LREC 2026

# The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026

## Workshop Proceedings

**Editors**
**Isuri Anuradha and Martin Wynne**

11 May 2026

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026



ISBN 978-2-493814-93-7

# Preface

Holocaust testimonies constitute one of the most significant and ethically complex bodies of primary source material available to historians, archivists, and language technologists alike. As digitisation efforts continue to expand the scale and accessibility of these collections, the intersection of natural language processing, speech technology, and digital humanities offers increasingly powerful tools for their analysis, preservation, and dissemination. This workshop, the second in the HTRes series following the inaugural edition at LREC-COLING 2024 in Turin, brings together researchers working at precisely this intersection.

The accepted papers reflect both the breadth and the maturation of the field. On the infrastructure and access front, Kučera et al. demonstrate that hybrid semantic-lexical retrieval consistently outperforms vector-only approaches across heterogeneous multimodal testimony collections, while Del Grosso et al. present an integrated digital environment for Italian survivor testimonies combining TEI encoding with embedding-based semantic retrieval. Dermentzi's EHRI Annotator advances this infrastructure work further, offering a web-based multilingual named entity recognition and entity linking tool that achieves an overall accuracy of 77.7On the language resource front, Brückner et al. introduce MalachNER, a ten-language NER dataset built from manually transcribed oral testimonies that captures the distinctive challenges of transcribed speech, and demonstrate that joint training on written and oral testimony achieves state-of-the-art results on both. A companion paper by the same group presents XLM-RoBERTa-malach, a transformer domain-adapted through continued pretraining on over 33,000 automatically transcribed and machine-translated Visual History Archive interviews, yielding meaningful NER improvements, particularly for domain-specific entities such as camps and ghettos across eight languages.

At the level of large-scale content analysis, Trainin et al. use discourse segmentation, topic modelling, and LLM-based classification to compare over 1,600 testimonies from the USC Shoah Foundation and Fortunoff archives, uncovering both structural differences and surprising convergences. Gagnier reveals systematic emotional and thematic arcs in 500 CORHOH testimonies through cross-modal trajectory analysis, and Jaff's corpus-scale diagnostic study of sentiment classifier disagreement on the same corpus introduces the ABC stability taxonomy, finding that inter-model divergence is driven primarily by boundary decisions around neutrality. On the audiovisual side, Mattingly and Bailey-Tomecek provide the first large-scale ASR evaluation on the Fortunoff Video Archive across eight languages, Mantaj et al. outline a multimodal workflow for detecting emotionally significant moments in testimony interviews, and Bleaman presents the first ASR system for Northeastern Yiddish, achieving 37.96

Taken together, these contributions reflect a field moving with confidence toward scalable, reproducible, and ethically grounded pipelines. The testimonies preserved in these archives are not data points but records of individual human experience, and any computational engagement with them must be pursued with corresponding care and humility. We thank CLARIN and EHRI for their ongoing support, the programme committee for their thorough reviews, and above all the survivors and witnesses whose testimonies make this work both possible and necessary.

**Organising Committee**

Isuri Anuradha, Lancaster University, UK
Martin Wynne, Oxford University, UK
Paul Rayson, Lancaster University, UK

**Programme Committee**

Angelo Mario Del Grosso, CNR-Istituto di Linguistica Computazionale "A. Zampolli", Italy
Elvira Mercatanti, CNR-Istituto di Linguistica Computazionale "A. Zampolli", Italy
Ignatius Ezeani, Lancaster University, UK
Ines Matres, University of Helsinki, Finland
Maria Dermentzi, King's College London, UK
Martin Bulin, University of West Bohemia, Czech Republic
Renana Keydar, The Hebrew University of Jerusalem, Israel
Robert Ehrenreich, United States Holocaust Memorial Museum, USA
William Mattingly, Yale University, USA

# Table of Contents

# Workshop Program

**Arrival and Registration**

**14:00–14:05** **Welcome and Introduction**

**14:10–15:30** **First session: Computational Methods for Historical Testimony: Search, Transcription, and Analysis**

14:10–14:30 *Integrating TEI Publication, Guided Exploration, and Vector Databases for Semantic Search in the Voci dall'Inferno Project*
Angelo Mario Del Grosso, Elvira Mercatanti, Carla Congiu and Marina Riccucci

14:30–14:50 *Towards Semantic Searching in Diverse Multimodal Collections*
Václav Kučera, Martin Bulín, Jan Švec and Pavel Ircing

14:50–15:10 *Automatic Transcription of Holocaust Testimonies in Yiddish: Orthographic Comparison and Cross-Domain Validation*
Isaac L. Bleaman

15:10–15:30 *From Consensus to Split Decisions: ABC-Stratified Sentiment in Holocaust Oral Histories*
Daban Q. Jaff

**15:30–16:30** **Second session: Poster session**

*EHRI Annotator: A Web-Based Tool for Named Entity Recognition and Linking in Holocaust-Related Texts*
Maria Dermentzi

*The Shape of Testimony: A Scalable Framework for Oral History Archive Comparison*
Renana Keydar, Amit Pinchevski and Itamar Trainin

*From Oral History to Structured Data: The MalachNER Dataset*
Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián and Pavel Pecina

*Emotions In Oral History Interviews: A Multimodal Approach to Holocaust Testimonies*
Nele Mantaj, Vaibhav Agarwal and Ines Matres

**16:00–16:30** **Coffee Break**

**16:30–17:30** **Third session:Modelling Holocaust Survivor Testimony: Language, Speech, and Emotion**

16:30–16:50 *Modeling the Language of Holocaust Survivors' Testimony with Domain-Adapted Transformers*
Christopher Brückner, Jan Lehečka, Jan Švec and Pavel Pecina

16:50–17:10 *Evaluating Automatic Speech Recognition for Holocaust Testimonies: A Large-Scale Analysis of Whisper Performance on the Fortunoff Video Archive*
William J.B. Mattingly and Christy Bailey-Tomecek

17:10–17:30 *Cross-Modal Modeling of Emotional and Thematic Trajectories in Holocaust Survivor Oral Histories*
Henry Gagnier

**17:30–18:00** **Pannel Discussion**

**Final Remark**

# Integrating TEI Publication, Guided Exploration, and Vector Databases for Semantic Search in the Voci dall'Inferno Project

**Angelo Mario Del Grosso, Elvira Mercatanti, Carla Congiu, Marina Riccucci**
Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Via Moruzzi, 1, Pisa, Italy
Università di Pisa, P.za Evangelista Torricelli, 2, Pisa, Italy
{angelomario.delgrosso, elviramercatanti}@cnr.it
c.congiu1@studenti.unipi.it, marina.riccucci@unipi.it

**Abstract**

This paper presents the *Voci dall'Inferno* digital environment, which integrates TEI-based encoding, web publication, and embedding-based semantic retrieval for testimonies by Italian Holocaust survivors. The corpus comprises written and oral sources encoded in XML-TEI through an ODD customization that documents provenance, structure, and interpretive layers. The web application, build on eXist-db, supports guided access, management, visualization, and exploratory analysis of encoded data. Within this infrastructure, we report a pilot semantic-retrieval study on references to Dante's *Divine Comedy*, using SentenceTransformers embeddings and a vector database to retrieve both literal and non-literal Dantean passages. Given the current corpus size, findings are interpreted as exploratory and method-oriented, and future large-scale validation will be conducted. We also address the ethical and legal constraints that shape access policies and long-term reuse in sensitive historical collections.

**Keywords:** Sentence similarity, web application, eXist-db, TEI-based digital archives, Divine Comedy

## 1. Introduction

Preserving, curating, and providing access to testimony archives, including both written and oral sources, is essential for historical, literary, and linguistic scholarship, as well as for ensuring that these stories remain available for research, education, and public memory.

Holocaust testimony archives remain an underused resource across a wide range of fields, including linguistics, oral history, sociology, and related disciplines. By preserving firsthand testimonies, these sources serve as repositories of collective memory and contribute to safeguarding intangible cultural heritage.

The lack of shared infrastructures and interoperability mechanisms, together with the legal and ethical complexity associated with access conditions and data governance, remains a major challenge (Calamai et al., 2021). These issues are particularly evident in the Italian context, where collections are frequently managed by individual researchers or small teams, and where institutions may lack trusted digital repositories and clear policies for archiving primary sources and scholarly analyses. As a result, historically and culturally significant testimonies risk being dispersed or even lost (Abete et al., 2025).

Testimony archives still struggle to be fully integrated into research-data management and open-science agendas, where long-term sustainability depends on shared protocols, interoperable infrastructures, and robust legal and ethical procedures (Calamai and Frontini, 2018). From this perspective, the (re)use of these sources is commonly associated with the adoption of the FAIR principles, which promote findability, accessibility, interoperability, and reusability as prerequisites for durable and responsible dissemination (Jong et al., 2018).

This paper builds on our contribution to the 2024 edition of the HTRes workshop (Anuradha et al., 2024), where we introduced the *Voci dall'Inferno* initiative and described the collection and curation of both oral and written resources, focusing mainly on non-literary testimony texts. We outlined the data-acquisition pipeline, annotation choices, quality-control procedures, and intended research and evaluation use cases (Del Grosso et al., 2024), with particular attention to investigating the presence and use of Dante's lexicon in the corpus. We also reported initial dataset statistics and observations, and discussed the limitations and future directions that motivate the present submission.

