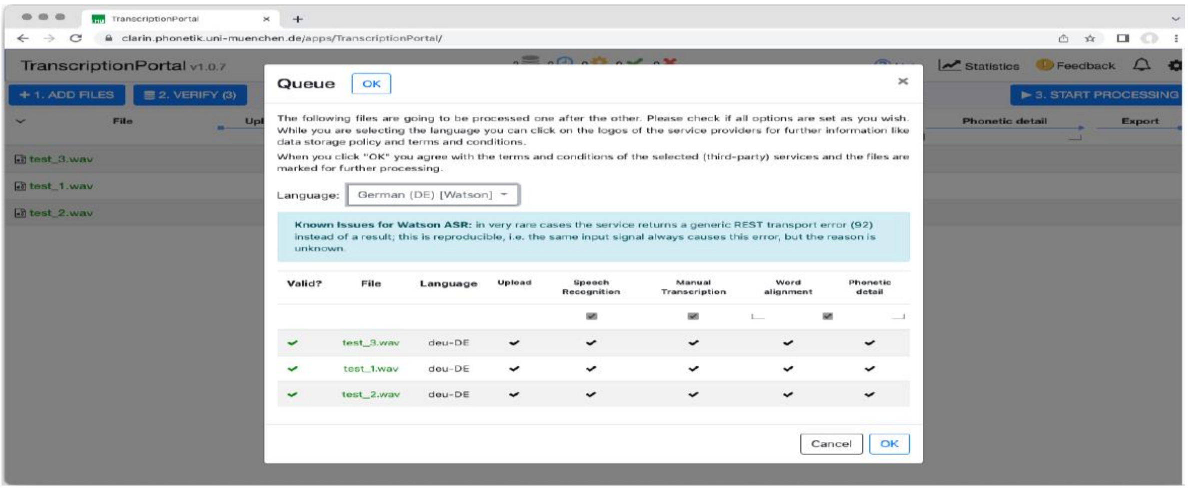
The services support 40+ languages, plus a language-independent mode based on phonemic transcripts. The list of available languages is accessible via drop down menus on the web page.

The following services may be of particular interest to Oral History scholars. They can be used without authentication.

- ChannelSeparator separates the individual channels of a stereo recording and retains in each channel only the voice of the dominant speaker.
- G2P (Grapheme to phoneme) converts an orthographic text to its phonemic representation. The service allows a customized specification of pronunciation rules for vernacular language, dialects, and common coarticulation phenomena (e. g. 'haben wir' (*we have*) → /hamva/ or /hama/ in German). These rules improve the performance of automatic word alignment.
- MAUS (Munich automatic segmentation) aligns an orthographic transcript in one of the available languages and regional variants) with the audio signal. The performance of MAUS depends on

[1] https://www.clarin.eu/content/clarin-nutshell
[2] https://www.clarin.eu/resource-families/oral-history-corpora
[3] https://speechandtech.eu/transcription-portal
[4] See https://lindat.cz/en/services

The First Workshop on Holocaust Testimonies as Language Resources @LREC-COLING-2024, pages 38–43
20-25 May, 2024.



The following files are going to be processed one after the other. Please check if all options are set as you wish. While you are selecting the language you can click on the logos of the service providers for further information like data storage policy and terms and conditions.

When you click "OK" you agree with the terms and conditions of the selected (third-party) services and the files are marked for further processing.

Language: German (DE) [Watson]

Known Issues for Watson ASR: in very rare cases the service returns a generic REST transport error (92) instead of a result; this is reproducible, i.e. the same input signal always causes this error, but the reason is unknown.

Valid?	File	Language	Upload	Speech Recognition	Manual Transcription	Word alignment	Phonetic detail
				☑	☑	☑	
✓	test_3.wav	deu-DE	✓	✓	✓	✓	✓
✓	test_1.wav	deu-DE	✓	✓	✓	✓	✓
✓	test_2.wav	deu-DE	✓	✓	✓	✓	✓

Figure 1: Selection of ASR language and processing steps in the transcription portal

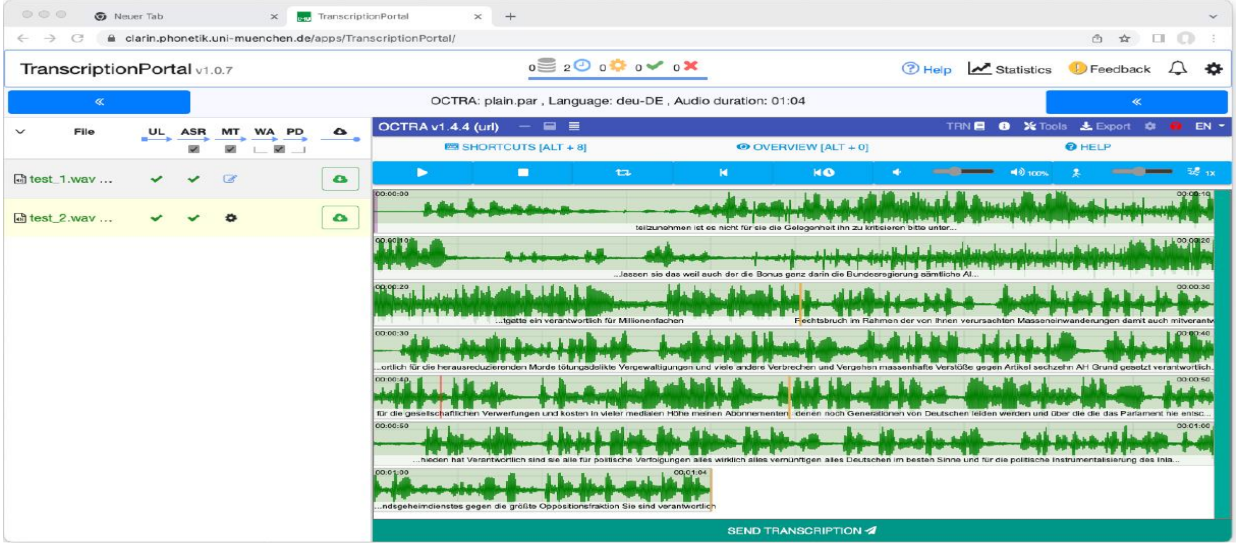
the language, the audio quality and the number of speakers; for German monologue recordings, it reaches 95% of the performance of human transcribers (Kipp et al. 1997).

- Octra is a graphical editor for orthographic transcription (Draxler & Pömp, 2022). It provides different views of the signal and the associated transcript, supports many input and output formats, and may split a recording into fragments to focus on relevant parts and/or distribute the workload. See further below for more details.
- Anonymizer automatically replaces the signal fragment corresponding to a given text by a beep, so that private information is effectively removed from both the recording and the transcript.

Automatic speech recognition is another web service. In contrast to the other services, it requires authentication as a member of academia, e.g. via the credentials necessary to login to a recognized academic institution (as recognized by CLARIN). The actual speech recognition is performed by external third-party providers, both academic and commercial, and thus may not be available if the privacy guidelines for a given recording or project do not allow data exchange with such providers.

Each service comes with a number of obligatory and optional parameters, e.g. the supported languages or file types, or output encodings and alphabets. and each service requires up- and downloading of data.

To reduce the number of file uploads and downloads, preconfigured service pipelines are available. This enables non-technical users to perform complex speech processing tasks without having to worry about low-level details. For example, the pipeline G2P→Chunker→MAUS→Anonymizer takes as input pairs of long audio files with an orthographic transcript plus a stop list of words, and returns a time-aligned transcript with the words from the stop list masked by a beep in the signal and a special symbol in the



OCTRA: plain.par , Language: deu-DE , Audio duration: 01:04

teilzunehmen ist es nicht für sie die Gelegenheit ihn zu kritisieren bitte unter...

...lassen sie das weil auch der die Bonus ganz darin die Bundesregierung sämtliche Al...

...tgette ein verantwortlich für Millionenfachen

Rechtsbruch im Rahmen der von Ihnen verursachten Masseneinwanderungen damit auch mitverantw

...ortlich für die herausreduzierenden Morde tötungsdelikte Vergewaltigungen und viele andere Verbrechen und Vergehen massenhafte Verstöße gegen Artikel sechzehn AH Grund gesetzt verantwortlich...

für die gesellschaftlichen Verwerfungen und kosten in vieler medialen Höhe meinen Abonnementen

denen noch Generationen von Deutschen leiden werden und über die die das Parlament nie entsc...

...hieden hat Verantwortlich sind sie alle für politische Verfolgungen alles wirklich alles vernünftigen alles Deutschen im besten Sinne und für die politische Instrumentalisierung des Inla...

...ndsgeheimdienstes gegen die größte Oppositionsfraktion Sie sind verantwortlich

Figure 2: Multi-line preview of audio file and transcript in the Octra editor within the Transcription Portal

transcript. This pipeline works in any of the languages supported by G2P or MAUS.

If the user has specified an email address, a link to download the results will be sent once processing is done – ideal for overnight processing.

Finally, all services feature help pages, and they may be used via REST calls from scripts and application programs. In fact, transcription editors such as ELAN (Wittenburg et al. 2006), EXMARaLDA (Schmidt & Wörner, 2014), or Octra access BAS web services in the background.

Transcription portal

The transcription portal was designed as a zero-configuration service for transcribing oral history recordings. Audio files are entered into the portal via drag & drop on the graphical user interface, The user then selects the language and which processing steps to apply (automatic speech recognition, manual correction, word alignment, export). See Figure 1.

The portal displays the status of every file in the workflow, and automatically calls all necessary services and tools in the workflow. See Figure 2.

At each step, the current state of the file can be examined and downloaded.

For subsequent in-depth analysis, the transcription portal supports exporting the transcript in a number of common formats, e.g. ELAN eaf, Praat TextGrid, or tabular or plain text for statistical and linguistic analysis.

Currently, the transcription portal relies on external service providers for the automatic speech recognition. In the ATRIUM project, AI-based speech recognition will be integrated into the portal. This will not only eliminate the need to access third party providers, but also increase the number of languages that can be accessed.

Octra Backend

Oral History recordings generally contain private information and thus strict requirements on data protection must be met. Octra Backend is a software for the management of transcription projects. It was designed with privacy as one of its key features. Octra backend operates in three scenarios:

a) in closed local area networks
b) in the intranet of a workgroup, a company or an institution, or
c) globally via the internet.

In all scenarios, only registered users may access data, and access is regulated by roles in a project. In scenario a), access is restricted to known machines in the local network, while in scenarios b) and c) privacy is ensured by encrypted communication. Audio and transcription data is managed by Octra Backend in its own protected file space, with file names hidden from outside viewers.

In Octra Backend, project administrators define tasks, e.g. manual correction of transcripts generated by ASR, or creating transcripts from scratch, and can assign these tasks to specific transcribers. Transcriptions are performed via Octra in the browser, and there is no need to open or save files, increasing process efficiency and reducing error-prone manual interactions.

The ATRIUM-project[5] starting in 2024 is an EU-funded infrastructure project targeted at archaeologists, with a section devoted to processing spoken language, namely for transcribing Oral History interviews. This task will focus on improving the user experience and the transcription performance, both in terms of quality and efficiency, of the existing transcription portal developed and maintained by BAS.

3. LINDAT for Oral Historians

The web-based ASR engine named UWebASR[6] has been deployed as a service within the LINDAT/CLARIAH-CZ portal in 2018. In order to be able to access this service – and other services that are available in the portal – the user has to authenticate herself/himself as a member of academia (in the same manner as for the use of BAS services mentioned above). The aim is to always provide the best possible ASR performance – we have therefore switched the underlying technology to state-of-the-art wav2vec models (Baevski, 2020) recently. The service is provided for recordings in English, Czech, Slovak and German. All the language-specific models are built from pre-trained models using an innovative 2-phase fine-tuning depicted in Figure 3 – the models are first fine-tuned to a target language in general and then to the specific domain, in our case the oral history interviews.

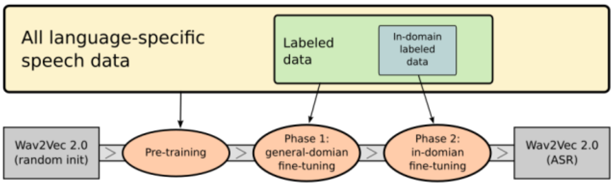

All language-specific speech data
Labeled data
In-domain labeled data
Wav2Vec 2.0 (random init)
Pre-training
Phase 1: general-domian fine-tuning
Phase 2: in-domian fine-tuning
Wav2Vec 2.0 (ASR)

Figure 3: The scheme of 2-phase fine-tuning

The fine-tuning procedure was the same for all the languages, one of the key differences is in the choice of pre-trained model. For English, the model from Meta AI (`wav2vec2-base`) was used. For Czech and German, we have pre-trained our own model from scratch. The Czech model named ClTRUS[7] was then

[5] https://archaeologydataservice.ac.uk/about/projects/atrium/

[6] https://lindat.mff.cuni.cz/services/uwebasr/

[7] https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS

re-used also as the base model for Slovak as this language is very similar to Czech.

Further details about the model training, including the description of the data used for fine-tuning, can be found in (Lehečka, 2023a) for English, Czech and German models and in (Lehečka, 2023b) for Slovak model.
Since the output from wav2vec models is a lower-cased, time-aligned continuous stream of words, further post-processing is needed in order to provide a high-quality human-readable text; this includes mainly case restoration, segmentation into sentences and adding of appropriate punctuation. All those post-processing steps are performed automatically, again using the latest NLP techniques employing the Transformer architecture – see (Švec, 2021) for details.The UWebASR service uses a simple HTTP API interface - the input data can be passed directly within the HTTP request or as a link to a file in the form of a URL. Live audio stream recognition from a given URL is supported as well. The output format includes plain text, machine-readable XML and JSON formats, and the WebVTT format for web captions. Recognition results (except TRS format) are streamed continuously.

While automatically generated transcripts and subtitles assist researchers in locating relevant interviews, they fall short in facilitating a comprehensive understanding of the entire testimony, whether through speech or text. Our innovative approach, again leveraging Transformer-based neural networks, seeks to bridge this gap. It not only aids in clearer navigation through lengthy testimonies but also transforms the listening experience from passive to interactive. By generating contextually relevant questions, our system enriches the interview monologues, allowing listeners to better orient themselves within the narrative and identify key segments of interest. These questions are designed to enhance understanding without altering the original meaning of the testimony, thereby maintaining the integrity of the historical record. Additionally, this method empowers users to engage more deeply with the material by posing their own inquiries, fostering a dynamic exploration of the rich narratives contained within the archives (Švec et al., 2024). This functionality is currently available outside of the LINDAT/CLARIAH-CZ portal in the test mode and will be integrated to the portal in the near future.

4. Do it yourself with Whisper

In Autumn 2022 OpenAI delivered Whisper[8] (Radford et al., 2022): an open source ASR toolkit that can handle about 100 different languages. OpenAI became famous with Chat GPT which was delivered a couple of months later and the enclosed LLM, but it also delivered a very good ASR engine that can be used in a (inter)national cloud environment (for example at a faculty), a local environment or at your local computer at home. The recognition velocity depends on the hardware used, but the recognition results are the same.

In the Netherlands, a first version of Whisper was tested by the end of 2022. We started to use it and made our colleagues in various ASR-projects enthusiastic. Whisper and its derivates, are basically a set of python instructions which can be installed on Windows, Linux, and Mac computers. Moreover, with a GPU in your computer, Whisper performs the recognition up to 10x faster.

OpenAI in the meantime, updated the environment and delivered in February 2023 a large model V2 (and some weeks later model V3) that increased the recognition results for most languages. Our impression is that V3 suffers more from hallucinations, which is confirmed by colleagues. For now, we therefore continue to use V2 as the "best" model.

The benefit of using Whisper is the open source character (more later) and the use of a powerful audio-conditional language model[9] (ACLM, Radford et al.).

When using Kaldi (Povey et al, 2011), we always had the problem that relatively unknown names or terms were not recognized. With Whisper, this is no longer an obstacle, because it uses an ACLM that “knows” these names and terms. Another advantage is that, unlike Kaldi, the recognized text includes punctuation, and capitalized words which greatly improves readability.

However, it should be noted that this applies to those languages that are "well" recognised, where the recordings are of good quality, people speak in a "standard" manner (no dialect) and contain no or little noise or background noise. If these conditions are less fulfilled, then, of course, the quality of transcriptions will also decrease. Yet Whisper surprises by often still offering a very usable result even in "worse" recordings.