The scope of this paper is twofold but methodologically unified: (i) to document an interoperable TEI-based archival and publication framework for heterogeneous Holocaust testimonies, and (ii) to present a pilot retrieval component for Dantean intertextuality built on sentence embeddings and vector search. Our main contribution is therefore the end-to-end integration of data curation, encoding, publication, and computational exploration, with explicit attention to reproducibility, interpretability, and governance constraints (Mercatanti et al., 2025a).

In this context, the *Voci dall'Inferno* initiative[1] contributes to ongoing efforts to preserve and make testimonies accessible by providing a structured, interoperable digital archive spanning both written and oral sources, and by offering tools for explo-

[1]The GitHub repository for the project is available at https://github.com/CoPhi/voci-inferno/.



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 1–11
11 May 2026. 

ration and analysis that support long-term reuse under appropriate legal and ethical constraints (Mercatanti et al., 2025b).

The remainder of the paper is structured as follows. Section 2 reviews related work and situates our contribution within TEI-based initiatives and semantic retrieval approaches. Section 3 outlines the methodology and data-processing workflow adopted in the project. Section 4 describes the *Voci dall'Inferno* web application and its main functionalities for guided exploration and analysis. Section 5 presents our semantic search component based on embeddings and vector databases, focusing on the automatic retrieval of explicit and implicit Dantean echoes. Finally, we summarize current limitations and future directions in Section 6.

## 2. Related Work

Like other digital archives, testimony collections require harmonized metadata practices, robust workflows from digitization to preservation, and technical solutions that support controlled access while maintaining visibility and reusability (Lin et al., 2020).

The design of our application is informed by the broader ecosystem of tools for the production, dissemination, and exploration of TEI-encoded content (Bénière et al., 2024). In particular, frameworks such as TEI Publisher[2] provide configurable web interfaces for browsing, searching, and rendering TEI documents, often leveraging XQuery/XSLT pipelines and XML databases. Client-side approaches such as CETEIcean[3] enable in-browser transformation of TEI to HTML, supporting lightweight publication workflows and interactive presentation. Related projects and toolkits (e.g., TEI Boilerplate[4]) similarly aim to facilitate the rendering of TEI documents on the web with minimal infrastructure.

For digital editions with advanced navigation and reading interfaces, EVT (Edition Visualization Technology)[5] is a widely used environment for publishing scholarly editions and integrating multiple views (e.g., diplomatic and interpretative transcriptions, facsimiles, and apparatus).

In this comparison, TEI Publisher is used as a conceptual benchmark for publication patterns and user-facing functionalities rather than as our deployment stack. Our implementation is based on eXist-db and XQuery while remaining compatible with architectural principles shared by TEI-native publication environments.

Recent advances in semantic retrieval are largely driven by Transformer encoders derived from BERT (Devlin et al., 2018), which enable queries and text units to be mapped to dense vector representations and compared in an embedding space (Venkatesh Sharma et al., 2024). In this setting, retrieval can be implemented with bi-encoder architectures and optionally complemented by re-ranking components for improved precision. Sentence-level models (e.g., Sentence-BERT and related *SentenceTransformers* variants) (Mayil and Jeyalakshmi, 2023) are now commonly adopted to support semantic similarity and passage retrieval, especially when the goal is to capture paraphrases and non-literal correspondences rather than exact lexical matching (Zhou et al., 2023).

Operationally, embedding-based retrieval is typically supported by vector indexes and specialized databases providing approximate nearest-neighbor search, metadata filtering, and hybrid lexical–semantic retrieval. Among widely used infrastructures, Weaviate[6] and LlamaIndex[7] offer integrated environments for storing vectors alongside structured metadata and executing similarity queries at scale, and they can be combined with application-layer logic to expose results. In the *Voci dall'Inferno* project, these technologies served as methodological references for developing the testimony archive, combining TEI-based publication workflows with semantic retrieval.

## 3. Methodology

Our methodology integrates (i) philological and linguistic work to curate heterogeneous testimonies, (ii) TEI-based modeling and encoding to represent textual phenomena and preserve interpretability, (iii) digital modules for guided web access and visual analytics, and (iv) computational modules for semantic retrieval of Dantean echoes using BERT-based embeddings.

**Data acquisition and transcription.** The corpus is built from heterogeneous sources, including written documents (e.g., diaries, memoirs, manuscripts, and printed texts) and oral testimonies (audio or audiovisual interviews); in the current release, testimonies used in retrieval experiments are in Italian. Source selection is based on resources currently available to the team (an unpublished oral-interview collection for spoken testimonies and unpublished private and family collections for written materials) and on explicit inclusion criteria: documented provenance, sufficient metadata for contextualization, legal feasibility for research use and controlled publication, and transcription/encoding feasibility within the project workflow. Exclusion

[2] https://teipublisher.com/index.html
[3] https://github.com/TEIC/CETEIcean
[4] http://teiboilerplate.org/
[5] https://github.com/evt-project/evt-viewer/
[6] https://weaviate.io/
[7] https://www.llamaindex.ai/


Figure 1: Example of an XQuery function used in the web application

criteria include uncertain provenance, unresolved rights constraints, and insufficient textual or audio quality for reliable encoding. To reduce circularity in downstream analysis, selection is not based on the prior presence of Dantean references. Oral sources follow a transcription workflow that combines manual revision with semi-automatic support, using speech-to-text tools such as the CLARIN transcription portal (Draxler et al., 2024) to produce transcriptions that are faithful to the spoken interaction and suitable for linguistic analysis.

**Domain-specific transcription language.** To ensure consistency and machine processability, we formalize transcription conventions as a domain-specific language (DSL) (Bambaci and Boschetti, 2020). Oral testimonies contain a complex set of verbal and non-verbal phenomena (e.g., interruptions, unfinished segments, repetitions, false starts, pauses, background noise, prosodic cues; and, for audiovisual sources, gestures and body movements) that must be explicitly identified and encoded. We therefore define a controlled set of transcription markers and a parsing strategy grounded in a context-free grammar approach, in line with established transcription ecosystems such as CHAT (MacWhinney, 2019) and DT2 (Bois, 1991), enabling formal recognition of units such as speaker changes, gaps, unclear spans, and prosodic variations. This design facilitates readability for encoders and downstream extraction of structured information.

**TEI-based encoding and customization.** All testimonies are encoded in XML-TEI through a unified data model that supports both oral and written sources while preserving source-specific characteristics. The model supports multiple annotation layers, including document structure, source alignment, entities, events, relations, and interpretive tags. Encoding decisions aim to (i) retain stable links to primary sources (including alignment to audio segments or line-based references where applicable), (ii) capture document structure (divisions, segments, and anchors), and (iii) represent the authorial and analytical layers needed by the project, such as additions, deletions, named entities, events, places, and relations. In the current implementation, core documentary and linguistic layers are represented inline in TEI for transparency and editorial traceability, while computationally derived layers can be externalized as stand-off annotations when needed. In this setup, stand-off annotation is used to represent named entities, relations among mentioned individuals, cited events, and quotations through six dedicated TEI lists: `listPerson`, `listPlace`, `listOrg`, `listRelation`, `listEvent`, and `listBibl`. An ODD customization governs element and attribute usage, together with project-specific constraints, ensuring consistency, validation, and long-term maintainability (Mercatanti et al., 2025c).

**Information extraction and linguistic analysis.** On top of the TEI backbone, we extract structured

information for guided access and analysis, including witness-level metadata, testimony provenance, and corpus-derived annotations for maps, timelines, and exploratory statistics. In addition, we curate and visualize lexical information derived from linguistic analysis of the testimonies, with particular attention to Dantean lexicon, quotations, and allusions. This extraction layer is explicitly separated from semantic-retrieval testing: editorial annotations (explicit quotations, implicit quotations, allusions) define the ground truth, whereas retrieval experiments are executed as a distinct computational phase. This separation is intended to limit circularity and to make evaluation assumptions explicit.

**Guided access and visual analytics in the web application.** Publication and exploration are implemented through a web application built on eXist-db. XQuery functions retrieve TEI-encoded content and assemble HTML views that provide guided access to witnesses, testimonies, and analytical outputs. The interface integrates interactive components for reading and listening to encoded testimonies, and for exploring encoded phenomena, named entities, relationships, maps, timelines, and corpus-level statistics. The architecture is modular and extendable. Scalability across different collections depends on harmonized metadata quality, governance policies, and indexing strategies.

**Embedding-based semantic retrieval of Dantean echoes.** To complement editorially curated links and lexicon-driven analysis, we integrate an embedding-based semantic retrieval module that surfaces non-literal correspondences between testimony fragments and passages of the *Commedia*. Textual units (e.g., verses, terzine, or longer segments) are normalized and represented with sentence-level Transformer embeddings, then indexed in a vector database to enable nearest-neighbor queries. Retrieved candidates are returned with bibliographic coordinates and metadata (cantica, canto, verse locus) to support verification and scholarly use. This module supports both explicit quotations (typically characterized by high lexical overlap) and implicit quotations or allusions, where semantic proximity is more informative than string matching.

**Ethical and legal considerations (data governance).** Given the sensitivity of Holocaust-related testimonies and the heterogeneity of archival provenance, methodological choices are complemented by governance measures that regulate access, documentation, and reuse. We adopt a privacy- and rights-aware approach to dissemination, ensuring that publication and computational processing remain aligned with applicable legal and ethical constraints while preserving the research value of the archive.

## 4. The *Voci dall'Inferno* web application

The digital corpus is accessible through the *Voci dall'Inferno* web application, developed as an integrated environment for managing, presenting, exploring, and analyzing encoded testimonies (Mercatanti et al., 2025a). Built on eXist-db, the application combines the HTML templating framework with XQuery functions to process XML-TEI documents and generate HTML fragments assembled into end-user pages. This architecture separates presentation from data-processing logic, improves maintainability, supports incremental scaling, and contributes to long-term sustainability. The platform is therefore extendable across collections that adopt compatible TEI and metadata profiles.

For example, Figure 1 shows an XQuery function (`app:contaTestimonianzeArchivio()`) that returns the number of testimonies and their categories.

The web application provides several features for consulting and exploring heterogeneous data derived from encoding. The current *Voci dall'Inferno* corpus includes 25 testimonies from 20 witnesses and reflects substantial variation in both source type and testimonial profile, including accounts by people who experienced the Lager firsthand and by others who were never deported. Importantly, corpus inclusion is not conditioned by prior Dantean evidence; this design supports a less circular setting for retrieval analysis. The testimonies currently processed, encoded, and semantically analyzed are Italian-language materials, consistently normalized through the same preprocessing pipeline.

To support guided access and conceptual clarity, the project adopts a hierarchical taxonomy that groups witnesses according to their historical experience (Fig. 2). At the top level, the collection is organized under *Witnesses* and divided into *Deported witnesses* and *Non-deported witnesses*. The *Deported* branch is further articulated into *Jewish deportees* and *Non-Jewish deportees*, the latter including *Italian Military Internees* (IMI) and *Italian Civil Internees* (ICI). The *Non-deported witnesses* branch includes Jewish partisans, currently represented in the archive by the testimony of Emanuele Artom.

Although still subject to refinement, this taxonomy provides a coherent organizational framework for managing the corpus's heterogeneity while clearly identifying the provenance and historical context of each testimony.

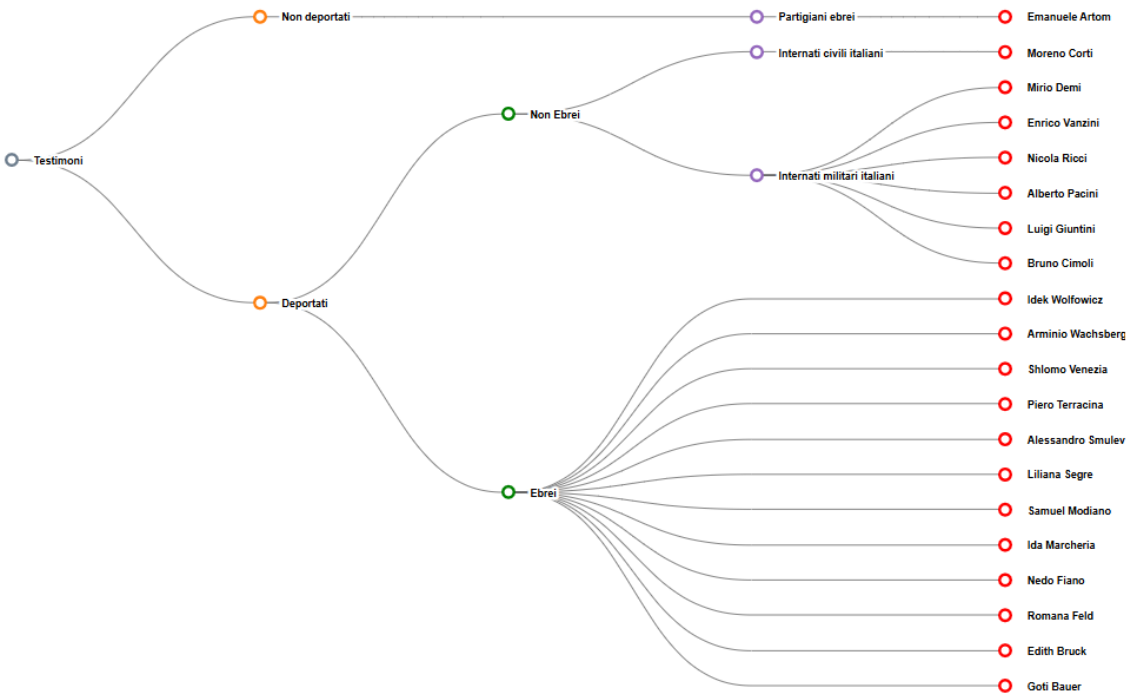


Figure 2: The *Voci dall'Inferno* project taxonomy of witnesses

After encoding, the application enables the extraction and visualization of heterogeneous data types, including written and spoken phenomena, named entities, maps, interpersonal relationships, timelines, and statistical analyses. Interactive visualizations are implemented with Highcharts, a JavaScript library that enhances accessibility and supports intuitive engagement with quantitative analyses.

The application is structured into nine main sections: Home, The Project, Voices, Search for a Witness, Dante, Statistics, Automatic Transcription, Events, and Bibliography (Fig. 3).


Figure 3: Homepage of the application

The core section, *Voices*, provides access to witnesses and their testimonies through an alphabetical navigation menu that allows users to filter results by the initial letter of a witness's surname and open individual profile pages.

For each witness, users are directed to a dedicated page displaying a brief biographical profile together with the list of encoded testimonies currently available for consultation. The page also provides visualizations to support analytical exploration of the encoded data, including (i) a directed, labeled graph of interpersonal relationships, (ii) two maps showing places mentioned and the witness's movements before, during, and after deportation (Fig. 4), and (iii) a timeline of the main events cited by the witness (Fig. 5).

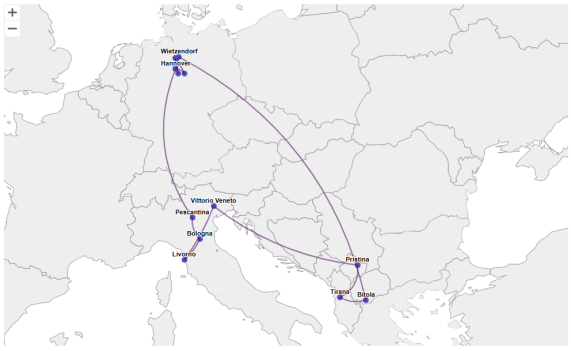


Figure 4: Map of the witness's movements

Upon selecting a testimony, users access a dedicated page presenting structured metadata extracted dynamically from the XML-TEI source through XQuery functions. The displayed information varies by testimony type (oral vs. written), enabling a differentiated representation of source-specific features. For oral testimonies, the interface

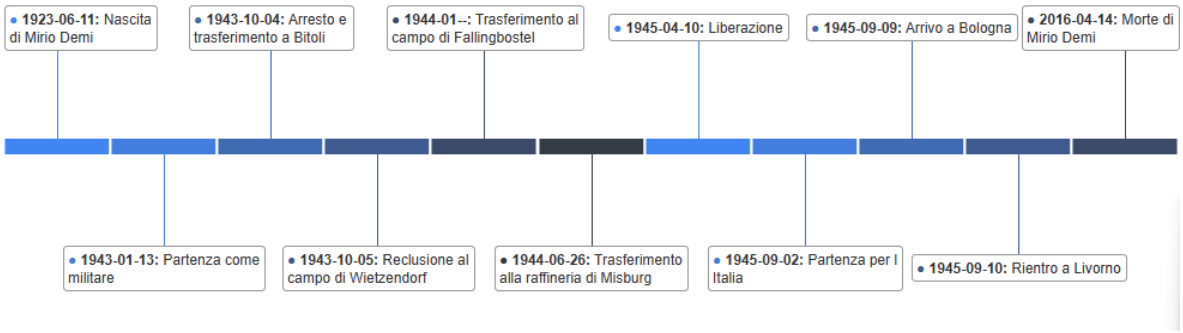


Figure 5: Timeline of the main events cited in the testimony

synchronizes the structured abstract (*regesto*), organized into segments, with the timeline, allowing users to read each segment and listen to the corresponding portion of the audio recording.

The transcription interface adapts to the specific type of resource. For written testimonies, an image-based mode displays the source image alongside its transcription (Fig. 6); alternatively, users can view the transcription alone, with a legend of encoded phenomena that can be interactively highlighted in the text. The same highlighting functionality is available for oral sources, with the legend dynamically tailored to the resource type (Fig. 7).

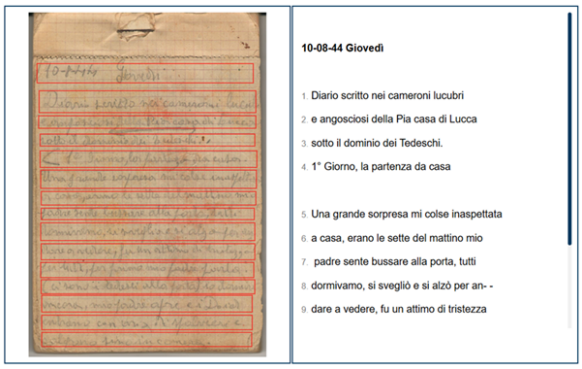


Figure 6: Alignment of the primary source image with its corresponding transcription

Each testimony is further enriched by a set of statistical visualizations, including charts on encoded phenomena, witness expression patterns, and named-entity distribution. Additionally, the interface provides an analysis of references to the *Divine Comedy*. Explicit and implicit quotations, allusions, and Dantean lexical references can be explored interactively: selecting a citation highlights the corresponding verses or passages of the poem referenced in the testimony.