Finally, the open-source character has the great advantage that not only OpenAI but also anyone with an eye for detail can use it to make it faster, better and richer. From February 2023, initiatives like WhisperX[10], Fast-Whisper[11], and others started working on improving Whisper's recognition procedure and making it more accessible.

For example, in February 2023, Jordi Bruin came up with MacWhisper[12], a very useful MacOS app using CPP conversion[13] of the original code. MacWhisper is an ideal tool for quickly creating a textual transcription (with or without timecode).

[8] https://openai.com/research/Whisper
[9] https://cdn.openai.com/papers/whisper.pdf
[10] https://github.com/m-bain/WhisperX
[11] https://github.com/SYSTRAN/faster-Whisper
[12] https://goodsnooze.gumroad.com/l/macWhisper
[13] https://github.com/ggerganov/Whisper.cpp

In October 2023 aTrain[14], a Windows-based version of Whisper, was delivered by the University of Gräz[15]. It is a fast and improved version of what was possible with Whisper (see Haberl et al., 2024).

Some "disadvantages" of Whisper are e.g. the not very accurate time estimation (when exactly was which word said) and the absence of speaker diarization. In the summer of 2023, Fast-Whisper already came with an even greater acceleration and in the autumn of 2023, WhisperX came with better time estimation and initial diarization. Since January 2024 diarization seems to be “a solved issue”.

SURF in the Netherlands[16] had taken the initiative to establish Whisper as a service so that anyone with a SURF account could use it. Since a substantial number of researchers work with “sensitive” material, many research groups have set up a Whisper installation in their own secure network.

Google is working hard on updating an even better language model (Chirp, Universal Speech Model[17]) that will be able to handle 1000 different languages, and so did Meta with SeamlessM4T[18], but for now Whisper and its derivative versions, is the best candidate for a very good, fast and relatively easy ASR engine.

5. Remaining challenges

As remarked, transformer based speech models perform exceptionally better than the classical modular speech architectures such as Kaldi. This is especially true if a clean orthographic transcription with appropriate punctuation is the target of the speech to text conversion, which is the case for many research purposes. However, notably for linguistic research, more detailed output may be relevant. Which may require more than just this. Interviews are a relevant source e.g. for discourse analysis where pause durations and disfluencies play a paramount role in various methodological approaches See e.g. Van den Heuvel and Oostdijk, 2016). However, it are these phenomena that typically remain under the radar in standard ASR output (Lopez, Liesenfeld & Dingemanse, 2022). Disfluencies in spontaneous speech include *repetitions* (e.g. the the), *corrections* (e.g. Show me the flights … the early flights), *restarts* (e.g. There’s a … Let’s go), *filled pauses* (e.g. um and uh), and truncations resulting in *partial words* (e.g. wou- and oper-) (Lou & Johnson, 2020). The efforts in creating the transformed based large speech models are typically directed towards suppressing these phenomena in the text output they generate. Nonetheless, lately, there has been a growing interest in the advancement of technology that can decode speech while considering disfluencies. This technology is aimed at studying interruptions and corrections in communication settings. It has found various applications in the medical field, such as identifying early signs of cognitive decline (Claza et al., 2021, stuttering (Mitra et al, 2021), and detecting disfluencies through diagnostic tasks (Rohanian et al., 2021). Researchers are employing a combination of traditional AM/LM architectures and newer end-to-end models utilizing bidirectional LSTMs working in offline (non real-time) mode. These models incorporate features like word probabilities, confidence scores, prosodic features, pause duration statistics, and a range of acoustic features including fluency and speaking rate, which are now becoming standard in this area of research (Huang et al., 2018).

Another challenge is speaker diarisation which attributes the text output to resp. the interviewer and the interviewee(s). See for a review Park et al, (2022). Whisper-X is providing speaker diarisation in its output (Bain et al. 2023) as mentioned, but its performance needs to be explored and improved.

6. Conclusion

In this contribution we sketched a number of initiatives and toolkit approaches to improve automatic speech recognition for oral history interviews whilst offering these in a safe, well protected data shield minimizing dataleaks. We address the transcription portal and the webservices associated with speech processing at BAS, speech solutions developed at LINDAT, how to do it yourself with Whisper, but also mentioned a number of remaining challenges.

7. Bibliographical References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-1246

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. ArXiv, abs/2303.00747.

Calzà, L., Gagliardi, G., Favretti, R. R., & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65, 101113.

Draxler, Chr. (2022) Automatic Transcription of Spoken Language Using Publicly Available Web Services. In: *FARE LINGUISTICA APPLICATA CON LE DIGITAL HUMANITIES, 2022*. https://bia.unibz.it/view/pdfCoverPage?instCode=39UBZ_INST&filePid=13284996250001241&download=true#page=28

Draxler, Chr., Van den Heuvel. H., Van Hessen, A., Calamai, S., Corti, L., Scagliola, S. (2020) A CLARIN Transcription Portal for Interview Data. In

[14] https://apps.microsoft.com/detail/9n15q44szns2?hl=en-US&gl=US

[15] https://doi.org/10.1016/j.jbef.2024.100891

[16] https://www.surf.nl/en

[17] https://cloud.google.com/speech-to-text/v2/docs/chirp-model

[18] https://ai.meta.com/blog/seamless-m4t/

Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020). pp. 3346-3352, http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.411.pdf

Draxler, Chr., Pömp, Julian (2022) OCTRA – An Innovative Approach to Orthographic Transcription. In *Proceedings INTERSPEECH 2022*, 5217-5218

Gref, M. (2022). *Robust Speech Recognition via Adaptation for German Oral History Interviews* (Doctoral dissertation, Universitäts-und Landesbibliothek Bonn). https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/10373

Haberl, A., Fleiß, J., Kowald, D., & Thalmann, S. (2024). Take the aTrain. Introducing an interface for the Accessible Transcription of Interviews. *Journal of Behavioral and Experimental Finance*, 41, 100891. https://doi.org/10.1016/j.jbef.2024.100891

Huang, H. Y., Choi, E., & Yih, W. T. (2018). Flowqa: Grasping flow in history for conversational machine comprehension. arXiv preprint arXiv:1810.06683.

Kipp, A., Wesenick, M.-B., Schiel, F. (1997) Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. In *Proceedings Eurospeech 1997,* 1023-1026, doi: 10.21437/Eurospeech.1997-358

Lehečka, J., Švec, J., Psutka, J.V., Ircing, P. (2023a) Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Proceedings INTERSPEECH 2023*, 201-205, doi: 10.21437/Interspeech.2023-872

Lehečka, J., Psutka, J.V., Psutka, J. (2023b). Transfer Learning of Transformer-Based Speech Recognition Models from Czech to Slovak.. *TSD 2023. Lecture Notes in Computer Science 2023*, vol 14102, doi: 10.1007/978-3-031-40498-6_29

Lopez, A., Liesenfeld, A., & Dingemanse, M. (2022). Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing? *In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. Potsdam.

Lou, P. J., & Johnson, M. (2020). End-to-end speech recognition and disfluency removal. arXiv preprint arXiv:2009.10298.

Mitra, V., Huang, Z., Lea, C., Tooley, L., Wu, S., Botten, D., Palekar, A., Thelapurath, S., Georgiou, P., Kajarekar, S., Bigham, J. (2021) Analysis and Tuning of a Voice Assistant System for Dysfluent Speech. In *Proceedings Interspeech 2021*, 4848-4852, doi: 10.21437/Interspeech.2021-2006.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72, 101317.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF*). IEEE Signal Processing Society.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Rohanian, M., Hough, J., & Purver, M. (2021). Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. arXiv preprint arXiv:2106.15684.

Švec, J., Lehečka, J., Šmídl, L., Ircing, P. (2021). Transformer-Based Automatic Punctuation Prediction and Word Casing Reconstruction of the ASR Output. TSD 2021. *Lecture Notes in Computer Science*, vol 12848, doi: 978-3-030-83527-9_7

Švec, J., Bulín, M., Frémund, A., Polák, F. (2024) Asking questions framework for oral history archives. *Accepted for ECIR 2024*

Scagliola, S., Corti, L., Calamai, S., Karrouche, N., Beeken, J., Van Hessen, A., Draxler, Chr., Van den Heuvel, H., Broekhuizen, M., Truong, K. (2020) Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow. In: Simov, K., & Eskevich, M.: *Selected Papers from the CLARIN Annual Conference 2019* Leipzig, 30 September - 2 October 2019. Linköping Electronic Conference Proceedings 172:15, pp. 126-136. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=172&Article_No=15

Schmidt, Thomas & Wörner, Kai (2014) 'EXMARaLDA', in Jacques Durand, Ulrike Gut, and Gjert Kristoffersen (eds), *The Oxford Handbook of Corpus Phonology* (2014; online edn, Oxford Academic, 4 Aug. 2014), https://doi.org/10.1093/oxfordhb/9780199571932.013.030

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559), https://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4

Van den Heuvel. H. & Oostdijk, N. (2016) Falling silent, lost for words ... Tracing personal involvement in interviews with Dutch war veterans. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC2016)*, Portorož, 23-28 May 2016. pp. 998-1001

Identifying Narrative Patterns and Outliers in Holocaust Testimonies Using Topic Modeling

Maxim Ifergan† **Renana Keydar**‡ **, Omri Abend**† **Amit Pinchevski**◊
† Department of Computer Science ‡ Faculty of Law and Digital Humanities
◊ Department of Communication and Journalism
Hebrew University of Jerusalem
{first_name}.{last_name}@mail.huji.ac.il

Abstract

The vast collection of Holocaust survivor testimonies presents invaluable historical insights but poses challenges for manual analysis. This paper leverages advanced Natural Language Processing (NLP) techniques to explore the USC Shoah Foundation Holocaust testimony corpus. By treating testimonies as structured question-and-answer sections, we apply topic modeling to identify key themes. We experiment with BERTopic, which leverages recent advances in language modeling technology. We align testimony sections into fixed parts, revealing the evolution of topics across the corpus of testimonies. This highlights both a common narrative schema and divergences between subgroups based on age and gender. We introduce a novel method to identify testimonies within groups that exhibit atypical topic distributions resembling those of other groups. This study offers unique insights into the complex narratives of Holocaust survivors, demonstrating the power of NLP to illuminate historical discourse and identify potential deviations in survivor experiences.

Keywords: Topic Modeling, Narrative, Testimonies, Holocaust

1. Introduction

In recent decades, significant efforts have been made to gather the accounts of the remaining Holocaust survivors. The passing of the last living witnesses and the beginning of the era of the post-testimony occurs simultaneously with technological developments in NLP.The wealth of testimonies in the archives presents a challenge: how to preserve the significance of individual stories within a vast collection of a thousand testimonies, while also giving voice to the collective body of testimonies in a manner that honors the individuality of each story. By employing techniques such as contextualized topic modeling and topic narrative analysis, we aim to uncover broad trends within the collection, while ensuring the preservation of the uniqueness and integrity of each personal narrative.

Despite advancements in NLP, representation of long texts still poses a challenge to state-of-the-art models (Piper et al., 2021; Castricato et al., 2021; Mikhalkova et al., 2020; Dong et al., 2023). Antoniak et al. (2019) pioneered the representation and visualization of narratives as sequences of interpretable topics. And while previous topic modeling analyses of Holocaust testimonies (Blanke et al., 2019) have provided valuable insights, they treated the corpus as a monolithic body of text, obscuring the unique narrative structure of individual testimonies. Furthermore, using non-contextualized topic modeling such as LDA ((Blei et al., 2001)) treated the text as a body of words without order. Recent advancements in topic modeling techniques such as BERTopic (Grootendorst, 2022), and other Contextualized Topic Modeling (Bianchi et al., 2020; Angelov, 2020; Pham et al., 2023) leverage language model representation to better identify and predict the text topics. While such methods were applied to Holocaust testimonies (Wagner et al., 2022), the main focus was on the segmentation of the testimonies for topic modeling. Our contributions are as follows:

- We apply a novel contextualized topic modeling approach, BERTtopic, to holocaust testimonies, revealing the main themes and their distribution.
- We examine the evolution of topics across aligned sections of testimonies, revealing a typical narrative scheme.
- We investigate how age and gender are expressed in the narrative structure of testimonies, highlighting distinctions between survivor subgroups.
- We introduce a novel method for identifying divergent testimonies, i.e., testimonies within a given group that exhibit atypical topic distributions, resembling patterns more characteristic of other groups. We demonstrate it in a case-study of different age groups.

We note that related contributions appear in an unpublished paper of ours (under review; anonymized) that uses an earlier contextualized topic model (CTM; Bianchi et al., 2020) for a similar process. The current paper uses a better performing model (Grootendorst, 2022) regarding topic di-

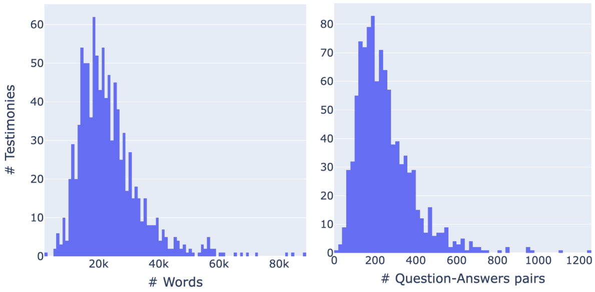


Figure 1: Testimonies number of words and number QA-pairs histogram.

versity and coherence and a more detailed and precise narrative analysis approach.

2. Corpus Level Statistics

This paper analyzes transcripts from the USC Shoah Foundation, a corpus containing 1000 oral testimonies in English. Survivors originated from over 30 countries, with a significant representation from Poland and Germany. The testimonies were recorded between 1996 and 2015, offering insights into the survivors' experiences decades after the events of the Holocaust. The length of the testimonies ranges from 3K to 88K words, with a mean length of 23K words. Each testimony contains an average of 250 questions, with the majority of question-answer pairs (95%) consisting of no more than 400 words. Fig. 1 illustrates the distribution of testimony lengths.

3. BERTopic: Topic Analysis

We use BERTopic to identify the topics within the corpus. Preprocessing involves the merging of consecutive very short sections (question-answer pairs <200 words) and the division of very long sections (>450 words) to mitigate potential outlier effects. BERTopic leverages all-MiniLM-L6-v2 (Wang et al., 2020) document embeddings and a TF-IDF based clustering approach, providing a context-aware analysis that surpasses traditional methods like LDA (Blei et al., 2001). For dimensionality reduction, UMAP (McInnes and Healy, 2018) is employed before clustering with HDBSCAN (McInnes et al., 2017). Unlike LDA, BERTopic dynamically determines the number of topics only by determining the minimum cluster size for HDBSCAN, resulting in greater flexibility. Our dataset yielded 58 topics, with approximately 4% outliers classified as "unknown topic". We set the minimum cluster size to be 50 sections.

To ensure interpretability, BERTopic extracts c-TF-IDF[1] word representations from each section's cluster, revealing the importance of words within each topic. The most representative word is selected for initial topic representation. A domain expert then manually reviews these word sets and assigns a descriptive title to each topic, ensuring both accuracy and clarity. Notably, the topics detected by the model align with those outlined in the USC Shoah Foundation's interviewer guidelines [2] but also extend way beyond them. The guidelines encourage interviewers to ask about pre-war life, family, religion, politics, community, and experiences of antisemitism. The model's successful detection of these themes confirms the effectiveness in identifying core key topics.

4. Narrative analysis

This study analyzes individual survivor testimonies as narratives – sequences of interpretable topics (Antoniak et al., 2019). We aim to construct comprehensive narratives from the corpus testimonies that enable comparisons without sacrificing their temporal structures. Several challenges arise in this analysis. First, each testimony comprises a large number of sections (250 on average), conflicting with the direct interpretation and visualization purposes. Secondly, variations in testimony length complicate direct comparisons of narrative structures.

To address this, we divide testimonies into a fixed manageable number of parts, defining the part's theme representation as the distribution of its sections' topics. This division requires considering the trade-off between preserving temporal detail and achieving clear visualization and comparison. A large number of parts yields more nuanced narratives but risks excessive details and redundancy, whereas fewer parts allow better interpretation and visualization at the expense of obscuring finer temporal shifts in topics. After careful examination and consultation with domain experts in Holocaust studies and digital humanities, we divide testimonies into 15 equal parts. This strikes a balance between the need for detail and the goals of clear visualization and comparisons of topic distribution across parts.