The web application features two analytical sections, *Dante* and *Statistics*, dedicated to exploring the encoded corpus. The *Dante* section examines the presence and distribution of references to the *Divine Comedy* within the testimonies. It provides visualizations and quantitative summaries of explicit and implicit quotations, allusions, and lexical references identified during transcription and encoding. The analysis is twofold: it investigates both the typology of Dantean references and the witnesses who most frequently employ Dante's language to articulate the Lager experience. Out of 25 testimonies (produced by 20 witnesses), 10 contain references to the poem, totaling 61 occurrences: 15 explicit quotations, 4 implicit quotations, 7 allusions, and 35 terms from the *Comedy* (Fig. 8). Among the 19 encoded quotations, the majority refer to the *Inferno* (16 cases), with only limited references to the *Purgatorio* (2) and the *Paradiso* (1).

The *Statistics* section provides broader corpus-level analyses. It presents visualizations and summary data concerning the composition of the archive, including the distribution of testimonies by witness category and the provenance of the two main categories of sources, oral and written. The section also documents the overall extent of the encoded material: to date, 18 hours, 35 minutes, and 48 seconds of oral recordings and 395 pages of written sources have been transcribed and encoded.

## 5. Semantic search with embeddings and vector databases

To complement guided access to the corpus, we are developing *Dante Similarity Search*, a semantic retrieval module that enables users to identify potential Dantean echoes across testimonies, including non-literal correspondences, through embedding-based similarity and vector search.

*Dante Similarity Search* is currently implemented as a prototype web application designed to detect echoes of Dantean language in concentration-camp survivor testimonies by linking prose frag-

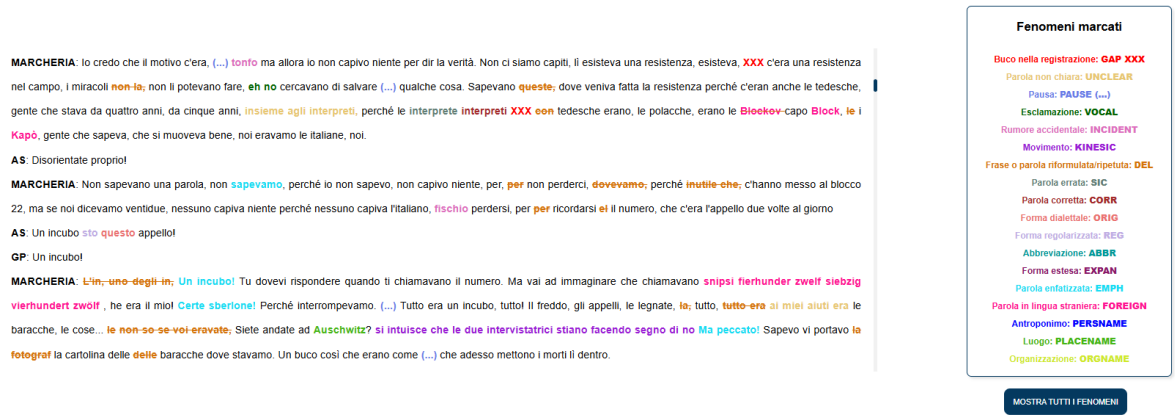


Figure 7: Transcription of an oral testimony with in-text highlighting of encoded phenomena

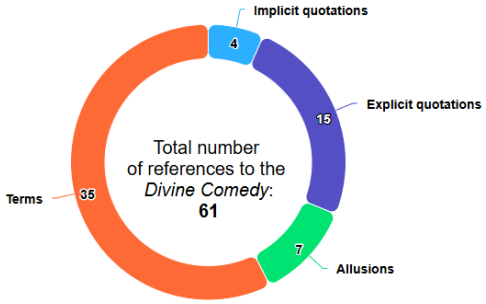


Figure 8: References to the *Divine Comedy*

ments to passages from the *Commedia*. Given a query, the system returns the closest candidates, ranked by similarity and enriched with metadata (*cantica*, *canto*, position), making results verifiable, citable, and suitable for subsequent human validation (Congiu et al., 2025) (Fig. 9). The approach goes beyond lexical overlap and targets semantic proximity that can surface non-literal correspondences.

The development of Dante Similarity Search draws inspiration from projects by William Mattingly[8] that aim to automatically identify biblical quotations in Latin. In Vulgata spaCy (2022)[9], the Clementine Vulgate is cleaned and organized into CSV format, and a spaCy pipeline is built by combining embeddings trained on the Patrologia Latina[10] (Bloom/floret) with two components: an EntityRuler to detect direct or partial quotations and a machine-learning model to detect quotations in context. A subsequent step links each occurrence to specific verses while addressing spelling and punctuation variability and incomplete quotations. Constraints such as requiring phrases of at least four words reduce false positives. With Weaviate Vulgate (2024), the approach is recast in terms of semantic similarity: verses are converted into vectors using SentenceTransformers (LaBSE[11]), indexed in Weaviate, and made accessible through a Streamlit web application that allows users to submit a text fragment and retrieve the most closely related verses.

The architecture integrates a vector-

[8] https://www.wjbmattingly.com/
[9] https://github.com/wjbmattingly/vulgata-spacy
[10] https://patristica.net/latina/
[11] https://huggingface.co/sentence-transformers/LaBSE

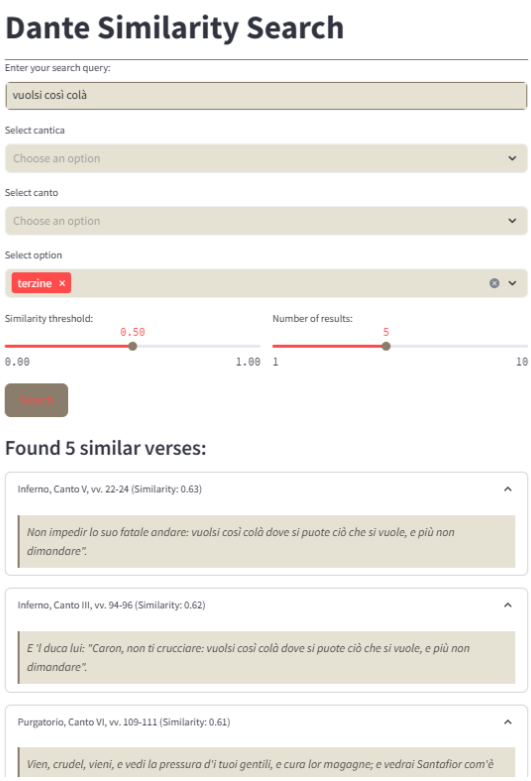


Figure 9: Dante Similarity Search

representation pipeline and a semantic retrieval module. Texts are normalized and transformed into contextual embeddings via *SentenceTransformers*[12]. We compared three candidate models in our pilot setting: `LaBSE`, `paraphrase-mpnet-base-v2`, and `all-mpnet-base-v2`. In our data, `LaBSE` produced weaker rankings for Dantean fragments, while `paraphrase-mpnet-base-v2` recovered some relevant passages but with lower ranking consistency. `all-mpnet-base-v2` provided the best overall trade-off in top-ranked relevance and ranking stability, and was therefore selected as the operational model. We note, however, that this remains a pragmatic choice for a pilot setup, and domain adaptation to historical/literary Italian remains a key next step. Vectors are then indexed in *Weaviate*[13], which supports efficient nearest-neighbor queries and structured metadata management needed to reconstruct the Dantean reference associated with each result. Interaction takes place through a *Streamlit*[14]-based interface, designed for entering queries and inspecting matches, including similarity scores and textual references. For testing purposes and to support external access and service sharing, the application can be exposed via *Cloudflare Tunnel*[15], simplifying deployment without requiring complex network configurations.

A central methodological component is the construction of datasets from the *Commedia*, transformed into collections of homogeneous, queryable textual units, each associated with bibliographic metadata. The choice of granularity directly affects interpretability and result quality: smaller units support precise anchoring but may be semantically fragile, whereas larger units introduce context and stability at the expense of precision. From this perspective, segmentation into *terzine* prioritizes semantic coherence, reduces ambiguity, and is effective when an echo is distributed across multiple lines; segmentation into single verses maximizes precision and is particularly suited to identifying explicit quotations, although it requires careful normalization in Python and filtering by metadata to limit spurious matches; segmentation into sentences offers a useful compromise, especially for paraphrases and reformulated echoes, preserving semantic relations that a single verse may not make explicit.

Overall, the workflow maps user input to the retrieval of the most similar Dantean candidates within a reproducible pipeline in which technical choices (model, textual units, search parameters)

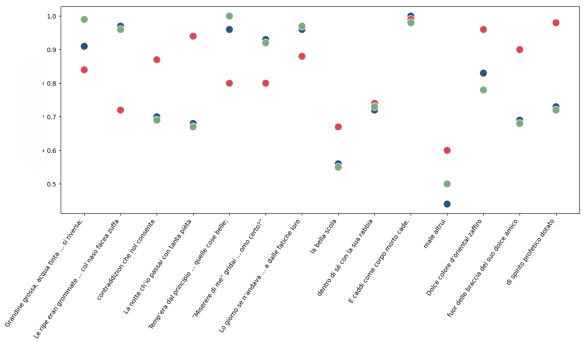


Figure 10: Similarity scores for a set of explicit Dante quotations identified in the testimonies. Each quotation is listed along the x-axis, while the y-axis represents a similarity score.

remain traceable and results interpretable. Given the current corpus size, we frame this component as a pilot study and avoid strong generalization claims. To support systematic validation, we will define an explicit protocol including *precision*, *recall*, and baseline comparisons. We will also set up experiments to distinguish explicit quotations, implicit quotations, and allusions. Editorial annotation and computational testing will be treated as separate phases to reduce circularity.

Indexing in Weaviate and inspection through Streamlit turn similarity into an operational tool: results become explorable and filterable, and can be critically evaluated thanks to metadata and textual coordinates. Quantitatively, explicit quotations are the most robust case: because testimony fragments often preserve literal portions of the *Commedia*, the system returns precise and reliable alignments (Fig. 10). Implicit quotations are more challenging: paraphrastic reformulations can preserve thematic resonance while substantially altering lexical form, making exact retrieval less stable even when top candidates remain semantically coherent with the input fragment.

A representative failure case concerns Nicola Ricci's diary quotation, "Si va verso la fame, si va verso il freddo, si va verso l'inferno," for which the system did not recover the expected match with *Inferno* III, 1–3 ("Per me si va ne la città dolente / per me si va ne l'etterno dolore / per me si va tra la perduta gente"). For a human reader, the allusion is immediate; for the model, it is an implicit intertextual signal that exceeds purely distributional similarity. This example clarifies a key limitation of the current setup and motivates the next step: fine-tuning on a *Commedia*-centered corpus enriched with commentaries, paraphrases, and curated intertextual links, in order to improve recognition of non-explicit echoes and reduce confusion between shallow lexical similarity and deeper semantic-literary correspondence.

[12] https://sbert.net/

[13] https://weaviate.io/

[14] https://streamlit.io/

[15] https://developers.cloudflare.com/cloudflare-one/networks/connectors/cloudflare-tunnel/

The output is not limited to on-screen consultation: retrieval results can be reintegrated into the encoding workflow in a structured form. Starting from a testimony fragment, the system stores the top candidates across different textual granularities (terzine, verses, and sentences), each with a similarity score and bibliographic coordinates (cantica, canto, verse range). This enrichment process is implemented as a Python routine over testimony XML files and is designed to remain auditable, reversible, and compatible with subsequent human validation. The enrichment procedure follows an explicit criterion: for each encoded quotation, the system searches across the three textual unit types (terzine, verses, and sentences) and, for each type, inserts the three closest results. The encoded documents are then enriched with the suggested references (Fig. 11).


Figure 11: For each encoded quotation, the system automatically searches terzine, verses, and sentences, adding the three closest matches with bibliographic references to the XML files in a structured format.

## 6. Conclusions and Further Directions

This paper presented an integrated framework for testimony curation, TEI encoding, web publication, and embedding-based semantic retrieval, together with a pilot study on Dantean echoes. The current results are promising but preliminary and should be interpreted as proof-of-concept evidence rather than as definitive performance claims. We delivered a first operational release of the *Voci dall'Inferno* web application, including dedicated views for exploring Dantean lexicon, quotations, and allusions, and we defined a reproducible workflow linking editorial annotation and computational analysis. Given the limited experimental scale, we prioritize methodological transparency and a clear separation between annotation and retrieval evaluation. Future work will focus on: (i) refining the transcription workflow by integrating more robust support for automatic recognition and DSL-based encoding across different types of primary sources; (ii) improving interpretability and methodological soundness by connecting the archive to major SSH research infrastructures, in particular the CLARIN-IT community; (iii) extending the system with additional visualization components (e.g., temporal and thematic facets) to better support scholarly workflows; (iv) expanding the benchmark and reporting full IR metrics; (v) measuring annotation reliability for explicit, implicit, and allusive categories; and (vi) evaluating multilingual and domain-adapted embedding models for historical and literary Italian.

## 7. Acknowledgments

The project has progressed primarily through the work of undergraduate and graduate students, mostly enrolled in the Digital Humanities and Italian Studies degree programs at the University of Pisa. To date, twenty-two students have completed their degree theses within the framework of *Voci dall'Inferno*. The initiative has also benefited from the support of several institutions and research infrastructures: the Department of Philology, Literature and Linguistics at the University of Pisa; the Institute for Computational Linguistics "A. Zampolli" of the National Research Council (CNR) in Pisa[16]; the CLARIN-IT research infrastructure[17]; the Collaborative and Cooperative Philology Lab (CoPhiLab)[18]; the CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)[19]; the Centro Interdipartimentale di Studi Ebraici (CISE)[20]; the Centro di Documentazione Ebraica Contemporanea (CDEC)[21].

### 7.1. Ethical considerations and limitations

The current version of the *Voci dall'Inferno* web application is accessible only to project members in a restricted environment. This choice reflects both the sensitive nature of the collection and the need to ensure responsible management of sources and content in compliance with data-protection requirements. The ethical and legal framework is particularly complex due to the heterogeneity of the corpus: while some testimonies derive from publicly accessible archives, others consist of unpublished materials from private family collections. Obtaining consent for research purposes is often challenging because many data subjects are difficult to contact due to age or because they are deceased. Furthermore, the testimonies span different historical periods, with some sources predating the entry into

[16] https://www.ilc.cnr.it/
[17] https://www.clarin-it.it/it
[18] https://cophilab.ilc.cnr.it/
[19] https://diptext-kc.clarin-it.it/
[20] https://cise.unipi.it/
[21] https://www.cdec.it/

force of the GDPR, which complicates the identification of lawful bases for processing.

An additional layer of complexity concerns the dual nature of the sources (oral and written), with oral testimonies presenting specific challenges for data governance. Audio recordings constitute a particularly sensitive category of research material, as they simultaneously function as historical documents, scientific sources, and biometric identifiers. This hybridity creates tensions between openness and protection that directly affect decisions about accessibility, reuse, and long-term sustainability within research infrastructures.

Recent collaboration with the ROADS project, which has defined a FAIR-by-design model for managing oral archives throughout the entire data life cycle (from data collection to publication and reuse), has enabled us to test and refine this approach on the *Voci dall'Inferno* archive. The ROADS framework provides a transferable model to guide the transition from project-based archives to FAIR, sustainable, and reusable research resources, ensuring compliance with data-protection requirements while respecting the sensitivity of the documented contexts. A key factor in the sustainability of the ROADS model is the involvement of legal experts embedded within participating institutions, who mediate between Open Science principles and data-protection constraints. Their contribution supports the legal robustness and transparency of research resources, highlighting how the long-term reuse of sensitive historical data depends not only on technical solutions but also on stable governance frameworks and legal accountability (Abete et al., 2026).

In practice, this work could inform a set of concrete strategies to be progressively implemented within the *Voci dall'Inferno* archive. These may include: (i) formalizing the distribution of responsibilities among stakeholders (e.g., via joint controllership agreements and the identification of a single contact point for data-subject requests), (ii) documenting provenance and consent status at the item level, while adopting a diligent-search approach for legacy materials collected before current standards, aimed at contacting data subjects or, where this is no longer possible, their potential heirs, (iii) introducing layered information and multi-level consent procedures in the case of newly collected data, allowing participants to make informed choices about different levels of access, dissemination, and reuse of their testimonies, and (iv) applying data-minimization measures (e.g., redacted public views and restricted access to particularly sensitive content), supported by logging mechanisms.

From this perspective, the model defined by the ROADS project represents a fundamental guideline for orienting future decisions concerning data management, consent documentation, and access policies. Its adoption could support the gradual transition from a project-based archive to a FAIR, sustainable, and reusable research resource, while ensuring compliance with data-protection requirements and respect for the sensitivity of the historical contexts represented in the corpus.

## 8. Bibliographical References


Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, and Monica Monachini. 2026. La filiera legale di ROADS. Una proposta FAIR per archivi orali analogici. Technical report, Zenodo.

Giovanni Abete, Cesarina Vecchia, Silvia Calamai, Alessandro Casellato, Sergio Canazza, Elvira Mercatanti, Monica Monachini, Roberta Ottaviani, Giulia Zitelli Conti, and Giada Zuccolo. 2025. On the lifecycle of Italian oral archives: the ROADS project. In *La voce della grammatica. Nuove prospettive sull'interazione tra fonetica e morfologia, sintassi, lessico*. Associazione Italiana di Scienze della voce.

Isuri Anuradha, Martin Wynne, Francesca Frontini, and Alistair Plum, editors. 2024. *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Luigi Bambaci and Federico Boschetti. 2020. Encoding the Critical Apparatus by Domain Specific Languages: The Case of the Hebrew Book of Qohelet. In *La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica.*, Quaderni di Umanistica Digitale, Milano. Università Cattolica del Sacro Cuore.