4.1. Typical Testimony Narrative Schema

This section examines the most common topics covered in each part of a Holocaust survivor testimony, as well as the variation in topic representation between the different testimony parts. The analysis is based on Fig. 3, which shows the distribution of topics across the 15 parts into which each testimony was divided.

[1] https://maartengr.github.io/BERTopic/api/ctfidf.html

[2] https://sfi.usc.edu/content/interviewer-guidelines

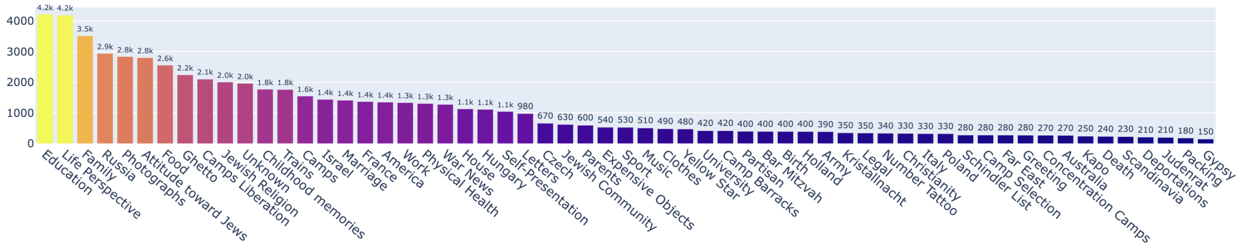
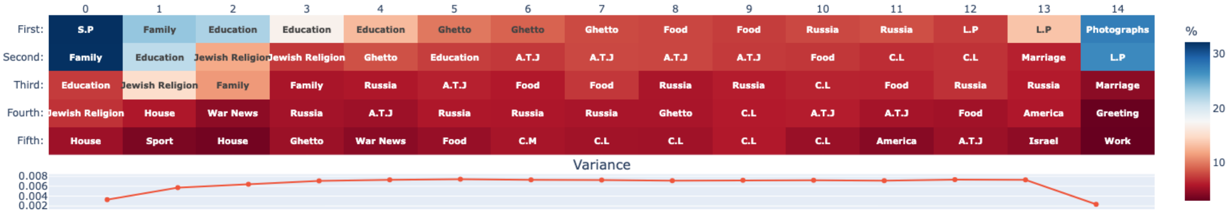


Figure 2: Corpus level QA-s topics histogram.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
First:	S.P	Family	Education	Education	Education	Ghetto	Ghetto	Ghetto	Food	Food	Russia	Russia	L.P	L.P	Photographs
Second:	Family	Education	Jewish Religion	Jewish Religion	Ghetto	Education	A.T.J	A.T.J	A.T.J	A.T.J	Food	C.L	C.L	Marriage	L.P
Third:	Education	Jewish Religion	Family	Family	Russia	A.T.J	Food	Food	Russia	Russia	C.L	Food	Russia	Russia	Marriage
Fourth:	Jewish Religion	House	War News	Russia	A.T.J	Russia	Russia	Russia	Ghetto	C.L	A.T.J	A.T.J	Food	America	Greeting
Fifth:	House	Sport	House	Ghetto	War News	Food	C.M	C.L	C.L	C.L	C.L	America	A.T.J	Israel	Work

Figure 3: The 5 most prevalent topics and topics variance for each part. A.T.J = Attitude toward Jews, S.P = Self-Presentation, C.L = Camps Liberation, L.P = Life Perspective, and C.M = Childhood Memories

The first part of all testimonies is dominated by the topic of self-presentation followed by the family topic. It is perhaps unsurprising that many testimonies begin this way, as survivors introduce themselves and their families to the interviewer. The fact that the self-presentation topic rarely appears later may not constitute a significant finding, but it is nevertheless of importance as it validates the model's analytic capability. The next two parts of the testimonies also reveal a number of common topics, associated with the description of community life before the war. These include family, education, religion, house, and sport. The latter part also contains the topic of war news, hinting at the events to come.

In contrast to the dominance of common topics at the beginning of the testimonies, the middle parts show greater variance in the topic distributions. Each part typically features several common topics with similar percentages (around 5-15%). This might reflect the diversity of experiences among Holocaust survivors. The middle part topics vary starting with ghettos and war news to concentration/death camps and food, resolving in the rise in dominance of the camps liberation topic.

In the final parts of the testimonies narratives once again the model identifies a few dominant topics. These include interview-related topics such as presenting family pictures and discussing life after the Holocaust topics. Additionally, topics related to life after the war emerged, such as immigration and establishing work and marriage in new countries.

In conclusion, the BERTopic model successfully identifies a typical structure for Holocaust survivor testimonies, particularly at the beginning and end. The middle sections show more variation, reflecting the different experiences of individual survivors.

4.2. Gender and Age as Expressed in Testimonies Narratives

This section introduces a method for comparing the narrative trajectories present within different groups of Holocaust survivor testimonies. We apply this method to investigate gender- and age-based differences in testimonial narratives.

To begin, we compute a typical testimony path for each group under consideration (e.g., male vs. female, young vs. older survivors). This 'typical' testimony schema represents the average topic distribution across the 15 fixed parts. Next, we perform t-tests for each part to quantify the differences in topic prevalence between groups. Topics with a substantial t-value (above 3.5%) and a low probability of such deviation arising by chance (p-value under 0.01) are flagged as characteristic of the group in which they are more prevalent.

Let's consider the age-based comparison between younger survivors, born 1925-1940, experiencing the Holocaust as children (522 testimonies), and older survivors, born 1902-1925, adults during the Holocaust (467 testimonies). Fig. 4 reveals interesting distinctions. Topics like "Childhood Memories" and "Food" dominate the middle parts of younger survivors' testimonies, while "Life Perspective" features in the final parts. Conversely, "Marriage", "Work", and "War News" are more prominent in the middle of older survivors' accounts. Interestingly, while education-related topics seem more prevalent at the beginning of older survivors' tes-

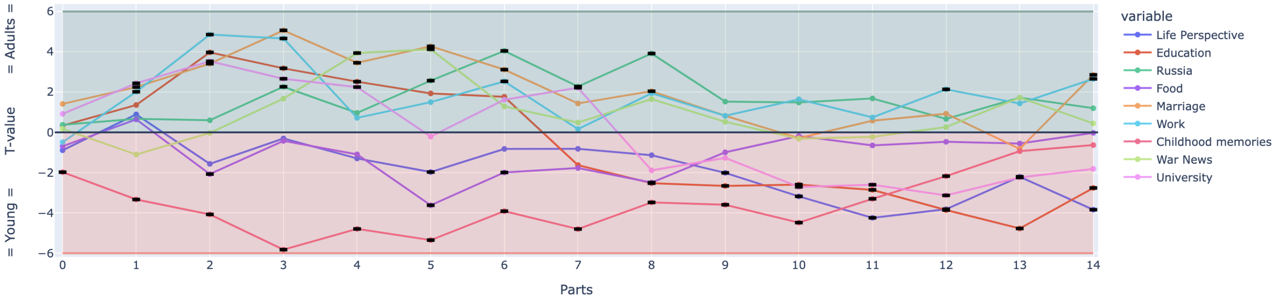


Figure 4: Adults vs. young survivors typical testimony t-test. The Black Point represents values with p-values under 0.01.

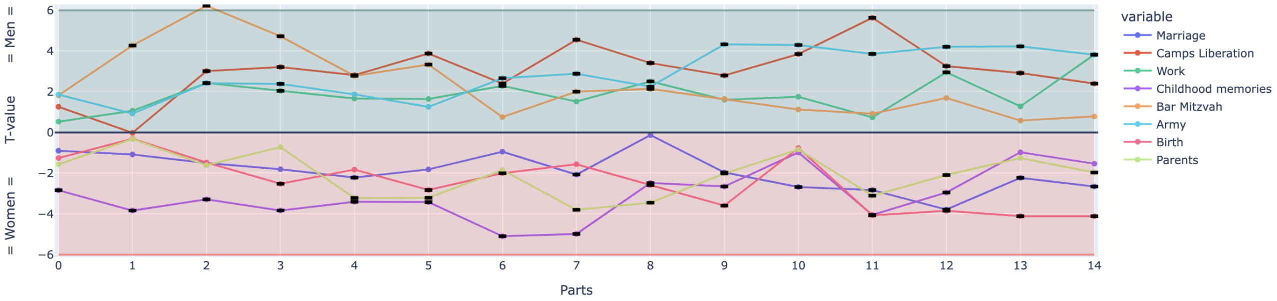


Figure 5: Men vs. women survivors typical testimony t-test. The Black Point represents values with p-values under 0.01.

timonies, they tend to re-emerge near the end for the younger group. Turning to the gender-based analysis, with a balanced corpus of 531 male and 469 female testimonies, Fig. 5 highlights potential differences. Topics like "Bar Mitzvah", "Army", "Camp Liberation", and "Work" are more characteristic of men's testimonies. In contrast, "Birth", "Childhood Memories", "Parents", and "Marriage" are more prevalent in women's testimonies. This analysis reveals how men and women may structure their narratives differently, particularly in the middle sections of their testimonies.

The USC Shoah Foundation's interviewer guidelines do not provide specific instructions for ordering topics or tailoring questions based on the subject's age or gender. This suggests that the observed differences in narrative structure between these groups are not a direct result of the guidelines. Rather, they may stem from the interviewers' individual approaches or the survivors' unique experiences and perspectives.

5. Exploratory Study Identifying Diverging Narratives

This study introduces a novel method for identifying testimonies within a specific group that exhibit topic distribution patterns more characteristic of another group. Our goal is to pinpoint narratives that stand out as atypical within their designated category. We achieve this by defining a scoring function that quantifies the similarity between a testimony's topic distribution and the typical narrative patterns of a different group. Let us formalize the scoring function which yields a high score for testimonies from A that resemble the narrative patterns typical of B. Let $t = (t_1, t_2, ..., t_{15})$ represent the testimony's topic distributions from group A, where each t_i is a vector of topic probabilities for part i. And, $C_A = \{(i_1, j_1), (i_2, j_2), ..., (i_n, j_n)\}$ denotes the characteristic topic-part pairs for group A, where i_x is a part index and j_x is a topic index. These pairs have high t-values (>3.5) and low p-values (<0.01) in the group comparison. C_B similarly represents the characteristic topic-part pairs for group B.

$$R_B = \sum_{(i,j) \in C_B} t_i[j] \cdot |Tvalue_B(i,j)|$$

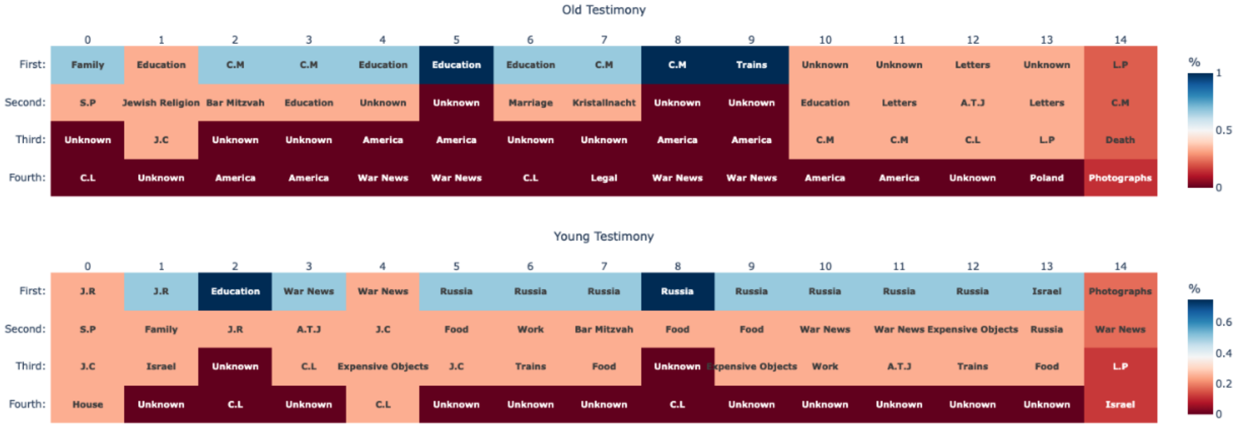
$$R_A = \sum_{(i,j) \in C_A} t_i[j] \cdot |Tvalue_A(i,j)|$$

$$S(t, C_A, C_B) = R_B - R_A$$

Finally, we apply an argmax operation to spot those testimonies exhibiting the highest resemblance to group B's typical narrative:

$$argmax_{t \in A} S(t, C_A, C_B)$$

When comparing older and younger survivor groups, Fig. 7 presents the distribution of resem-



Old Testimony

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
First:	Family	Education	C.M	C.M	Education	Education	Education	C.M	C.M	Trains	Unknown	Unknown	Letters	Unknown	L.P
Second:	S.P	Jewish Religion	Bar Mitzvah	Education	Unknown	Unknown	Marriage	Kristallnacht	Unknown	Unknown	Education	Letters	A.T.J	Letters	C.M
Third:	Unknown	J.C	Unknown	Unknown	America	America	Unknown	Unknown	America	America	C.M	C.M	C.L	L.P	Death
Fourth:	C.L	Unknown	America	America	War News	War News	C.L	Legal	War News	War News	America	America	Unknown	Poland	Photographs

Young Testimony

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
First:	J.R	J.R	Education	War News	War News	Russia	Russia	Russia	Russia	Russia	Russia	Russia	Russia	Israel	Photographs
Second:	S.P	Family	J.R	A.T.J	J.C	Food	Work	Bar Mitzvah	Food	Food	War News	War News	Expensive Objects	Russia	War News
Third:	J.C	Israel	Unknown	C.L	Expensive Objects	J.C	Trains	Food	Unknown	Expensive Objects	Work	A.T.J	Trains	Food	L.P
Fourth:	House	Unknown	C.L	Unknown	C.L	Unknown	Unknown	Unknown	C.L	Unknown	Unknown	Unknown	Unknown	Unknown	Israel

Figure 6: Outliers testimonies. A.T.J = Attitude toward Jews, S.P = Self-Presentation, C.L = Camps Liberation, L.P = Life Perspective, J.C = Jewish Community, J.R = Jewish Religion, and C.M = Childhood Memories

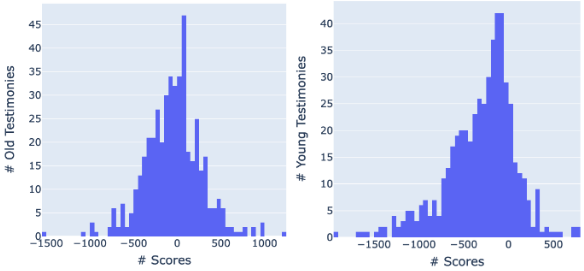


Figure 7: Testimonies scores histogram.

blance scores for testimonies within the groups. The uneven distribution favoring negative scores reveals that higher scores tend to be related to non-conforming narratives. Using this method, Fig. 6 highlights two specific examples: a younger survivor whose narrative strongly resembles the older group, and vice-versa emphasizing topics characteristic of the opposite group.

6. Conclusion and Future Work

This study applies NLP techniques to explore the complex narratives within the USC Shoah Foundation's Holocaust testimonies. Contextualized topic modeling with BERTopic reveals key themes and their distributions within the corpus. And, by aligning testimonies into fixed parts, we unveiled a common narrative trajectory along with age- and gender-based variations. Our method detects divergent testimonial narratives, identifying those within one group that exhibit topic patterns characteristic of another group. Future Work will extend the analysis by comparing survivor narratives across corpora[3] and other testimonial archives to identify both shared and distinct narratives patterns.

7. Acknowledgements

The authors acknowledge the USC Shoah Foundation - The Institute for Visual History and Education for its support of this research. This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Alfred Landecker Foundation.

8. Bibliographical References

Dimitar Angelov. 2020. Top2Vec: Distributed Representations of Topics. *ArXiv*, abs/2008.09470.

Maria Antoniak, David M. Mimno, and Karen E. C. Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proceedings of the ACM on Human-Computer Interaction*, 3:1 – 27.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Annual Meeting of the Association for Computational Linguistics*.