Sarah Bénière, Floriane Chiffoleau, and Laurent Romary. 2024. TEI specifications for a sustainable management of digitized holocaust testimonies. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 10–17, Torino, Italia. ELRA and ICCL.

John W. Du Bois. 1991. Transcription design principles for spoken discourse research. *Pragmatics*, 1:71–106.

Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, (47):339–354.

Silvia Calamai, Stefania Scagliola, Fabio Ardolino, Christoph Draxler, Arjan van Hessen, and Henk van den Heuvel. 2021. Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant. In *Selected Papers from the CLARIN Annual Conference 2021, virtual event, September 27-29, 2021*, volume 189 of *Linköping Electronic Conference Proceedings*, pages 1–9. Linköping University Electronic Press.

Carla Congiu, Angelo Mario Del Grosso, and Marina Riccucci. 2025. Verso l'implementazione di un sistema di riconoscimento diallusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in project. In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025*, Verona. AIUCD. Num Pages: 663.

Angelo Mario Del Grosso, Marina Riccucci, and Elvira Mercatanti. 2024. The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of project Project. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Pavel Ircing, and Jan Lehečka. 2024. Speech technology services for oral history research. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 38–43, Torino, Italia. ELRA and ICCL.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann E. Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sierman, Dina V. Sokolova, Martina Stockhause, and John Westbrook. 2020. The TRUST Principles for digital repositories. *Scientific Data*, 7(1):144.

Brian MacWhinney. 2019. Chat manual.

V.Valli Mayil and T.Ratha Jeyalakshmi. 2023. Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP. In *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5.

Elvira Mercatanti, Carla Congiu, Angelo Mario Del Grosso, and Marina Riccucci. 2025a. Voci dall'Inferno: a Web application to study and analyze the Lager testimonies. In *DH2025 - Building Access and Accessibility, Open Science to all Citizens*, Lisbon, Portugal. Zenodo.

Elvira Mercatanti, Angelo Mario Del Grosso, and Marina Riccucci. 2025b. Voci dall'Inferno: Dante per esprimere l'indicibile: Un'applicazione digitale per esplorare le testimonianze non letterarie dei sopravvissuti ai Lager. *Umanistica Digitale*, (20):527–562.

Elvira Mercatanti, Marina Riccucci, and Angelo Mario Del Grosso. 2025c. Voci dall'Inferno: a TEI-Based Digital Archive for finding Dante in Concentration Camp Testimonies. In *"New Territories". Text Encoding Initiative Conference and Members' Meeting 2025*, Kraków, Poland. Zenodo.

K. Venkatesh Sharma, Pramod Reddy Ayiluri, Rakesh Betala, P. Jagdish Kumar, and K. Shirisha Reddy. 2024. Enhancing query relevance: leveraging SBERT and cosine similarity for optimal information retrieval. *International Journal of Speech Technology*.

Ya Zhou, Ning Zhao, Guimin Huang, Nanxiao Deng, and Qingkai Guo. 2023. Sentences Similarity Model Based on Fusion of Semantic, Syntactic and Word Order Multi-Features. In *2023 4th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pages 121–124.

# Towards Semantic Searching in Diverse Multimodal Collections

**Václav Kučera, Martin Bulín, Jan Švec, Pavel Ircing**
Department of Cybernetics @ Faculty of Applied Sciences @ University of West Bohemia in Pilsen
Univerzitní 8, 301 00 Pilsen, Czech Republic
vaclavk@students.zcu.cz, {bulinm, honzas, ircing}@kky.zcu.cz

**Abstract**

Digital humanities projects increasingly rely on heterogeneous collections of multimodal data, including video testimonies, scanned documents, and photographs. Despite the growing availability of such archives, researchers face challenges in efficiently locating relevant content due to the diversity of formats and the lack of unified retrieval methods. In this work, we present a general framework for semantic search over collections of multiple modalities. The framework integrates specific parsers and transforms all inputs into textual representations leveraging services like automatic speech recognition (ASR), optical character recognition (OCR), and generative-AI-based image captioning. Text is subsequently segmented into overlapping chunks, indexed in a vector database, and enriched through an automatic question generation (AQ) pipeline to create ground-truth queries for evaluation. We evaluate the framework on a constructed dataset derived from Holocaust-related archives, comparing two retrieval strategies (pure vector search vs. hybrid semantic-lexical search) under two chunking scenarios. Results demonstrate that hybrid search consistently outperforms vector-only retrieval, achieving high recall across modalities, and that semantic search is feasible even with diverse and noisy input sources. This framework provides a robust foundation for exploring complex multimodal archives, facilitating access to content that would otherwise remain difficult to discover.

**Keywords:** multimodal data, semantic search, digital humanities, vector retrieval

## 1. Introduction

Large multimodal archives encompassing diverse data types began to be amassed on a significant scale during the 1990s, when recording technologies and storage capacities became sufficiently affordable for institutions to preserve substantial volumes of material.

However, the challenge of efficiently retrieving relevant information from these extensive corpora emerged almost immediately thereafter. One prominent example is the USC Shoah Foundation's Visual History Archive (USC Shoah Foundation), which preserves authentic testimonies from Holocaust survivors and witnesses. This collection comprises approximately 52,000 interviews conducted between 1994 and 1999, totaling over 115,000 hours of video material recorded in 32 languages. The sheer volume of this material renders the identification of relevant content exceedingly challenging.

This challenge prompted the MALACH project (2001–2007), which sought to enhance access to these oral histories by advancing automatic speech recognition (ASR) and information retrieval (IR) technologies.

The initial approach in the MALACH project treated ASR and IR as independent tasks: audio was transcribed into text using state-of-the-art ASR systems, segmented into documents, and subjected to standard document-oriented IR. This strategy rapidly revealed significant limitations. Besides the poor performance of the ASR systems (roughly 40% Word-Error-Rate – WER – across languages), the IR systems had is own issues stemming from oversimplified designs, notably fixed-length sliding windows for segmenting continuous transcripts into pseudo-documents, bypassing the more complex task of topical coherence detection. Evaluated in CLEF campaigns in 2005–2007 using detailed topics specifying user information needs, these systems achieved dismal mean Generalized Average Precision (mGAP) scores (Pecina et al., 2007), attributable to both ASR errors and inadequate segmentation. Standard bag-of-words methods failed to leverage distinctions between relevant and non-relevant material mentioned in the search topic specifications. Ultimately, this era yielded disjointed ASR-IR pipelines with poor results and no user-friendly graphical interface for non-experts.

In the second "epoch" of research, we redefined the paradigm for ASR and IR system design to better align with the practical demands of searching continuous speech transcripts. Recognizing the absence of discrete documents in automatically transcribed streams, we shifted focus to identifying precise replay points—specific timestamps marking the onset of topic-relevant discussion—enabling direct playback of corresponding video segments. Departing from prior document-oriented IR systems, which relied solely on lexical overlap without semantic processing, we adopted a spoken term detection (STD) paradigm. In STD, queries comprise single words or short phrases submitted against a fixed collection, inverting the traditional keyword search model.

This transition also fostered tighter integration be-



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 12–19
11 May 2026. 

tween ASR and IR components. STD indexes incorporated not only the highest-probability ASR transcriptions but also competing hypotheses weighted by their estimated probabilities, enhancing detection robustness. Numerous ML methods were employed for STD over the years, details about the latest incarnation using the Transformer architecture can be found in (Švec et al., 2023). In this era we have also developed several iteratively improved versions of the graphical user interface (GUI), empowering non-expert users to submit queries and instantly replay pertinent testimony segments.

The latest set of techniques — named the "Asking Questions" (AQ) framework shifts to proactive, generative content enrichment. STD enabled efficient pinpointing of exact-term matches via integrated ASR-IR indexing but remained limited to user-initiated lexical queries, yielding timestamps for replay without semantic expansion or contextual guidance. In contrast, AQ generates contextually grounded, time-aligned question-answer pairs directly from transcripts, filtered for semantic coherence, to create navigable "open-set topics" that supplement lengthy monologues. This transforms passive listening into interactive exploration, anticipating user needs rather than reacting to explicit terms, while preserving testimony authenticity; it yields sparse, high-quality questions (one every 2 minutes post-filtering), outperforming STD in facilitating thematic discovery across unstructured speech (Bulin et al., 2025).

Advances in state-of-the-art optical character recognition (OCR) algorithms and large language models (LLMs) now enable the integration of previously overlooked data sources within these collections, including scanned textual documents and photographs. The unification of such diverse modalities within a single retrieval framework constitutes the primary contribution of this paper.

### 1.1. Related Work

Work on semantic retrieval has evolved from dense neural models for open-domain question answering, which replace keyword matching with learned vector representations of text (Karpukhin et al., 2020), to more fine-grained interaction mechanisms that improve semantic matching within purely textual collections (Khattab and Zaharia, 2020). Subsequent research extended retrieval beyond text by learning shared embedding spaces for images and language (Radford et al., 2021), and more recently by training multimodal large language models to act as universal retrievers across mixed text–image inputs (Lin et al., 2024). While these approaches advance semantic and cross-modal search, they typically assume relatively clean data and jointly trained embedding models. In contrast, our framework is designed for arbitrary settings: it is capable of processing heterogeneous and potentially noisy archival materials (e.g., video testimonies, scanned documents, photographs) by converting all modalities into textual form and combining semantic vector search with lexical matching, thereby prioritizing transparency, robustness, and practical applicability, for instance, in real-world cultural heritage collections.

## 2. Evaluation Data and Methodology

The proposed framework is designed to be domain-independent and applicable to heterogeneous multimodal collections. For demonstration purposes, we constructed an evaluation dataset grounded in the Holocaust domain. All experiments were conducted in English; however, the framework itself is language-agnostic.

We selected 17 publicly accessible testimony recordings from the public part of the *USC Shoah Foundation Dataset* (USC Shoah Foundation), each approximately 2.5 hours in length. The recordings were processed using our automatic speech recognition (ASR) engine, producing time-aligned transcripts. This constituted the first modality: video testimonies represented as textual segments.

During the interviews, witnesses frequently present photographs or documents to the camera. These were automatically extracted, resulting in 299 images. To further diversify the dataset, we selected 108 scanned historical documents from the *Arolsen Archives* (Arolsen Archives). Hence, in total, we obtained 407 images constituting the second modality.

All images were processed using parsers described in Sec. 3: Optical Character Recognition (OCR) was applied to extract textual content, and Large Language Model (LLM) captioning was used to generate semantic descriptions. After this step, all modalities were transformed into textual form that was subsequently segmented according to the chunking strategy described in Sec. 3.2. In total, we obtained 2,420 chunks serving as retrieval units.

### 2.1. Automatic Question Generation

To enable scalable evaluation without manual annotation, we employed the Asking-Questions (AQ) framework (Švec et al., 2024). For each out of the original 2,420 chunks, the framework generates multiple semantically grounded questions (approximately three per chunk) and may internally create sub-chunks aligned with each generated query, as shown in Table 1.

This process resulted in 7,249 query – sub-chunk pairs. Each generated query is considered relevant to its corresponding (sub-)chunk as well as to the

| | |
|---|---|
| original chunk | *One notable development is the organic connection which now exists between the SS and the Police. In 1956, when he was @ppointed Chief of the German Police, HIMMLUR was enabled to effect that fusion between the two forces he controlled which is & marked feature of Germany's internal security organisation today, fs J All senior police officers and many of the junior officers a@re also members of the SS, holding rank in both organisations. The official policy is to recruit new menbers of the police force solely from the SS. 'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.* |
| sub-chunk 1 | *One notable development is the organic connection which now exists between the SS and the Police. In 1956, when he was @ppointed Chief of the German Police, HIMMLUR was enabled to effect that fusion between the two forces he controlled which is & marked feature of Germany's internal security organisation today, fs J All senior police officers and many of the junior officers a@re also members of the SS, holding rank in both organisations. The official policy is to recruit new menbers of the police force solely from the SS.* |
| query 1 | `When was HIMMLUR appointed Chief of the German Police?` |
| sub-chunk 2 | *The official policy is to recruit new menbers of the police force solely from the SS. 'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.* |
| query 2 | `According to the official policy, from what organization were new police force members recruited?` |
| sub-chunk 3 | *'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.* |
| query 3 | `What organizations had effectively merged?` |

Table 1: Example output of the AQ framework on a selected ASR result sample, illustrating the generation of evaluation queries and the extraction of relevant sub-chunks (used in scenario B).

original chunk, forming the ground truth for a known-item retrieval task.

#### 2.1.1. Retrieval Setup

Chunks were indexed in a vector database using the proposed framework described in Sec. 3. For each generated query, we computed its embedding and performed top-$k$ retrieval.

We evaluated two scenarios:

(A) *Original chunk indexing:* Only the 2,420 original chunks were indexed. The AQ framework was used solely to generate evaluation queries (three per original chunk on average), which are all considered to semantically correspond to the same original chunk.

(B) *Sub-chunk indexing:* 7,249 AQ-generated sub-chunks were indexed, resulting in a one-to-one mapping between queries and indexed items.

Scenario A better reflects realistic deployment, where question generation is not part of the production pipeline.

### 2.2. Evaluation Metrics

We report Recall@$k$ (for $k \in \{1, 3, 5, 10\}$) and Mean Reciprocal Rank (MRR@$10$).

Recall@$k$ measures whether at least one relevant item appears among the top-$k$ retrieved results and so can be interpreted as the probability of finding the correct segment within the first $k$ returned results.

$$\text{Recall@}k = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\{\text{rank}_q \leq k\}, \qquad (1)$$

where $\text{rank}_q$ denotes the rank of the first relevant item for query $q$.

Mean Reciprocal Rank (MRR@10) evaluates how highly the first relevant result is ranked, considering only the top 10 retrieved items. If no relevant item appears within the top 10 results, the reciprocal rank is defined as zero.

$$\text{MRR@10} = \frac{1}{|Q|} \sum_{q \in Q} \begin{cases} \frac{1}{\text{rank}_q}, & \text{if } \text{rank}_q \leq 10, \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

### 2.3. Metrics Limitations

Although each generated query is guaranteed to be relevant to its originating chunk, it may also be semantically relevant to other chunks in the collection. Since no exhaustive manual relevance annotation was performed, such additional relevant matches remain undetected. Consequently, the reported metrics may slightly underestimate the true semantic retrieval performance.

## 3. Multimodal Search Framework

The concept of the proposed framework, illustrated in Fig. 1, is designed to provide a unified and extensible system for retrieving information from heterogeneous collections of data. Users interact with the system through a user-friendly web-based interface, allowing both conventional textual queries and exploration of the indexed multimodal content.

### 3.1. General Data Parser

The foundation of the proposed multimodal search framework is a modular data parsing pipeline designed to transform heterogeneous source files into a unified semantic representation. This pipeline, implemented in Python, uses a routing mechanism based on file extensions to dispatch documents to specialized parsers. Each parser is responsible for extracting textual information and, where applicable, spatial or temporal metadata, ensuring that the semantic context is preserved across different

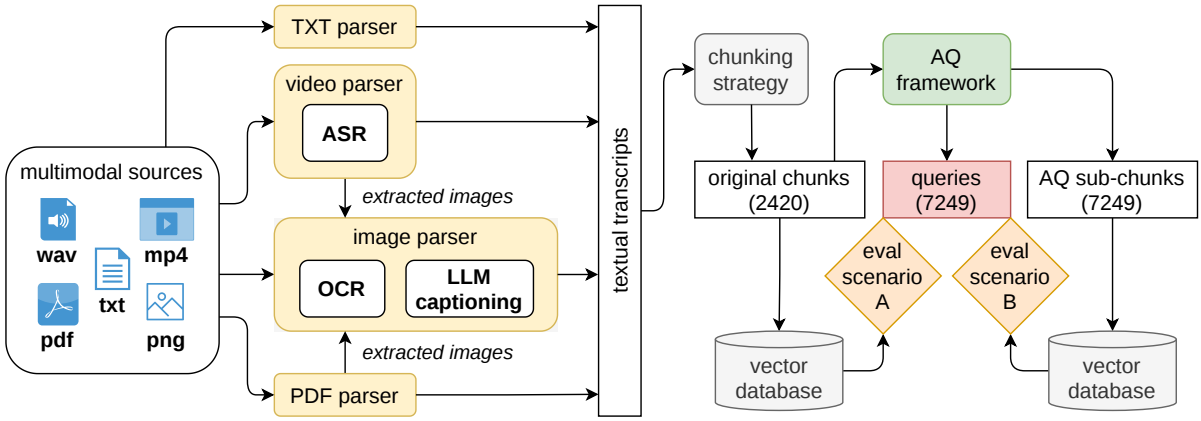


Figure 1: Overall concept and evaluation of the presented multimodal search framework.

modalities. The PDF parser (Section 3.1.3) is included in the pipeline described here to support future applications, but it is not evaluated on the dataset considered in this study.

#### 3.1.1. Audio Recordings

Audio files, typically in WAV or MP3 formats, are processed through an Automated Speech Recognition (ASR) module. The system utilizes the *UWebASR* API (Lehečka et al., 2023) to perform transcription of spoken content. The resulting transcript is processed according to the chunking strategy described in Section 3.2. Each resulting segment preserves its temporal boundaries (start and end timestamps), enabling precise localization within the audio stream during retrieval. This enables the search engine to pinpoint the exact segment within the audio file during retrieval.

#### 3.1.2. Images

Images are processed using a method that captures both their literal details and semantic content.

- *Optical Character Recognition (OCR):* We use *Tesseract OCR* (Smith, 2007) to extract any textual information present within the image. This is particularly crucial for scanned documents, infographics, or slides. Extracted text blocks are further processed using the unified chunking strategy described in Section 3.2. Each chunk retains its associated bounding box coordinates, enabling spatial localization during retrieval.
- *Semantic Captioning:* To capture the visual content of the image, we leverage a Large Language Model (LLM), specifically OpenAI's GPT-4o (OpenAI, 2024). The model generates a comprehensive textual description of the image content and identifies key objects along with their relative positions (top, center, bottom, etc.). The generation is guided by the constraints defined in the chunking strategy (Section 3.2), ensuring that each description forms a single indexable unit compatible with the embedding model.

#### 3.1.3. PDF Documents

PDF documents are handled by a structure-aware parser based on the *PyMuPDF* library (Artifex Software, Inc., 2026). The parser decomposes the document into text blocks and images.