Tobias Blanke, Michael Bryant, and Mark Hedges. 2019. Understanding memories of the Holocaust - A new approach to neural networks in the digital humanities. *Digit. Scholarsh. Humanit.*, 35:17–33.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Louis Castricato, Stella Biderman, David Thue, and Rogelio Cardona-Rivera. 2021. Towards

[3] Yale Fortunoff Archive

a model-theoretic view of narratives. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 95–104, Virtual. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Lunyi Li, and Wayne Xin Zhao. 2023. A survey on long text modeling with transformers. *ArXiv*, abs/2302.14502.

Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, abs/2203.05794.

Leland McInnes and John Healy. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, abs/1802.03426.

Leland McInnes, John Healy, and S. Astels. 2017. HDBSCAN: Hierarchical Density Based Clustering. *J. Open Source Softw.*, 2:205.

Elena Mikhalkova, Timofei Protasov, Polina Sokolova, Anastasiia Bashmakova, and Anastasiia Drozdova. 2020. Modelling narrative elements in a short story: A study on annotation schemes and guidelines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 126–132, Marseille, France. European Language Resources Association.

Chau Minh Pham, Alexander Miserlis Hoyle, Simeng Sun, and Mohit Iyyer. 2023. TopicGPT: A Prompt-based Topic Modeling Framework. *ArXiv*, abs/2311.01449.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies. In *Conference on Empirical Methods in Natural Language Processing*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *ArXiv*, abs/2002.10957.

A. Topic list

Topic Title	Top-15 words in topic
Life Perspective	children’, ’think’, ’holocaust’, ’thank’, ’say’, ’life’, ’grandchildren’, ’years’, ’want’, ’experiences’, ’people’, ’much’, ’god’, ’thats’, ’never’
Photographs	picture’, ’taken’, ’thank’, ’photograph’, ’left’, ’name’, ’right’, ’photo’, ’next’, ’daughter’, ’son’, ’crew’, ’sister’, ’year’, ’wife’
Family	name’, ’mother’, ’father’, ’born’, ’family’, ’mothers’, ’remember’, ’sister’, ’brother’, ’fathers’, ’lived’, ’years’, ’sisters’, ’brothers’, ’grandfather’
Education	school’, ’jewish’, ’antisemitism’, ’hitler’, ’jews’, ’remember’, ’teacher’, ’hebrew’, ’yiddish’, ’german’, ’friends’, ’language’, ’teachers’, ’polish’, ’years’
Russia	russian’, ’russians’, ’russia’, ’people’, ’away’, ’germans’, ’army’, ’take’, ’train’, ’told’, ’come’, ’want’, ’food’, ’says’, ’already’
Jewish Religion	synagogue’, ’holidays’, ’used’, ’shabbos’, ’remember’, ’religious’, ’shabbat’, ’shul’, ’friday’, ’father’, ’passover’, ’home’, ’holiday’, ’jewish’, ’family’
Self-Presentation	name’, ’born’, ’birth’, ’spell’, ’interview’, ’please’, ’date’, ’english’, ’today’, ’1997’, ’conducting’, ’interviewer’, ’language’, ’1998’, ’maiden’
Ghetto	ghetto’, ’people’, ’germans’, ’jews’, ’work’, ’food’, ’place’, ’little’, ’lived’, ’remember’, ’jewish’, ’house’, ’already’, ’street’, ’away’
Food	food’, ’bread’, ’soup’, ’eat’, ’potatoes’, ’little’, ’day’, ’used’, ’people’, ’water’, ’piece’, ’work’, ’something’, ’hungry’, ’put’
Marriage	married’, ’wedding’, ’met’, ’husband’, ’wife’, ’marry’, ’israel’, ’years’, ’meet’, ’name’, ’marriage’, ’laughs’, ’come’, ’want’, ’family’
Israel	’israel’, ’palestine’, ’zionist’, ’british’, ’kibbutz’, ’organization’, ’jewish’, ’people’, ’army’, ’hebrew’, ’tel’, ’war’, ’come’, ’also’, ’years’
France	’french’, ’france’, ’paris’, ’belgium’, ’people’, ’brussels’, ’germans’, ’train’, ’german’, ’antwerp’, ’war’, ’border’, ’think’, ’vichy’, ’little’
Physical Health	’hospital’, ’sick’, ’doctor’, ’typhus’, ’camp’, ’fever’, ’people’, ’doctors’, ’typhoid’, ’day’, ’couldnt’, ’food’, ’take’, ’put’, ’work’
Hungary	’hungarian’, ’hungary’, ’hungarians’, ’budapest’, ’jews’, ’romania’, ’romanian’, ’jewish’, ’people’, ’war’,’many’, ’labor’, ’started’, ’germans’, ’father’
America	’ship’, ’states’, ’united’, ’york’, ’boat’, ’new’, ’visa’, ’america’, ’come’, ’arrived’, ’american’, ’quota’, ’affidavit’, ’days’, ’papers’
Camps	’auschwitz’, ’birkenau’, ’camp’, ’people’, ’gas’, ’work’, ’saw’, ’day’, ’barracks’, ’train’, ’see’, ’barrack’, ’right’, ’knew’, ’women’
House	’room’, ’house’, ’apartment’, ’kitchen’, ’rooms’, ’lived’, ’bedroom’, ’big’, ’living’, ’floor’, ’remember’, ’home’, ’dining’, ’describe’, ’used’
Camps Liberation	’camp’, ’german’, ’american’, ’germans’, ’saw’, ’americans’, ’day’, ’see’, ’people’, ’soldiers’,’planes’, ’war’, ’prisoners’, ’started’, ’liberated’
Trains	’train’, ’people’, ’cattle’, ’trains’, ’wagon’, ’camp’, ’march’, ’long’, ’water’, ’many’, ’station’, ’car’, ’food’, ’days’, ’see’
Work	’job’, ’business’, ’worked’, ’work’, ’money’, ’working’, ’company’, ’store’, ’bought’, ’make’, ’years’, ’week’, ’started’, ’factory’, ’want’
Attitude Toward Jews	’says’, ’jewish’, ’told’, ’come’, ’jews’, ’mother’, ’want’, ’away’, ’german’, ’see’, ’knew’, ’people’, ’look’, ’gave’, ’man’
Music	’music’, ’sing’, ’singing’, ’song’, ’songs’, ’opera’, ’played’, ’piano’, ’sang’, ’play’, ’remember’, ’used’, ’yiddish’, ’voice’, ’laughs’
Childhood memories	’mother’, ’father’, ’remember’, ’think’, ’parents’, ’see’, ’sister’, ’never’, ’really’, ’knew’, ’always’, ’happened’, ’mean’, ’couldnt’, ’tell’
Czech	’prague’, ’czech’, ’czechoslovakia’, ’theresienstadt’, ’train’, ’people’, ’see’, ’come’, ’army’, ’left’, ’knew’, ’stayed’, ’transport’, ’home’, ’want’
Letters	’letters’, ’letter’, ’wrote’, ’mother’, ’war’, ’sister’, ’parents’, ’sent’, ’write’, ’cross’, ’found’, ’red’, ’brother’, ’knew’, ’family’
Bar Mitzvah	’bar’, ’mitzvah’, ’remember’, ’torah’, ’synagogue’, ’mitzvahed’, ’school’, ’jewish’, ’rabbi’, ’hebrew’, ’shul’, ’family’, ’mitzvahs’, ’father’, ’13’

Topic Title	Top-15 words in topic
War News	’radio’, ’poland’, ’jews’, ’polish’, ’war’, ’news’, ’jewish’, ’people’, ’german’, ’knew’, ’germans’, ’germany’, ’heard’, ’poles’, ’warsaw’
Greeting	’thank’, ’yourn’, ’much’, ’muchrn’, ’sharing’, ’welcome’, ’welcomern’, ’add’, ’youre’, ’mrs’, ’testimony’, ’concludes’, ’thanks’, ’say’, ’want’
Legal	’trial’, ’trials’, ’court’, ’nuremberg’, ’crimes’, ’witnesses’, ’case’, ’courtroom’, ’witness’, ’judge’, ’cases’, ’justice’, ’defense’, ’evidence’, ’interpreter’
Sport	’play’, ’soccer’, ’used’, ’played’, ’sports’, ’school’, ’games’, ’remember’, ’sport’, ’friends’, ’swimming’, ’playing’, ’ball’, ’club’, ’liked’
Partisan	’partisans’, ’partisan’, ’group’, ’russian’, ’germans’, ’forest’, ’killed’, ’people’, ’army’, ’fighting’, ’woods’, ’fight’, ’food’, ’knew’, ’thats’
Far East	’shanghai’, ’japanese’, ’chinese’, ’china’, ’japan’, ’hongkew’, ’refugees’, ’people’, ’war’, ’see’, ’money’, ’ship’, ’american’, ’boat’, ’harbor’
Yellow Star	’star’, ’wear’, ’yellow’, ’wearing’, ’stars’, ’jewish’, ’david’, ’jews’, ’remember’, ’armband’, ’wore’, ’germans’, ’jew’, ’people’, ’think’
Number Tattoo	’number’, ’tattooed’, ’tattoo’, ’numbers’, ’auschwitz’, ’camp’, ’arm’, ’barracks’, ’given’, ’birkenau’, ’tattooing’, ’people’, ’triangle’, ’put’, ’prisoners’
Camp Selection	’mengele’, ’selection’, ’auschwitz’, ’gas’, ’camp’, ’right’, ’people’, ’saw’, ’side’, ’twins’, ’selected’, ’see’, ’left’, ’told’, ’experiments’
Schindler List	’schindler’, ’factory’, ’list’, ’schindlers’, ’plaszow’, ’oskar’, ’goeth’, ’brxfcnnlitz’, ’people’, ’camp’, ’brunnlitz’, ’inaudible’, ’working’, ’knew’, ’auschwitz’
Jewish Community	’jewish’, ’town’, ’population’, ’jews’, ’lived’, ’community’, ’city’, ’people’, ’synagogue’, ’families’, ’big’, ’school’, ’area’, ’lot’, ’business’
Expensive Objects	’jewelry’, ’money’, ’gold’, ’ring’, ’things’, ’silver’, ’coins’, ’hide’, ’give’, ’take’, ’put’, ’buy’, ’remember’, ’sold’, ’whatever’
Army	’army’, ’training’, ’basic’, ’infantry’, ’drafted’, ’fort’, ’draft’, ’corps’, ’sergeant’, ’officer’, ’unit’, ’military’, ’citizen’, ’british’, ’service’
Concentration Camps	’bergenbelsen’, ’buchenwald’, ’camp’, ’people’, ’belsen’, ’barrack’, ’dead’, ’arrived’, ’saw’, ’barracks’, ’remember’, ’block’, ’liberated’, ’prisoners’, ’many’
Kristallnacht	’kristallnacht’, ’synagogue’, ’november’, ’happened’, ’1938’, ’remember’, ’father’, ’arrested’, ’night’, ’school’, ’mother’, ’home’, ’glass’, ’jewish’, ’day’
Italy	alian’, ’italy’, ’italians’, ’athens’, ’switzerland’, ’rome’, ’mussolini’, ’people’, ’camp’, ’modena’, ’train’, ’germans’, ’bari’, ’chichibo’, ’nonantola’
Birth	’baby’, ’hospital’, ’doctor’, ’child’, ’pregnant’, ’husband’, ’mother’, ’father’, ’sick’, ’born’, ’cancer’, ’died’, ’told’, ’home’, ’never’
Parents	’father’, ’mother’, ’come’, ’says’, ’take’, ’little’, ’told’, ’saw’, ’knew’, ’see’, ’place’, ’away’, ’want’, ’ill’, ’thought’
Scandinavia	’sweden’, ’denmark’, ’danish’, ’swedish’, ’danes’, ’copenhagen’, ’stockholm’, ’king’, ’people’, ’malmo’, ’jews’, ’goteborg’, ’jewish’, ’swedes’, ’government’
University	’university’, ’college’, ’school’, ’degree’, ’years’, ’new’, ’high’, ’york’, ’job’, ’engineering’, ’masters’, ’worked’, ’work’, ’social’, ’graduate’
Australia	’australia’, ’melbourne’, ’australian’, ’sydney’, ’boat’, ’fremantle’, ’ship’, ’arrived’, ’come’, ’australians’, ’people’, ’job’, ’english’, ’friends’, ’years’
Holland	’dutch’, ’holland’, ’amsterdam’, ’westerbork’, ’jews’, ’people’, ’rotterdam’, ’german’, ’camp’, ’germans’, ’war’, ’jewish’, ’happened’, ’nazis’, ’germany’
Christianity	’catholic’, ’church’, ’priest’, ’baptized’, ’communion’, ’religion’, ’jewish’, ’catholicism’, ’mother’, ’school’, ’convert’, ’never’, ’convent’, ’prayers’, ’think’
Deportations	’deported’, ’deportation’, ’people’, ’deportations’, ’jews’, ’day’, ’work’, ’ghetto’, ’happened’, ’heard’, ’think’, ’told’, ’believe’, ’sent’, ’knew’
Clothes	’shoes’, ’clothes’, ’shower’, ’hair’, ’shaved’, ’camp’, ’clothing’, ’naked’, ’put’, ’auschwitz’, ’pair’, ’showers’, ’barracks’, ’barrack’, ’women’
Packing	’take’, ’clothing’, ’little’, ’suitcase’, ’remember’, ’things’, ’knitting’, ’left’, ’made’, ’used’, ’yarn’, ’everything’, ’mother’, ’maybe’, ’something’
Judenrat	’judenrat’, ’judenrate’, ’ghetto’, ’police’, ’jewish’, ’people’, ’jews’, ’germans’, ’council’, ’orders’, ’killed’, ’away’, ’gestapo’, ’work’, ’town’
Gypsy	’gypsies’, ’gypsy’, ’camp’, ’block’, ’people’, ’auschwitz’, ’mengele’, ’lager’, ’saw’, ’jews’, ’gassed’, ’barracks’, ’eichman’, ’see’, ’told’

Topic Title	Top-15 words in topic
Kapo	’kapos’, ’kapo’, ’prisoners’, ’camp’, ’block’, ’barrack’, ’people’, ’german’, ’political’, ’killed’, ’work’, ’put’, ’prisoner’, ’somebody’, ’remember’
Poland	’warsaw’, ’polish’, ’lublin’, ’place’, ’train’, ’find’, ’army’, ’walked’, ’also’, ’anyway’, ’station’, ’says’, ’poles’, ’people’, ’praga’
Camp Barracks	’barrack’, ’barracks’, ’beds’, ’bunk’, ’people’, ’bunks’, ’slept’, ’camp’, ’wooden’, ’straw’, ’bed’, ’sleeping’, ’cold’, ’sleep’, ’little’
Death	’cemetery’, ’buried’, ’grave’, ’jewish’, ’people’, ’put’, ’died’, ’stone’, ’family’, ’find’, ’mass’, ’tombstones’, ’graves’, ’place’, ’stones’
Unknown	’people’, ’remember’, ’jewish’, ’see’, ’think’, ’father’, ’come’, ’knew’, ’little’, ’german’, ’camp’, ’day’, ’mother’, ’told’, ’thats

Tracing the deportation to define Holocaust geometries.
The exploratory case of Milan.

Giovanni Pietro Vitali, Laura Brazzo
Université de Versailles Saint Quentin en Yvelines - Université Paris-Saclay, Fondazione CDEC
(Contemporary Jewish Documentation Center)
47 Boulevard Vauban - 78047 Guyancourt Cedex, Piazza Edmond Jacob Safra, 20125 Milano MI, Italia
giovannipietrovitali@gmail.com, laurabrazzo@cdec.it

Abstract

This paper presents a pilot project conducted in collaboration with the Fondazione CDEC to shed light on the historical dynamics of the arrests and deportations of Jews from Italy to foreign concentration camps between 1943 and 1945. Led by a multidisciplinary team, including a Digital Humanities expert, an archivist, a GIS developer, and an education manager, the project aimed to rework archival information into data visualisation models utilising a subset of data from the CDEC LOD dataset of the victims of the Holocaust in Italy to construct detailed visual representations of deportation routes.

Drawing inspiration from previous projects like the Atlas of Nazi-Fascist Massacres and research on Holocaust testimonies, this project sought to create interactive maps, network and graphs illustrating the paths of forced transfers endured by arrested Jews, particularly focusing on those born or arrested in Milan. Despite challenges such as incomplete or imprecise data, the team managed to reconstruct deportation routes and classify transport convoys, enhancing the understanding of this dark period in history. The visualisations, along with detailed repositories and links provided on GitHub, serve as valuable research tools for both scholarly and educational purposes, offering users varying levels of granularity to explore historical events and timelines. Through meticulous data analysis and visualisation techniques, this project contributes to ongoing efforts to preserve and understand the tragic events of the Holocaust, emphasizing the importance of archival work and interdisciplinary collaboration in historical research.

Keywords: Data Visualisation, Spatial Humanities, Jewish Deportation

1. Introduction

This paper aims to take stock of a pilot project that seeks to address the need to put the historical dynamics that characterised the arrests of Jews, living in Italy (Italians and foreigners) and their subsequent deportation to concentration and extermination camps between 1943 and 1945, back at the centre of the scientific debate. After the Armistice the 8th of September 1943 and subsequent Nazi occupation of Italy, deportation of Jewish people started and more than 8,000 Jews were deported from Italy to Nazi concentration camps. This project was carried out in close collaboration with the Fondazione CDEC (Contemporary Jewish Documentation Center) and is the result of a work that starts with the reworking of archival data, also from the Fondazione CDEC, to their transformation into data visualisation models. All steps of this work were shared and discussed by a four-person working team:

- Giovanni Pietro Vitali, Associate Professor in Digital Humanities at the Université de Versailles Saint Quentin en Yvelines - Université Paris-Saclay who was in charge of coordinating the work, structuring the data, and creating a link between the archive activities and the final creation of the models.
- Laura Brazzo, vice director of the CDEC Foundation and head of its historical archives, who was in charge for the analysis and revision of the archive materials.
- Simone Landucci, GIS (Geographic Information System) developer who worked on the creation of the JavaScript models that enable the geographical visualisation of maps.
- Patrizia Baldi, education Activities Manager at CDEC, who supported the historical and critical reflection around the creation of models.

On the occasion of International Holocaust Remembrance Day, the work was officially published on the Fondazione CDEC website at this link: https://www.cdec.it/milano-mappe-sugli-arresti-e-le-deportazioni-degli-ebrei-1943-1945/.[1]

Before delving into previous projects that paved the way for GIS applications in the study of deportation and outlining the specificities of our work along with its potential in deportation analysis, it is important to note that this type of study faces significant challenges. This difficulty primarily stems from the near-complete destruction of transport lists containing the names of deported Jews.[2]

A concluding introductory remark pertains to technical matters. All technologies utilised in our endeavour are entirely open-source. The project

[1] These tools are currently available only in Italian. However, the development team of these visualisations intends to create an English-language version with the aim of expanding the scope of data to encompass the entire database of CDEC Foundation.

[2] Only two convoys left from Milan bound for Auschwitz, and one bound for Bergen Belsen. The other convoys which left Milan were directed to the concentration camps (Polizei und Durchgangslager) of Fossoli, Modena (from February until August 1944) and then to Bolzen (until December 1944). From Fossoli and then Bolzano most of the convoys departed for Auschwitz. The total number of convoys that left from the territories of the Italian Social Republic was 20 (Picciotto Fargion, 2002: 58-65).

development team has thoroughly annotated both the data and code, thereby facilitating its reuse by any individual requiring it.

2. State-of-the-art

This work has been possible reusing data that the CDEC Foundation made available through its endpoint (http://dati.cdec.it/lod/shoah/website/html). Via the CDEC endpoint the full dataset of Jewish people arrested and deported from Italy can be queried, downloaded and reused (CC License 4.0) (Brazzo and Rodriguez, 2019). Behind the creation of this dataset, as well as behind the work presented in these pages, there is clearly *Il libro della memoria. Gli Ebrei deportati dall'Italia (1943-1945)* by Liliana Picciotto (Picciotto Fargion, 2002) that represents the reference point for all studies dealing with Jewish deportation from Italy. This book, published for the first time in 1991, is the central reference of this paper. In this monumental work the full list of both Jewish deportees and victims of massacres in Italy is recorded. This list is made of more than 8000 individual records; data come from a variety of sources including already-mentioned four Nazi-list of transports[3] out of the 20 transports which left from the RSI (Picciotto Fargion, 2002: 58-61).[4]

One of the primary sources of inspiration for this project was the *Atlante delle stragi Nazifasciste* [Atlas of Nazi-Fascist Massacres] (http://www.straginazifasciste.it/), published in 2012. This research, funded by the German government, has been coordinated by the National Institute for the History of the Liberation Movement in Italy (INSMLI, http://www.reteparri.it/), and the National Association of Italian Partisans (ANPI, https://www.anpi.it/). The Atlas of Nazi-Fascist Massacres consists of an online database and related materials (documentaries, iconographic documents, videos) on some ~~specifical~~ specific historical recorded episodes: war massacres carried out by the Nazis and fascists principally during the German occupation period. The efforts dedicated to the development of this Atlas have culminated in the creation of a volume that stands as one of the pioneering endeavours in applying Spatial Humanities methodologies to conduct historiographic analysis of a significant phenomenon such as the spreading of violence during WWII: *Zone di guerra, geografie di sangue. L'Atlante delle stragi naziste e fasciste in Italia (1943–1945)* [War zones, blood geographies. An Atlas of the Nazi and Fascist massacres in Italy (1943-1945)] by Paolo Pezzino and Gianluca Fulvetti (Fulvetti and Pezzino, 2016).

Expanding upon the groundwork laid by CDEC's exploration of Holocaust testimonies, as well as Picciotto's research and the data compiled by the Ferruccio Parri National Institute's atlas of massacres in Milan, a significant study was released in 2021. For the first time, this paper interlinked wartime massacres and deportations, offering a comprehensive understanding of these intertwined historical events: *Visualizing Second World War Violence Through an Atlas of Nazi-Fascist Repression* by Giovanni Pietro Vitali (Vitali, 2021). In this essay, the author pioneers the use of data visualisation technologies and novel approaches to dataset creation, which serve as the foundation for the maps and networks presented in this paper.

Finally, it is important to mention previous scholarly efforts in addressing the challenge of visualizing deportation in Italy. Scholars such as Alberto Giordano, Tim Cole, and Maël Le Noc have undertaken significant work in this area. Their research focused on clustering Jewish arrests in Italy, culminating in the development of a model capable of tracing the movements of family members from the moment of arrest to their arrival at concentration camps (Le Noc *et al.*, 2020). Prior to this, Cole and Giordano collaborated with Anne Kelly Knowles on an innovative volume titled *GeoGraphies of the Holocaust* (Knowles *et al.*, 2014) which stands as a pioneering example of digitally assisted spatial humanities analysis within the field of Holocaust studies.[5]

This paper aims to present a project that aligns with prior research in Spatial Humanities and Deportation, with the objective of contemplating the interrogation of our understanding of the Holocaust phenomenon through archival sources and scholarly endeavours.

3. The project

The first goal of this project was to craft an elaborate map delineating the routes of involuntary displacements endured by Jews subjected to arrest and subsequent deportation to Nazi concentration and extermination camps. Utilising already available data, archival materials and data visualisation, the team sought to construct a research instrument primarily for scholarly endeavours, while also bolstering educational and training initiatives at CDEC

[3] The four transports are: (1) 5 April 1944/Auschwitz; (2) 16 May 1944/Bergen Belsen; (3) 16 May 1944/Auschwitz; (4) 26 June 1944/Auschwitz. They can all be viewed via the Digital Library at this link: https://digital-library.cdec.it/cdec-web/storico/search/result.html?query=Transportliste&titoloStorico=&contenutoStorico=&startDate=&endDate=&personeStorico=&luoghiStorico=&entiStorico=
The convoys departing from Trieste (Operationszone Adriatisches Küstenland, are numbered from 21T to 43T while the unique convoy departing from Rhodes is numbered 44R).

[4] RSI stands for Repubblica Sociale Italiana [Italian Social Republic], a puppet state under Nazi-German influence during the latter stages of World War II. Established following the German occupation of Italy in September 1943, RSI endured until the surrender of German forces in Italy in May 1945. The presence of German troops fuelled significant opposition across Italy, sparking widespread resistance and ultimately precipitating the Italian Civil War.

[5] In this volume, the third chapter in particular: Retracing the *'Hunt for Jews' A Spatial-Temporal Analysis of Arrests during the Holocaust in Italy* by Alberto Giordano and Anna Holian (Knowles *et al.*, 2014: pp. 53-86).

Foundation as well as at the Shoah Memorial of Milan. Therefore, the decision was made to launch this pilot project using data related to the arrests of Jews in Milan, encompassing both those born in Milan and arrested elsewhere across Italy. Recognising the unique aspects of the phenomenon under study - deportation - and its geospatial data representation, the project team deemed it essential to empower users of the visualisations with control over both the historical events and their chronological context.

The *Graphs, Maps and Trees* (Moretti, 2005), and networks approach that we proposed, recalling Franco Moretti's distant reading methodology (Moretti, 2013) applied to History. The graphical tools we have developed encompass various types of dataviz approaches - graphs, maps, trees and networks - and are designed to underscore the identified relationships between individuals, locations, and time periods, offering varying levels of detail granularity. This approach allows users to explore specific aspects based on their choices and research needs of the tools we have created. Through our work, the case studies we have conducted can be examined using both a close and distant reading approach, depending on the user's specific inquiries. Our underlying philosophy is to encourage users to 'ask the data graphically', enabling them to explore the information in a visual and interactive manner according to their preferences and research objectives.

The data sample we used for our project includes information on two case studies:

- **Jews born in the city of Milan** (and arrested in Milan or the rest of Italy), 166 persons
- **Jews arrested in the city of Milan** (born in Milan or the rest of Italy or Europe), 278 persons.

While these are two separate sets of data, there is an overlap of 41 cases between them, thus bringing the total number of cases examined to 403.

The two datasets were acquired by executing two separate SPARQL queries on the entire LOD dataset of the Victims of the Holocaust in Italy (http://dati.cdec.it/lod/shoah/website/html). The query criteria for the data were the birthplace in the first case and the place of arrest in the second case.

The dataset of the Victims of the Holocaust in Italy is the structured digital version of the textual information reported in *Il Libro della Memoria*, This volume is the result of a multi-year research conducted by CDEC, led by Liliana Picciotto, on the persecution and deportation of the Jews from Italy. The MS Access database that formed the basis of *Il Libro della Memoria* underwent subsequent processing starting in 2013: the data were massively transformed into RDF format based on established ontologies and a specifically crafted OWL domain ontology (*Shoah Ontology*) that formally describes the concepts and relationships proper to the persecution and deportation of the Jews from Italy.

The RDF dataset of the Victims of the Holocaust in Italy is exposed since 2014 on a SPARQL endpoint where queries, downloading and reusing of data by third parties are enabled. This dataset formally titled "Shoah Victim Names" was included in the LOD Cloud Diagram of the University of Mannheim in September 2014 (https://lod-cloud.net/).

The research on the Victims of the Holocaust in Italy relies on a diverse array of sources, meticulously documented in a dedicated introductory section of *Il Libro della Memoria*. Among these sources is the collection of handwritten paper cards created by CDEC in the early 1970s at the beginning of the research on Jewish deportations from Italy, used to record information about each of the deportees. Information was systematically gathered from sources of that time, including testimonies of the relatives of the victims or survivors.

These cards often provide granular details regarding specific locations – such as the address of residence and the exact place of arrest - which have not been included in *Il Libro della Memoria* or subsequent databases. Data about these locations have been incorporated into the two subsets used for our project.

The inclusion of such information (when available) has provided the precise georeferencing of where people lived and/or were arrested - whether it was at their own home, on the street, or at another location (e.g., at the home of friends or acquaintances who were hiding them).

In short, the two datasets upon which the project is based, incorporate information sourced from both the entire dataset of the Victims of the Holocaust in Italy queried via the SPARQL endpoint and the archival handwritten paper cards.

3.1 Repositories and links

The visualisations showcased in this paper have been archived in the repository space of Giovanni Pietro Vitali's GitHub profile: https://github.com/digitalkoine. The datasets, as well as the R and HTML, CSS and JavaScript code, along with the web pages developed, have been released under the MIT license within the repositories of this profile.

The initial visualisation is a graph illustrating the timeline of arrests carried out by Fascist and Nazi authorities against Jews in Milan during the specified period. The graph is accessible through the following link: https://digitalkoine.github.io/chronology_arrests_milan/.[6]

[6] GitHub repository: https://github.com/digitalkoine/chronology_arrests_milan

One of the most important issues of the deportation is the classification of the trains that took the Jews out of Italy. In the specific case of this paper, we used the numbering of convoys proposed by Liliana Picciotto with departure and arrival dates.[7] In order to eliminate any ambiguity surrounding the description of the transports, a dedicated website was created. On this site, we have included the information necessary to identify each of the convoys represented in the maps and network displays: https://digitalkoine.github.io/convogli_lager/.[8] Below the description of each convoy, the users will find a list of the deported Jews, categorised based on the two case studies. By clicking on the name of each deportee, the user is directed to their personal and persecution data, accessible via the LOD browser LodView (e.g., http://dati.cdec.it/lod/shoah/person/418/html).

In terms of visualisations, each of the two case studies consists of a map and a network, both interactive:

- **Jews born in the city of Milan**
  Map (map_milan_deportees):[9] https://digitalkoine.github.io/map_milan_deportees/;
  Network (network_milan_deportees):[10] https://digitalkoine.github.io/network_milan_deportees/;
- **Jews arrested in the city of Milan**
  Map (map_milan_arrested):[11] https://digitalkoine.github.io/map_milan_arrested/;
  Network (network_milan_arrested): https://digitalkoine.github.io/network_milan_arrested/.[12]

3.2 Data and code

The visualisations we have crafted arise from a thorough examination of the historical intricacies surrounding arrests and deportations during the Holocaust. To accurately depict each stage of every deportee's journey, bespoke models were tailored for the two case studies. The objective was to explore digital solutions, facilitating the application of the same methodology to all deportees in the CDEC endpoint in the future. The project team had engaged in prior discussions concerning the potentialities afforded by visualisation techniques and the nature of the data it possessed, including associated challenges. Through a comprehensive exploration of sources and available options, the team harmonised the data and selected appropriate coding methodologies. This process was conducted with a synergistic approach, ensuring the interoperability of data across all chosen programming environments, primarily R and JavaScript.

3.2.1 Data

The approach to data was a crucial aspect of our methodological framework. Through the SPARQL endpoint, we obtained the basic and necessary information to construct visualizations. In order to accurately depict the deportation routes, comprehensive data detailing each individual's journey across all its stages was indispensable.

In addition to biographical data, the pertinent information encompasses details pertaining to the various phases of each individual's journey, commencing from the moment of their arrest:

- First and last name
- Place and date of birth
- Address of residence
- Place and date of arrest
- Period of imprisonment (if any)
- Period of internment (ddmmyyyy to ddmmyyy) in prison and/or transit camp (until August 1944 at the camp of Fossoli di Carpi, and then at the camp of Bolzano)
- Place and date of departure of convoy
- Place and date of arrival of convoy

The data available are in some cases very precise, in others less so. For instance, sometimes the date of the arrest is only known with the month without the day. Or again, the perpetrators of the arrest are only known for 99 out of 278 cases. The provenance of these data is the milestone research of Liliana Picciotto, namely the already-mentioned *Il Libro della Memoria.* Information provided by this book are rich and complex: last known place of residence (e.g. Milan); place and date of arrest (e.g. Milan, 11 November 1943), perpetrators of the arrest (Fascist, Nazi) transfer(s) from prison to prison before deportation; Convoy Departure place and date, Arrival place and date; fate. This complex of information has been structured through both consolidated ontologies (FOAF, Event, BIO, Schema.org and DCMI, for biographical information) and the OWL domain Shoah Ontology (http://dati.cdec.it/lod/shoah/reference-document.html for data regarding persecution) in a large dataset that has been used for our pilot project. However, the mapping part of the project we were constructing necessitated a higher degree of precision, particularly concerning the dates and locations of arrests. This precision could, in certain instances, be deduced from the scrutiny of the already mentioned handwritten paper cards Consequently, it was feasible, in numerous instances, to revisit the archives and reconcile discrepancies within the data,

[7] A different numbering system to classify convoys to prison and extermination camps is proposed by Italo Tibaldi in *Compagni di viaggio* (Tibaldi, 1994) and in *Calendario della deportazione italiana* (http://www.associazioni.milano.it/aned/tibaldi_calendar.htm).

[8] GitHub repository https://github.com/digitalkoine/convogli_lager

[9] GitHub repository: https://github.com/digitalkoine/map_milan_deportees

[10] GitHub repository: https://github.com/digitalkoine/network_milan_deportees

[11] GitHub repository: https://github.com/digitalkoine/map_milan_arrested

[12] GitHub repository: https://github.com/digitalkoine/network_milan_arrested

or integrate with additional information. Our work has tried to remedy data gaps such as the case of dates. By going back into the archives, we managed to resolve certain conflicts or in other cases we reconstructed a plausible date that would allow us to digitally design the journey of each deportee. Where, for example, a date was only indicated by the year, we provided a possible option that would allow the map to function and that it was coherent with the rest of the data, especially the dates. All acts aimed at rectifying inconsistencies in cases with gaps in the dataset were resolved by relying on contextual historical data such as the dates of arrival and departure of convoys. We duly noted this particular aspect for each individual. Regrettably, it was not feasible to reconstruct all deportation routes.

- **For the Jews born in the city of Milan** (and arrested in Milan or elsewhere in Italy), our 166 persons dataset had to exclude 4 people: Gutenberger Elda, Pisetzky Dorotea, Spiro Eva, and Volterra, Nissim. From now on, we will refer to this dataset as **DS1**.

- **For the Jews arrested in the city of Milan** (and born in Milan or elsewhere in Italy or abroad), our 278 persons dataset had to exclude 19 people: Adler Oscar Zeliko, Adler Zora, Americano Carolina, Araf Marco, Percowiez Adolfo, Rabinoff Anna, Milul Isacco Gino, Dana Salomone, Dana Samuele, Guastalla Luciano, Lemberger Marcella, Lemberger Wolf, Lenghi Walter, Rosenbaum Elena, Foà Aldo, Romano Ferdinando Vittorio, Gutenberger Elda, Leoni Giulia, Voghera Augusta. From now on, we will refer to this dataset as **DS2**.

As a result, DS1 decreased from 166 to 162 and DS2 from 278 to 259 deportees.

DS1 is composed of fewer people but with more complex histories, and the errors multiply. This table provides a summary of the dataset fields that have undergone rectification. We incorporate accuracy information, which pertains to the entirety of the data scrutinised for each respective field outlined in the dataset, juxtaposed with the errors detected within each single field as delineated in the subsequent table lines:

Fields (original name Digital Library)	**Errors on 1782 fields**	**Accuracy**
dateOfBirth	3	0,0017
arrestPlace	8	0,0045
arrestDate	36	0,02
transferDate	85	0,048
convoyDepatureDate	16	0,0089
convoyArrivalDate	17	0,0095
labelToNaziCamp	16	0,0089
dateofDeath	56	0,031

Table 1: Table of the accuracy of data concerning Jews born in Milan - DS1

This table indicates that the data pertaining to Jews born in Milan and detained either inside or outside the city are notably more deficient, and also less numerous compared to DS2. Nevertheless, it is important to note that the aforementioned gaps in the data within the CDEC endpoint did not impede the possibility to devise methodologies for visualising the deportation routes of individuals affected by these discrepancies. However, this observation prompts consideration regarding the ongoing necessity for the analysis and categorisation of historical source material, particularly regarding the Second World War and, more specifically, the Holocaust (Vitali, 2022). It is evident that there exists a requirement to revisit primary sources and rectify the data through increased dialogue between archives' documentation and digital technologies.

With regard to the data concerning Jews arrested in Milan, it appears that the primary challenge encountered concerning these historical records lies in the dates of arrest, which exhibit the highest frequency of errors. The table presenting accuracy metrics illustrates the following:

Fields (original name Digital Library)	**Errors on 2849 fields**	**Accuracy**
dateOfBirth	1	0,00035
arrestsDate	35	0,013
detentionDate	7	0,0025
campDate	5	0,0017
convoyDepatureDate	1	0,00035

Table 2: Table of the accuracy of data concerning Jews arrested in Milan – DS2

It is of note that the entirety of errors within the reference dataset underscores the richness of information present in the most comprehensive dataset. Notably, all inaccuracies in the dataset pertain to dates. It is intriguing to observe that the dataset focusing on individuals arrested in Milan exhibits greater accuracy compared to the dataset encompassing a broader scope, such as individuals born in Milan but arrested beyond the Lombard capital. Upon examining the accuracy data, one discerns the intricacies involved in reconstructing the history of these deportees, likely attributable to the complexity of their movements.

The design and manipulation of the data were conducted concomitantly, not through a process of conforming the datasets to predetermined specifications, but rather by refining the programming to align the models with the data themselves - an established focal point determined by the project team. As the data underwent harmonisation, the coding evolved in tandem, adapting to accommodate the evolving datasets.

3.2.2 Code

Undoubtedly, the primary objective of the team engaged in this project was to develop interactive maps. The imperative was to craft an interactive map capable of addressing the intricacies inherent in the historical phenomenon of deportation. The sole viable approach to adhere to open-source principles involved utilising HTML, CSS, and particularly JavaScript.[13] The provided HTML code served the purpose of constructing interactive web maps that visualise historical data pertaining to the deportation of individuals from Milan or arrested in Milan during World War II. It leveraged several JavaScript libraries and plugins to achieve various functionalities. Firstly, the maps are built using Leaflet.js, a widely-used JavaScript library for creating interactive maps. This library is imported from a Content Delivery Network (CDN). Additionally, Turf.js, a JavaScript library for spatial analysis, is utilized for geospatial operations such as buffering and intersecting. To enhance user interaction and visualisation, several Leaflet plugins are incorporated. Leaflet.Coordinates is employed to display mouse coordinates on the map, while Leaflet.Search facilitates searching for specific features on the map. Leaflet.GeometryUtil enables geometric operations, and Leaflet.AlmostOver assists in handling mouseover events. Moreover, the maps design are enriched with features like multiple style layers, SVG shape markers, arrowheads for lines, and polyline decorators, all achieved through respective Leaflet plugins. OverlappingMarkerSpiderfier-Leaflet is employed to manage overlapping markers effectively. For user convenience, EasyButton.js is integrated to add customizable buttons to the map interface. Additionally, ISO8601.js aids in parsing ISO 8601 durations, crucial for handling time-related data. Furthermore, the map relies on Leaflet.TimeDimension, a plugin enabling time capabilities such as time sliders and animations. This allows users to visualise temporal aspects of the deportation data. The HTML code includes references to various GeoJSON files containing geographic data layers, such as birthplaces, arrest locations, detention camps, and convoy routes. Moreover, CSS files are utilized for styling map elements, while JavaScript files define map interactions, such as popups when hovering over features and searching for specific individuals. In summary, the HTML code amalgamates diverse JavaScript libraries and plugins to craft an engaging and informative web maps showcasing historical deportation data during World War II.

Concerning the adoption of R programming language, this one was selected due to the straightforwardness exhibited by specific packages in crafting interactive networks, such as visNetwork, or graphs, such as plotly. The simplicity inherent in programming networks and graphs with R was the rationale behind this selection.

The Graph's code utilises basically the plot_ly function to generate an elaborate area chart depicting comprehensive statistics on Jewish arrests in Milan between 1943 and 1945. This visualisation comprehensively illustrates diverse categories of arrests, encompassing those carried out by Italians and Germans, as well as arrests with unidentified authorship. The layout of the plot is meticulously customised to enhance technical aspects such as axis labels, title positioning, and annotation alignment, culminating in a visually refined and informative presentation.

The networks' code begins by loading essential libraries for data manipulation and visualisation, including visNetwork and tidyverse, followed by the importation of pertinent data on nodes, edges, and additional contextual information from CSV files. Subsequently, utilising the visNetwork function, the code generates a dynamic network visualisation, allowing for customizations in dimensions and the incorporation of explanatory text. Configuration options are set to enable user interaction, such as node highlighting, selection by identifiers or groups, and tooltip management. Finally, a randomised seed is introduced to ensure a consistent layout for the network visualization.

3.3 Graphs

The graph is a scatter plot with an underlying area, displaying the temporal trend of arrests for the different groups of individuals. Each group is represented by a different colour, and the area under the curve represents the total number of arrests up to that point. Annotations on the chart provide additional information about the temporal period considered and the source of the data used (information and legend on maps and graphs are provided in Italian only so far). In summary, the script generates an interactive chart that offers a visual representation of the statistics of Jewish arrests in Milan during World War II, divided into groups of arrest authors.

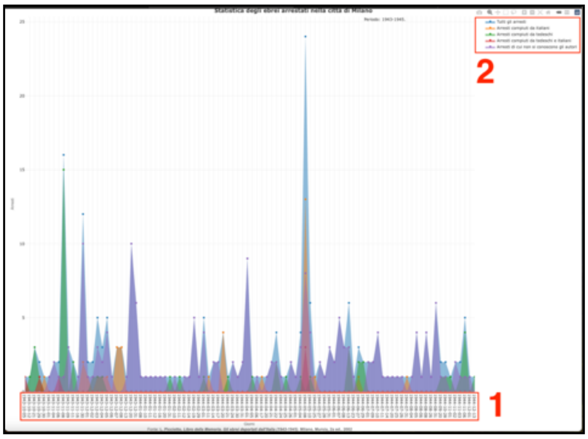


Figure 1: Interactive chronology of Jewish arrests 1943-45.

[13] In the current state of the art, there exists no readily available method to employ leaflet libraries in Python or R for effortlessly generating geometrically intricate time charts without resorting to services like shinyapps.

How to use the graph. This graph is fully interactive and through the zoom function allows you to select a chosen area that corresponds to a specific period. The dates of the Jewish arrests are on the x-axis (Number 1, Figure 1) and the number of arrests on the ordinates. The 5 lines in the graph correspond to the 5 options listed in the legend/option menu (Number 2, Figure 1). By clicking on the lines represented in the legend, one can highlight or hide the trends of arrests in Milan in order to visualise who the perpetrators are. More than one option can be selected and displayed at a time in order to make comparisons within the lines of arrest trends.

3.4 Maps

The two maps are web applications provide intuitive user interfaces enabling users to explore the data interactively. Information is organised into various layers, each representing a specific aspect of the deportees' lives, such as birthplaces, arrest locations, detention camps, and deportation routes. Each entity on the maps is associated with an informative popup providing details about the individuals or locations. Users can control which data layers are displayed and utilise a time function to explore the data over time. Furthermore, a search function allows users to look up specific individuals within the dataset. In summary, the maps offer an effective means of exploring and gaining a deeper understanding of the phenomenon of deportations from Milan during the Second World War.

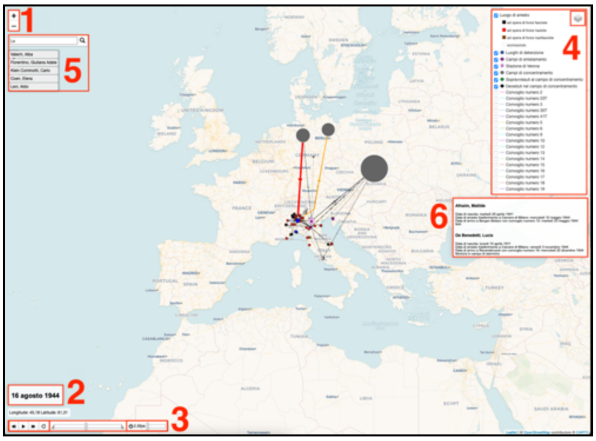


Figure 2: Interactive map of deportation explained with legends.

How to use the maps. The two maps are digital tools that can be browsed like any multimedia object. Figure 2 is an example representing both maps. It highlights the tools allowing the user to see the information he or she is looking for. The maps are dynamic and visualise the evolution of historical events day by day. They show the route of each individual person from the moment of his or her arrest at a certain location to his or her arrival in the concentration and/or extermination camp. On the screen, the places of passage that characterised the persecution route of these people are constantly highlighted and traced. Each point or line, which 'switches on and off' with the passage of time, can be clicked on to display information about the person it represents. These are the instructions for using the maps schematised in Figure 2:

(1) Zoom tool and positioning on the map

(2) Date box: The date constantly changes based on the passage of time given by the time bar.

(3) Time bar: This is the tool for starting and stopping the flow of time. The dates displayed day by day have been defined by default between 8 September 1943 and 25 April 1945. As can be seen from the image, the time bar is divided into three groups of buttons; each block corresponds to a different functionality:

- In the first group, the play/rewind/fast forward buttons allow the user to start/stop or advance the time scroll.

- The cursor in the middle block allows the user to manually move the time or date displayed forwards or backwards. Using the play/rewind/fast forward buttons, it is possible to manually move forward or backward the time and thus the events that the map represents.

- In the right-hand group of the time bar, a slider allows the user to change the speed of time scrolling.

(4) Legend/options menu: the legend explains the meaning of lines and dots and colours that are displayed on the map. It is opened by clicking on the layer selector icon. The legend also functions as an options menu: it enables lines of convoy trajectories and/or the places that marked each person's persecution route to be viewed at the same time.

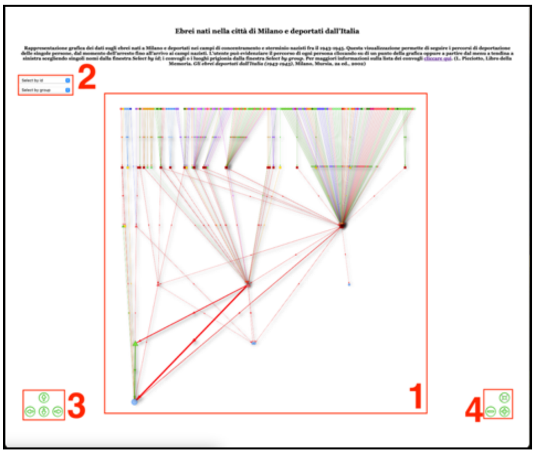
(5 & 6) Search/Selection Tool: allows the user to select the persecutory journey of a single person, from those represented on the map. By typing, letter by letter, the searched name, the tool gradually suggests one or more options. The biographical information on the chosen person(s) is displayed in the box marked No. 6. At the same time, the line identifying the person(s) on the card changes colour:

- turns orange if on the date on the card the deportation of the person has not yet taken place

- turns red if on the date indicated on the card by the time bar the person has already been deported.

3.5 Networks

The networks allow users to explore the deportation routes of individual persons from the moment of their arrest to their arrival at Nazi camps. Users can interact with the visualisation by clicking on nodes to highlight the deportation routes of specific individuals. Additionally, a dropdown menu enables users to select specific names, convoys, or places of imprisonment for further exploration. The aim of this visualisation is to provide insights into the deportation of Jews from Milan during the specified period. It serves as a tool for understanding the historical context and human impact of these events.



Ebrei nati nella città di Milano e deportati dall'Italia

Rappresentazione grafica dei dati sugli ebrei nati a Milano e deportati nei campi di concentramento e sterminio nazisti fra il 1943-1945. Questa visualizzazione permette di seguire i percorsi di deportazione delle singole persone, dal momento dell'arresto fino all'arrivo ai campi nazisti. L'utente può evidenziare il percorso di ogni persona cliccando su di un punto della grafica oppure a partire dal menu a tendina a sinistra scegliendo singoli nomi dalla finestra *Select by id*; i convogli e i luoghi prigionia dalla finestra *Select by group*. Per maggiori informazioni sulla lista dei convogli cliccare qui. (L. Picciotto, Libro della Memoria. Gli ebrei deportati dall'Italia (1943-1945), Milano, Mursia, 2a ed., 2002)

Figure 3: Interactive network visualisation of deportation explained with legends.

How to use the networks. The network analysis designed for this research are developed in the form of dendrograms and allow a further visualisation of the deportees' journeys and the relationships established between them and the places that characterised their deportation. Both network graphs are fully interactive, which means that by clicking on any element of the graph this element lights up, highlighting all the other points connected to it. Figure 3 schematises how to act with both dendrograms:

(**1) Development area of the graph**, nodes and edges are displayed in this area. The aligned dots at the top of the graph represent each deportee, the squares the first places of imprisonment (e.g. police stations or penitentiaries - in the network of those born in Milan), the diamonds the places of arrest (in the network of those arrested in Milan), the triangles the internment camps (*Polizei- und Durchgangslage* e.g. Fossoli or Bolzano) and finally the blue dots at the bottom of the network are the extermination and/or Concentration camps (Auschwitz, Bergen Belsen, Buchenwald, Ravensbrük). By clicking on each point/person, lines light up representing the path of persecution and deportation suffered by that person. Clicking on a point that defines a place will 'light up' all those representing the people who were arrested in that place. By hovering the mouse over each point, it is possible to view the information associated with that place or person via a pop-up.

(2) Drop-down menus, a tool enabling the focused display of places, people (id menu), or convoys (group menu). Selecting an item from these menus will illuminate it and all others connected to it. For example, in the case of arrestees in Milan, clicking on an arrest location will highlight all those arrested there. If, on the other hand, the user clicks on a specific convoy, he or she will immediately be able to see all the people who were transported there.

(3) Arrow buttons: with this tool you can navigate the network left and right, just like with a mouse.

(4) Graph zoom options.

4. Conclusions

This dataviz project is a valuable example of the potential unlocked by making research data available for public reuse. In this case, data about the victims of the Holocaust in Italy have been used to trace the movements of each individual and to test geo-mapping and networks tools. These case studies highlight how far we still are from obtaining a complete view of the events leading up to deportation. The lack of data primarily concerns the events that occurred before their arrest. This renders our map slightly incomplete when attempting to trace the entire pathway of each individual from his/her birthplace to the deportation site. For instance, information concerning the arrests of Jews in Milan during this period underscore the imperative for sustained research efforts aimed at augmenting the available data. As depicted in the ensuing graph, the identity of the perpetrator remains undisclosed for the majority of arrests conducted in Milan throughout this timeframe.

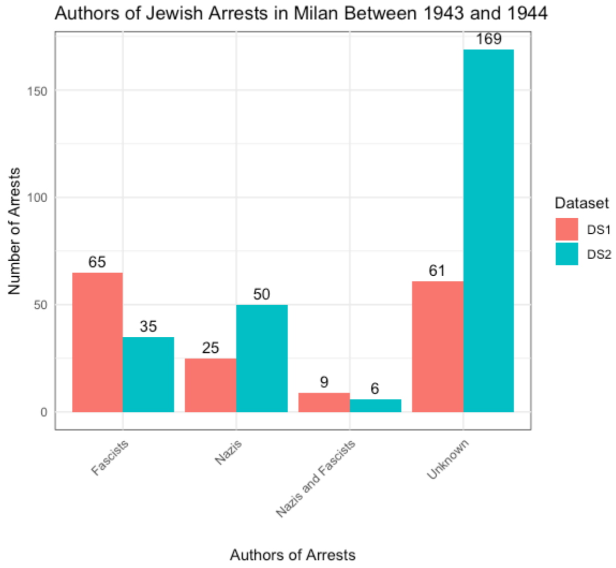


Figure 4: Authors of Jewish Arrests in Milan Between 1943 and 1944

Nevertheless, this project offered the opportunity to utilize sets of information such as addresses of residence or exact places of arrest, which allow us to at least begin the schematization of both the modalities of the arrests (house-to-house, for example, or by surveillance operations) and the urban areas where people were arrested.

The collection of data and the digital tools we have created during the research help us, again, to take stock of the state of the art of what we know, and impose new challenges on us, such as trying to fill the information gap of the 61 (DS1) + 169 (DS2) cases of arrests of unknown perpetrators.

Despite this partial incompleteness of the data, our geographical and network model succeeds in giving new insight into the dynamics of deportation experienced by the people included in our two case studies. The visualisations our team has developed make possible a comprehensive overview of the origins, forced movements, and final destinations for each of the individuals arrested (whether in Milan or elsewhere in Italy), deported, and subsequently annihilated in concentration and extermination camps. The maps, along with the associated graphs, allow for a deeper understanding of the places where Jews were rounded up and of the way the Nazis organised their transports to concentration and extermination camps. The combination of space and time in relation to individuals - never before tested in such a manner - makes this project relevant for both educational and research purposes. Such overviews and details at the same time would not have been possible, neither querying data from the endpoint neither browsing through over a thousand pages of the *Il Libro della memoria*.

It is worth noting that the data about the Victims of the Holocaust (in Italy as well as elsewhere) are still subject to continuous revisions, with integration or correction of previously provided information. Consequently, data on Jews deported from Italy, available on the CDEC SPARQL endpoint, may vary over time due to possible further revisions. To mitigate discrepancies between the data presented in our maps and graphs and the information obtained by querying the SPARQL endpoint, one of our future objectives is to establish a direct linkage between the tools and the data on the endpoint. The other objective is certainly applying our model to the entire dataset of the Victims of the Holocaust in Italy.

In conclusion, deportation was a tragic page in the history of the twentieth century that took on the scope of a collective memory (Erll *et al.*, 2008: 109-118) and at the same time represented the individual drama of thousands of people. This project aims precisely to unite collective and individual memory of the Holocaust through the use of the digital; considering how these tools, and data visualisation in particular, can play an important role in recalling the collective memory (Koçak, 2017). In our maps and networks, the user can witness the global phenomenon represented by the deportation but, at the same time, can go and find each deportee one by one and follow what was the terrible ordeal they went through individually. The maps and the networks of this project were designed precisely to capture this dichotomy between the history of the individual and that of the many. We wanted to give scholars and the general public the possibility both to read/browse the general phenomenon of deportation in a distant reading approach and to be able, by zooming in or by a simple name search in the sidebar, to follow each individual deportee in a close reading approach.

The dynamism of our models and its capability to combine distant and close reading approaches demonstrate how the digital can create a bridge between the historical representation of collective and individual memories.

5. Bibliographical References

Brazzo, L., and Rodriguez, K. J. (2019) Data Sharing, Holocaust Documentation and the Digital Humanities: Best Practices, Case Studies and Benefits. *Umanistica Digitale* (4). *https://doi.org/10.6092/issn.2532-8816/9035* [Accessed 8 September 2022].

Erll, A., Nünning, A., and Young, S. B. eds. (2008) *Cultural memory studies: an international and interdisciplinary handbook*. Berlin ; New York: Walter de Gruyter.

Fulvetti, G., and Pezzino, P. eds. (2016) *Zone di guerra, geografie di sangue: l'Atlante delle stragi naziste e fasciste in Italia: (1943-1945)*. Bologna: Società editrice Il mulino.

Anne Kelly Knowles, Tim Cole, and Alberto Giordano eds., Geographies of the Holocaust (2014). Indiana University Press Available at: *https://www.jstor.org/stable/j.ctt16gzbvn*. [Accessed 25 February 2024].

Koçak, D. Ö. (2017) Collective Memory and Digital Practices of Remembrance. In: Friese, H., Rebane, G., Nolden, M., and Schreiter, M. (eds.) *Handbuch Soziale Praktiken und Digitale Alltagswelten*. Wiesbaden: Springer Fachmedien Wiesbaden *https://doi.org/10.1007/978-3-658-08460-8_36-1* [Accessed 3 March 2024].

Le Noc, M., Giordano, A., and Cole, T. (2020) The Geography of the Holocaust in Italy: Spatiotemporal Patterns of Arrests for Families and Individuals and a Conceptual Model. *The Professional Geographer* 72, (4) 575–585. *https://doi.org/10.1080/00330124.2020.1758572*.

Moretti, F. (2005) *Graphs, maps, trees: abstract models for a literary history*. London ; New York: Verso.

Moretti, F. (2013) *Distant reading*. London ; New York: Verso.

Picciotto Fargion, L. (2002) *Il libro della memoria: gli ebrei deportati dall'Italia (1943-1945)*. 2. ed. Milano: Mursia.

Tibaldi, I. (1994) *Compagni di viaggio: dall'Italia ai lager nazisti: i trasporti dei deportati, 1943-1945*. Milano: F. Angeli.

Vitali, G. P. (2021) Visualizing second world war violence through an Atlas of Nazi–Fascist Repression. *Digital Scholarship in the Humanities* fqab070. *https://doi.org/10.1093/llc/fqab070*.

Vitali, G. P. (2022) Storia contro Narrazione, progetti e futuri digitali per la storiografia della seconda guerra mondiale. *Occupied Italy* 2, (2) 13.

Zero-shot Trajectory Mapping in Holocaust Testimonies

Eitan Wagner[†] **Renana Keydar**[‡] **Omri Abend**[†]
[†] Department of Computer Science [‡] Faculty of Law and Digital Humanities
Hebrew University of Jerusalem
{first_name}.{last_name}@mail.huji.ac.il

Abstract

This work presents the task of Zero-shot Trajectory Mapping, which focuses on the spatial dimension of narratives. The task consists of two parts: (1) creating a "map" with all the locations mentioned in a set of texts, and (2) extracting a trajectory from a single testimony and positioning it within the map. Following recent advances in context length capabilities of large language models, we propose a pipeline for this task in a completely unsupervised manner, without the requirement of any type of labels. We demonstrate the pipeline on a set of ≈ 75 testimonies and present the resulting map and samples of the trajectory. We conclude that current long-range models succeed in generating meaningful maps and trajectories. Other than the visualization and indexing, we propose future directions for adaptation of the task as a step for dividing testimony sets into clusters and for alignment between parallel parts of different testimonies.

Keywords: Mapping, Trajectory, Testimonies, Holocaust

1. Introduction

The location trajectory, i.e., the sequence of locations in which the story takes place, is an essential aspect of a story. The significance of location in a story is crucial, as placing a story in a specific setting is often seen as a defining characteristic that sets narrative texts apart from other types of writing (Piper and Bagga, 2022).

However, despite the abundance of Natural Language Processing (NLP) research on describing locations in texts, few efforts have been made to extract the progression or sequence of locations from a narrative story (Wagner et al., 2023). As a structured prediction task with a large class set, the ability to obtain data that is sufficient for generalization is very limited.

In this work, we present the task of zero-shot trajectory mapping and design a pipeline for it with long-context large language models. Zero-shot trajectory mapping involves both the extraction of the locations for each document (as a "trajectory") and the identification of the relationship between the locations (creating a "map"). We have no prior list of locations and the map is constructed based on the given texts only. Thus, the task is unsupervised in two ways – the set of locations must be inferred from a set of unannotated texts, and the trajectory of each text must be extracted without supervision.

Our research primarily centers on transcribed Holocaust survivor testimonies, which are provided in English. The significance of this dataset in the examination and remembrance of the Holocaust cannot be emphasized enough. As the last surviving witnesses inevitably pass away, there is an urgent necessity to find new approaches to engage with the extensive collection of Holocaust testimonies housed in records. Utilizing NLP technology for the analysis of these testimonies has recently been strongly recommended (Artstein et al., 2016; Wagner et al., 2022). By leveraging NLP, researchers can extract valuable insights from the vast array of testimonies (comprising tens of thousands) instead of limiting themselves to small-scale, predominantly manual studies. Additionally, we assert that Holocaust testimonies hold distinct value for NLP due to the combination of a multitude of accounts within a relatively confined domain of topics and locations. This quality sets them apart from typical narrative datasets (Sultana et al., 2022).

We describe and run a full pipeline for zero-shot trajectory mapping, using GPT4. [1] We show the resulting map on ≈ 75 testimonies and provide examples of the trajectories on this map. Based on the resulting maps, we describe future directions for alignment between testimonies.

Trajectory extraction is valuable for visualization and trajectory clustering (Bian et al., 2018). Characterizing a story by a sequence of locations is also beneficial as a backbone for alignment between different stories–an important task in its own right (see, e.g., Ernst et al., 2022). In general, successful location extraction indicates aspects of long-range narrative understanding, which is a highly active field in NLP (Yao et al., 2022; Bertsch et al., 2024).

2. Previous Work

Narrative Analysis. Narrative schema analysis aims to capture the core of event sequences, providing a condensed sequential timeline of a lengthy story. This overview helps in aligning relevant parts

[1] https://openai.com/research/gpt-4

and identifying common topic paths, as demonstrated by Antoniak et al. (2019) in their study on birth stories using segment-wise topic modeling.

To extract an interpretable sequential progression it was assumed necessary to divide the long story into shorter segments (Wagner et al., 2023). However, recent advances in NLP introduced significant increases in context lengths of models (Wang et al., 2024), allowing the extraction of sequences as an end-to-end task.

Recent studies have highlighted the importance of event locations in narrative analysis. Piper et al. (2021) provided a definition of narratives that included a focus on event locations. Soni et al. (2023) introduced a task involving grounding characters in specific locations. Kumar and Singh (2019) extracted event locations from individual events, such as those found in tweets. Wagner et al. (2023) expanded on this concept by examining trajectories of locations throughout entire narratives, utilizing a predetermined set of coarse-grained categories.

Trajectory Modeling in Transportation. Some works seek to extract document-level trajectories in transportation. Mathew et al. (2012) applied Hidden Markov Models (HMM) to human location trajectories. Sassi et al. (2019) utilized convolutional neural networks on location embeddings as an alternative to HMMs. Lui et al. (2021) employed LSTM-based models for predicting pedestrian trajectories. These works focus on locations given as coordinates and not as natural-text descriptions, which allow for a more thematic level of representation and comparison (Wagner et al., 2023).

Narrative Cartography Many works investigated the mapping of narratives. Reuschel and Hurni (2011) presented methods for the visualization of location maps. Their methods show differences between the maps in fiction and non-fiction. Mai et al. (2022) develop toolboxes for enrichment of geographic data, based on knowledge graphs.

These works are primarily based on a location ontology, thus limiting the scope to domains with sufficient prior knowledge. In our work, we propose a completely unsupervised method, allowing its application without any prior knowledge.

3. Task Definition

Our setting is the following: given a set of texts $\mathbf{x}^1, \mathbf{x}^2, \dots \mathbf{x}^k$, each divided into initial segments, $\mathbf{x}^i = x^i_1, x^i_2, \dots, x^i_n$, we wish to predict: (1) one directed graph $G = (V, E)$, where the vertices V are all the locations (name+type) in the set of texts and the edges E are the relationships between them (e.g., New York is in the United States); (2) for each $\mathbf{x}^i$, a path on the graph G, describing the trajectory in this text. The path should have additional vertex labels for the indices within the text of this location (e.g., segments 17-21) and edge labels for the method of transportation, if applicable (e.g., "by foot", "by plane" etc.). Roughly, we can say that the first part of the task corresponds to the creation of a "map" and the second part corresponds to the action of "mapping" within it.

It is instructive to compare this task to traditional Named Entity Recognition (NER) for location categories. NER is a prediction task at the phrase level that ignores the relationship between different locations or even between mentions of the same locations. Therefore, the first part of out task can be seen as a combination of NER and Entity Relation Extraction (focusing on the containment relation). The second part of our task is completely different as it requires a structured sequence as an output. Prediction is at the document level, with possible dependencies throughout the entire document. This property requires strong long-context capabilities which were not necessary for traditional NER.

3.1. Data

Our main data consists of 1000 Holocaust survivor testimonies, received from the Shoah Foundation (SF).[2] All interviews were conducted orally by an interviewer, recorded on video, and transcribed as time-stamped text. The lengths of the testimonies range from 2609 to 88105 words, with a mean length of 23536 words.

We note that the SF testimonies are divided into segments and contain highly detailed labels. Due to the extremely large set of labels we opted to use the text only and attempt zero-shot inference only. We arbitrarily chose a set of 74 testimonies and run them through our pipeline.

4. Zero-shot Trajectory Extraction

Recent advances in LLMs lead to a substantial increase in the context window that can be inputted into the models [3]. This makes it possible to input a whole testimony and perform location tracking as an end-to-end task.

For this we used GPT4-turbo-preview, which has a context length of $128K$ tokens [4]. The price for the experiment was ≈ 60 $.

We remark that the end-to-end task differs from supervised location tracking (Wagner et al., 2023) in multiple aspects: (1) Zero-shot extraction is not limited by granularity – it extracts countries, cities,

[2] https://sfi.usc.edu/

[3] https://www.anthropic.com/news/claude-2-1

[4] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

and also different types of locations (like "the forest") (2) Zero-shot extraction considers only locations that are mentioned in the text. This is also a limitation since different texts might be more specific with the names that are mentioned, leading to a longer trajectory.

4.1. Pipeline

Our pipeline is constructed of 4 steps: per-testimony location-graph extraction, per-testimony path extraction, combining all graphs, and visualizing each path in the combined graph.

Here we describe the details for each step.

Per-testimony location-graph extraction. For each testimony, we first extract a graph of the mentioned locations and their relationships.

We used gpt-4-turbo-preview with highly detailed instructions. The prompt was the following:

> I'll give you a Holocaust testimony.
> I want you to give me a JSON representing the graph of the mentioned locations (proper and common) and any known relations between them. Locations can be GPEs (like country or city) or significant facilities (like army camps, ghettos, concentration camps and death camps).
> Some important points:
> 1. Make sure the nodes contain locations only and not anything else (no nodes for events or people).
> 2. Give the nodes a type based on the type of location. The types should include: City, Country, Village, Ghetto, Army Camp, Concentration Camp, and Death Camp.
> 3. Keep the graph as full as possible, so, for example, if a place in a city in country is mentioned, there should be nodes for the place, the city, and the country. Separate a district from a city description into two nodes.
> 4. The graph should include relations between locations (i.e., A is in B). Make sure that the direction of an edge is that of inclusion if relevant (that is, if A is in B then the edge should be from A to B).
> 5. Make sure to avoid double entries.
> 6. Give me the graph as JSON dictionary, with a the "nodes" field indicating a list of nodes and "edges" indicating a list of edges. These nodes and edges should be in a format that can be create a python networkx graph. Make sure the nodes are given as a list of tuples, in which the first value is the name and the second is a dictionary with the type (as described above) The edges should be in a list of tuples, each containing two names (see example).
>
> Here is an example (from a different testimony):
> ‘‘‘json
>
> "nodes": <Here we provide an example list of locations>,
> "edges": <Here we provide an example relations between the locations>
>
> ‘‘‘
>
> This should all be based on the text.
>
> Testimony: <Here we add the testimony divided into numbered segments>

Per-testimony trajectory extraction. Following the answer about the locations, another request is made with the following prompt:

> Now, can you give a graph with the trajectory of the witness' movements? That is, give a list of location where he is. All location nodes should be nodes from the networkx graph you gave before. The nodes should have a field noting the sentence number in the text in which the witness was in that location.
> The edges should be between each adjacent node by order of the testimony. For each edge, add the method of transportation can be inferred from the text. Methods include: By foot, By car, By train, By plane. If the method is unknown give Unknown.
> Give me a graph in JSON format (like in the example).
>
> For example:
> ‘‘‘json
> "nodes": <Here we provide an example list of locations with their place in the testimony>,
> "edges": <Here we provide an example relations between adjacent locations, with the method of transportation>
> ‘‘‘

Combining the graphs into a map. To combine the obtained graphs into one global map, we first need to make sure that each location has only one label. Once we have one name per node, we can use the name as the identifier and create a

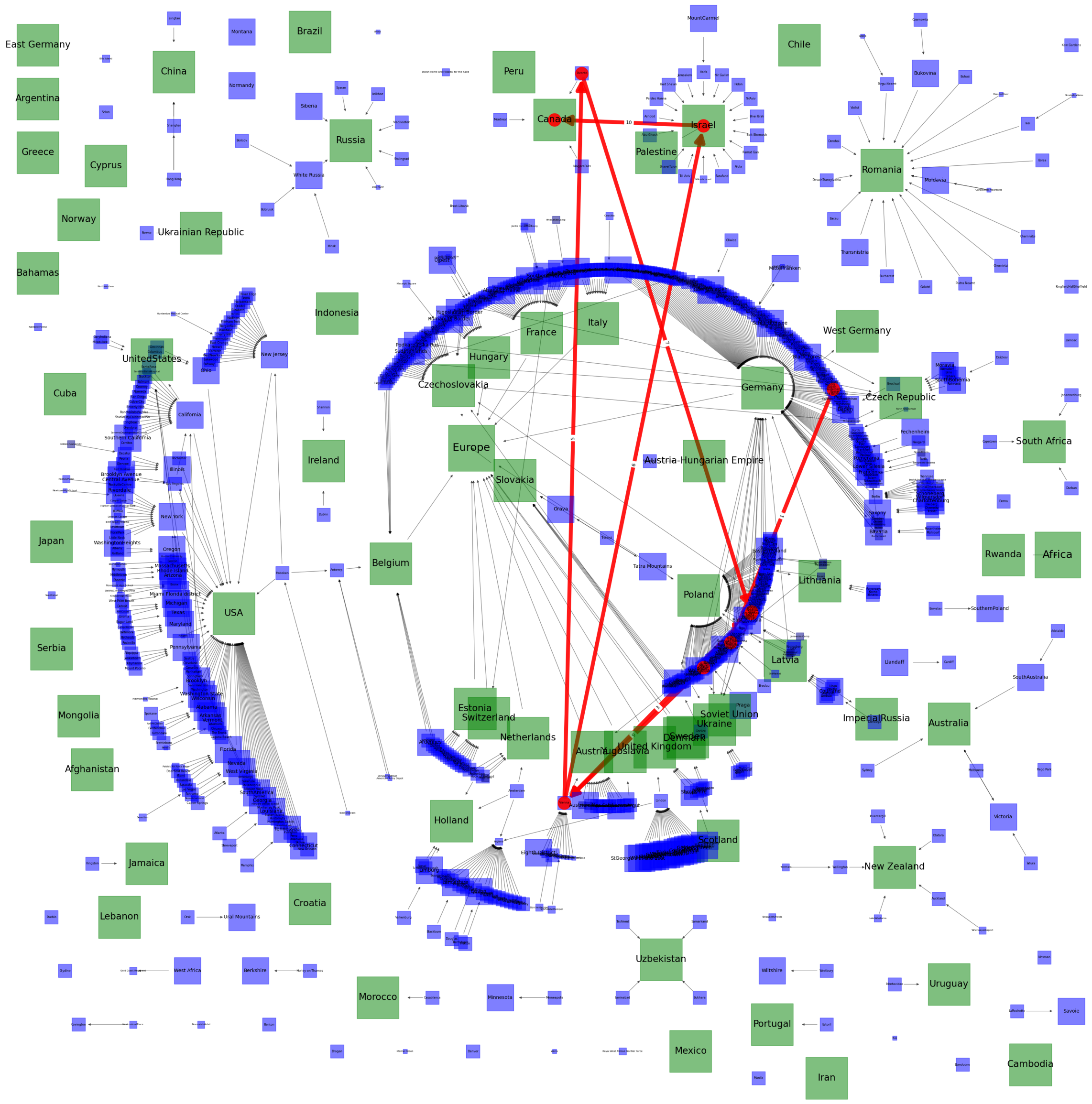


Figure 1: An example location map with a path extracted from a single testimony. Green nodes represent countries and blue nodes represent smaller locations. The path is in red.

graph with the new set of names and with all edges (removing doubles).

To create a list of double names, we again used GPT4.

We used the following prompt:

> I'll give you (in JSON format) a list of place names. I want you to see if there are any places that appear twice but with different names.
> Give me a JSON with a list of lists, where the inner list is the multiple names that describe the same place (and both appear in the input). No need to return unique names (i.e., lists with one element).
> Convert names only if you are positive that they are the same, e.g., different spellings or a longer description of the same place (like US, USA, America etc.).
> Make sure to maintain the exact spelling that appeared, including special characters. Make sure to give only the JSON format with no additional text.
>
> For example, if the input is:
> ‘‘‘json

<Here come some examples of lists of names describing the same place> “‘

Here is the input:
<Here comes a sorted list of the locations>

We manually proofed the resulting list leading to minor changes.

After aligning the node names, all nodes and edges were used to create a large map. We note that we applied some simple heuristic rules to sparsify the edges – we discarded edges between nodes from the same type (e.g., no edge from country to country) and edges that went against the type hierarchy (i.e., we discarded edges from country to city or from continent to country).

Plotting the maps and paths With the graphs and paths that we obtained, we used the Networkx[5] package for visualization.

4.2. Results

Here we present the statistics of the outputs and some examples of the resulting maps and paths.

We ran the pipeline on 74 testimonies from the Shoah Foundation. The average number of locations extracted from each testimony was ≈ 23 nodes and the average number of relationship edges was ≈ 17. The resulting graph had 883 nodes and 838 edges. The average length of the trajectories was 11.

In Figure 1 we display a view of the full map and a trajectory on it. Countries were enlarged for readability. In Figures 2 and 3 we show snippets around specific countries.

5. Future Work

The outputs from our pipeline can be useful on their own, such as for visualization or indexing. Moreover, the obtained map has theoretical qualities that can be further developed for additional uses.

For a pair of locations, we can define meaningful similarity measures that are based on the graph. For example, we use the distance from a common ancestor (so that two towns in Poland will be closer to each other than to a city in USA). In addition, since we extracted the types of locations, we might want to put special emphasis on Holocaust-specific locations (like ghettos and camps).

Provided with a point-wise distance measure (i.e., the distance between two locations) we can derive a trajectory-wise distance. For example, we can use Dynamic Time Warping (Vintsyuk, 1968) built upon the point-wise distance. This type of measure has the benefit of generating an optimal alignment between the trajectories, which in itself can be highly beneficial for Holocaust studies.

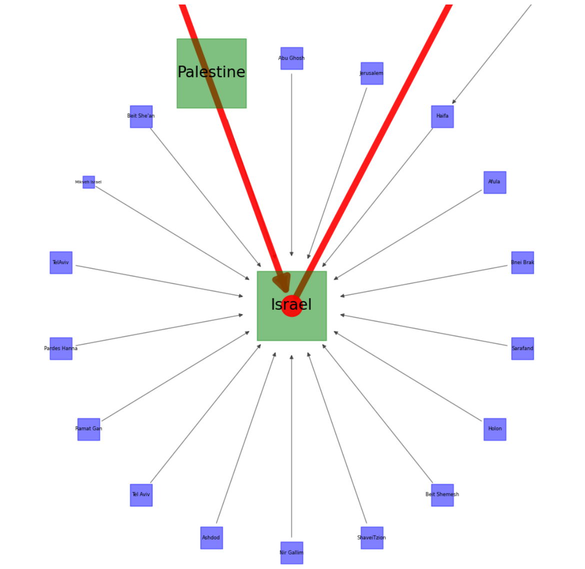


Figure 2: Snippet from the map that included Israel and locations within it. The path is from a trajectory going through Israel.

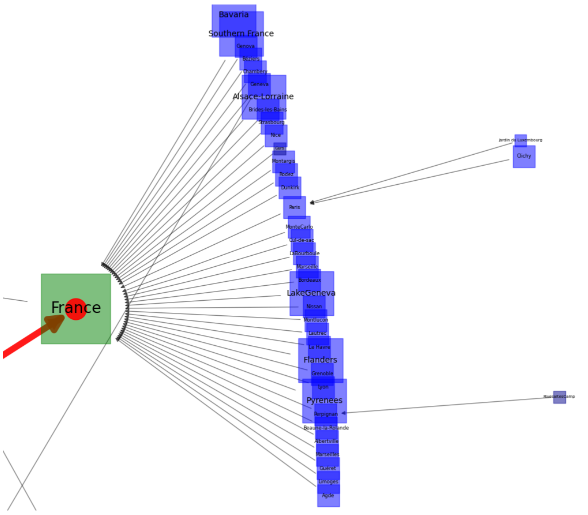


Figure 3: Snippet from the map that included France and locations within it. The path is from a trajectory concluding in France.

Providing a distance measure also allows us to perform unsupervised clustering based on the trajectories. The ability to cluster and align between testimonies has important implications for Holocaust research.

6. Conclusion

We presented and defined the task of zero-shot trajectory extraction. We built and demonstrated a pipeline for the task, based on GPT4. Our demon-

[5] https://networkx.org/

stration shows that the new models are capable of extracting meaningful trajectories from full testimonies without the necessity to break them into segments. These results suggest new ideas both for computational narrative analysis and specifically for Holocaust research.

Ethical Considerations

We followed the guidelines given by the archive. Although so the testimonies were not given anonymously, no identifying details will be included in our analysis. Our codebase and scripts will be released, but they will not contain any data from the archives. The data and trained models used in our work will not be shared with third parties without the archives' consent. To browse and research the testimonies, permission can be requested from the SF archive.

Acknowledgments

The authors acknowledge the USC Shoah Foundation - The Institute for Visual History and Education for its support of this research. This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Alfred Landecker Foundation.

7. Bibliographical References

Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Comput. Surv.*, 54(5).

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Ron Artstein, Alesia Gainer, Kallirroi Georgila, Anton Leuski, Ari Shapiro, and David Traum. 2016. New dimensions in testimony demonstration. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 32–36, San Diego, California. Association for Computational Linguistics.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2024. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Jiang Bian, Dayong Tian, Yuanyan Tang, and Dacheng Tao. 2018. A survey on trajectory clustering analysis.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Federico Dassereto, Laura Di Rocco, Giovanna Guerrini, and Michela Bertolotto. 2020. Evaluating the effectiveness of embeddings in representing the structure of geospatial ontologies. In *Geospatial Technologies for Local and Regional Development: Proceedings of the 22nd AGILE Conference on Geographic Information Science 22*, pages 41–57. Springer.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-level clustering for multi-document summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1765–1779, Seattle, United States. Association for Computational Linguistics.

Evelyn Gius and Michael Vauth. 2022. Towards an event based plot model. a computational narratology approach. *Journal of Computational Literary Studies*, 1(1).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, and Annekea Schreiber. 2021. Shared task on scene segmentation@ konvens 2021.

Xuke Hu, Yeran Sun, Jens Kersten, Zhiyong Zhou, Friederike Klan, and Hongchao Fan. 2023. How can voting mechanisms improve the robustness and generalizability of toponym disambiguation? *International Journal of Applied Earth Observation and Geoinformation*, 117:103191.

Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2022. Location reference recognition from texts: A survey and comparison. *arXiv preprint arXiv:2207.01683*.

Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1287–1296, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Mayank Kejriwal and Pedro Szekely. 2017. Neural embeddings for populated geonames locations. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 139–146. Springer.

Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, 33:365–375.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Andrew Kwok-Fai Lui, Yin-Hei Chan, and Man-Fai Leung. 2021. Modelling of destinations for data-driven pedestrian trajectory prediction in public buildings. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1709–1717.

Gengchen Mai, Weiming Huang, Ling Cai, Rui Zhu, and Ni Lao. 2022. Narrative cartography with knowledge graphs. *Journal of Geovisualization and Spatial Analysis*, 6(1):4.

Wesley Mathew, Ruben Raposo, and Bruno Martins. 2012. Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 911–918, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Semi Min and Juyong Park. 2019. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025.

Pinelopi Papalampidi, Kris Cao, and Tomas Kocisky. 2022. Towards coherent and consistent use of entities in narrative generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17278–17294. PMLR.

Andrew Piper and Sunyam Bagga. 2022. Toward a data-driven theory of narrativity. *New Literary History*, 54(1):879–901.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Anne-Kathrin Reuschel and Lorenz Hurni. 2011. Mapping literature: Visualisation of spatial uncertainty in fiction. *The Cartographic Journal*, 48(4):293–308.

Abdessamed Sassi, Mohammed Brahimi, Walid Bechkit, and Abdelmalik Bachir. 2019. Location embedding and deep convolutional neural networks for next location prediction. In *2019 IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium)*, pages 149–157.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, (Preprint):1–44.

Sandeep Soni, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative texts.

Sharifa Sultana, Renwen Zhang, Hajin Lim, and Maria Antoniak. 2022. Narrative datasets through the lenses of nlp and hci.

Chenyu Tian, Yuchun Zhang, Zefeng Weng, Xiusen Gu, and Wai Kin Chan. 2022. Learning fine-grained location embedding from human mobility with graph neural networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Taras K Vintsyuk. 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4(1):52–57.

Eitan Wagner, Renana Keydar, and Omri Abend. 2023. Event-location tracking in narratives: A case study on holocaust testimonies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8789–8805, Singapore. Association for Computational Linguistics.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6809–6821, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024. Beyond the limits: A survey of techniques to extend the context length in large language models.

Bingsheng Yao, Ethan Joseph, Julian Lioanag, and Mei Si. 2022. A corpus for commonsense inference in story cloze test. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3500–3508, Marseille, France. European Language Resources Association.

Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer. 2021. Detecting scenes in fiction: A new segmentation task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177, Online. Association for Computational Linguistics.

Zeyu Zhang and Steven Bethard. 2023. Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution.

8. Language Resource References

Author Index