- *Text Blocks:* Extracted text blocks are processed according to the chunking strategy described in Section 3.2. Structural metadata, including page numbers and bounding boxes, are preserved for each resulting unit. Short, non-informative segments (e.g., page numbers or artifacts) are filtered out to maintain the quality of the index.
- *Embedded Images:* Images embedded within the PDF are extracted and processed through the image parsing pipeline described in Section 3.1.2. This ensures that diagrams, charts, and illustrations within a document are fully indexed both by their textual content (via OCR) and their visual semantics (via captioning).

#### 3.1.4. Plain Text

Plain text files are parsed in a way that preserves their structural organization. Logical paragraphs are detected using line breaks and subsequently processed using the chunking strategy described in Section 3.2. The parser maintains the absolute line numbers associated with each resulting chunk, allowing direct referencing to the original source.

#### 3.1.5. Video Recordings

Video files are treated as complex multimodal data requiring both temporal and visual analysis.

- *Audio Transcription:* The audio stream is extracted using *FFmpeg* (FFmpeg Developers, 2026)and processed through the ASR module, identical to the standalone audio parser described in Section 3.1.1.

- *Visual Frame Analysis:* Visual content is obtained by extracting keyframes at a specified sampling rate (selected to be 0.1 frame per second). Each frame is then processed by the OpenAI GPT-4o model to generate semantic descriptions, as described in Section 3.1.2. This approach enables the system to index video content based on both spoken information and visual information over time.

### 3.2. Chunking Strategy

To ensure compatibility with the limited context window of embedding models, the pipeline employs a unified token-aware chunking strategy across all modalities. All textual data are segmented using a sliding window approach with a fixed length of 256 tokens and an overlap of 32 tokens. The tokenizer of the underlying embedding model, *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019), is used to guarantee that each chunk remains within the model's context window.

This strategy is applied across all modalities while respecting their logical structure:

- *Plain Text and OCR:* Chunks are created within logical boundaries such as paragraphs or OCR blocks to maintain semantic coherence.

- *Audio and Video:* Transcripts are segmented temporally, where each textual chunk preserves its corresponding start and end timestamps.

- *Semantic Captions:* For images and video frames, the chunk size acts as a constraint for the generative model, ensuring that descriptions are concise and ready for indexing as single units.

This multi-layered approach preserves necessary context and metadata (spatial and temporal) while providing the level of detail required for precise semantic retrieval.

### 3.3. Retrieval with Semantic Vectors

The core of our retrieval system is based on dense vector representations. We use the *all-MiniLM-L6-v2* sentence-transformer model (Reimers and Gurevych, 2019), which maps text chunks into a 384-dimensional dense vector space. This model provides a favorable trade-off between embedding quality and computational efficiency, making it suitable for large-scale semantic search.

For retrieval, we explore two approaches: a purely vector-based search using the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin, 2018) for efficient nearest-neighbor search, and a hybrid strategy that combines dense embeddings with traditional lexical features based on TF–IDF (Spärck Jones, 1972) to also capture exact keyword matches.

(i) *HNSW:* The index is implemented using the `IndexHNSWFlat` class from *Faiss* library (Douze et al., 2024). It is constructed with a connectivity parameter `M`=32 and a construction expansion factor `efConstruction`=200. During retrieval, the search expansion factor is set to `efSearch`=64 to balance efficiency and recall.

(ii) *Hybrid search:* The index combines retrieval scores from the dense HNSW index and a sparse TF-IDF index using a weighted sum:

$$S_{\text{hybrid}} = \alpha \cdot S_{\text{dense}} + (1 - \alpha) \cdot S_{\text{sparse}}, \quad (3)$$

where $S_{\text{dense}}$ and $S_{\text{sparse}}$ are the scores from the dense and sparse retrievers, respectively, and $\alpha \in [0, 1]$ is a weighting parameter (set to 0.5 in our experiments). Both indices apply $L_2$ normalization, ensuring comparable cosine similarity scores in the range $[0, 1]$.

The sparse TF–IDF index is constructed using the `TfidfVectorizer` class from the scikit-learn library (Pedregosa et al., 2011), configured with `min_df`=1, `max_df`=0.9, `max_features`=10000, and a list of stemmed English stopwords obtained using the Snowball stemmer provided by the *NLTK* library (Bird et al., 2009).

## 4. Results

From a broader perspective, three evaluation layers can be identified for the proposed framework: (1) content extraction quality (ASR, OCR, LLM-based captioning), (2) chunking strategy and information representation, and (3) cross-modal semantic retrieval.

In this paper, we focus exclusively on the third layer, i.e., the ability of the system to retrieve the correct multimodal segment given a textual query. It should be noted, however, that retrieval performance is inherently influenced by the quality of upstream processing stages.

We evaluated two retrieval strategies described in Section 3.3: (i) vector-based retrieval using

HNSW indexing and (ii) hybrid search combining semantic vectors with keyword-based matching. Furthermore, we report results for the two evaluation scenarios introduced in Section 2.1.1: (A) original chunk indexing and (B) sub-chunk indexing.

### 4.1. Scenario A: Original Chunk Indexing

Results for the more realistic setup (7,249 queries vs. 2,420 original chunks) are shown in Table 3.

In this scenario, multiple queries map to the same larger chunk, resulting in a slight decrease in performance compared to Scenario B, where each (shorter) sub-chunk corresponds to a single query. As expected, hybrid search consistently outperforms HNSW across all metrics.

Importantly, Recall@10 remains above 0.60 for video transcripts and above 0.95 for OCR documents in the hybrid configuration. This indicates that even under realistic indexing conditions, the system is capable of retrieving the correct multimodal segment within a small set of top-ranked results.

### 4.2. Scenario B: Sub-chunk Indexing

Table 2 reports results for the AQ-level indexing setup (7,249 queries vs. 7,249 indexed sub-chunks).

Across all modalities, hybrid search consistently outperforms pure HNSW vector retrieval. The improvement is particularly visible in Recall@1 and MRR, indicating that hybrid search more frequently ranks the correct segment at the very top of the result list. This suggests that combining lexical matching with semantic similarity helps stabilize retrieval when queries are closely aligned with the original wording of the source segment. This behavior is captured in Table 4, which presents the top-5 retrieved chunks for the query. In the table, the desired chunk is highlighted in bold. When using the pure HNSW index, this chunk does not appear even among the top-10 results. In contrast, under the hybrid setup, the required chunk is ranked second.

OCR-based documents achieve the highest scores overall (e.g., Recall@10 above 0.94 and MRR around 0.80 in the hybrid setup). This can likely be attributed to the relatively well-structured and information-dense nature of scanned documents, where generated questions often correspond to explicit factual statements.

In contrast, video transcripts (ASR modality) show lower Recall@1 and MRR. This may reflect the more narrative and less structurally explicit nature of spoken testimonies or finding semantically close passages from another testimonies, as the structure of the interviews is unified. Although the required chunk according to the evaluation protocol is not retrieved, several returned passages indicate clear semantic relevance to the query, as assumed in Section 2.3.

Caption-based image representations provide useful semantic summaries for pictures with unique activities, however, LLM-based descriptions of scanned documents, for instance, can confuse the retrieval system significantly. Therefore, we additionally report an evaluation of the LLM-based captioning parser excluding scanned documents, denoted as *Capt.** in Tables 2 and 3. The results show a clear improvement in performance under this setting.

### 4.3. Efficiency Considerations

Due to offline preprocessing (ASR transcription, OCR, caption generation, chunking, and indexing), online retrieval is computationally lightweight. On a standard CPU machine, average query latency remains below 500 ms for both retrieval strategies.

The computationally intensive stages are data parsing and index construction, which are performed only once during preprocessing. Parsing the complete dataset of 424 files including `.txt`, `.jpg`, and `.png` files) required 36 minutes in total on a standard CPU machine. Index construction scales linearly with the number of text chunks, with an average build time of approximately 2.25 s per 100 chunks.

## 5. Conclusion

In this paper, we have presented a general framework for semantic search across heterogeneous multimodal collections. Our evaluation, conducted on a constructed dataset from the Holocaust domain, demonstrates that the system is capable of retrieving relevant segments across multiple modalities, including ASR transcripts of video testimonies, OCR-processed documents, and LLM-generated image captions.

Two retrieval strategies were compared (HNSW vector search and hybrid semantic-lexical search), and two evaluation scenarios were explored (original chunk vs. sub-chunk indexing). The results indicate that hybrid search consistently outperforms pure vector-based retrieval, and that even under realistic indexing conditions, the system retrieves the correct segment within a small set of top-ranked results with high reliability. Overall, the results demonstrate that semantic search over heterogeneous multimodal collections is feasible and reasonably robust, even when different modalities exhibit varying levels of textual quality and structural consistency.

| Source | Queries | Chunks | HNSW | | | | | Hybrid search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | k1 | k3 | k5 | k10 | MRR | k1 | k3 | k5 | **k10** | MRR |
| ASR | 5700 | 1530 | 0.17 | 0.30 | 0.36 | 0.44 | 0.25 | 0.32 | 0.49 | 0.56 | **0.64** | 0.42 |
| OCR | 716 | 533 | 0.60 | 0.75 | 0.80 | 0.87 | 0.69 | 0.73 | 0.88 | 0.92 | **0.96** | 0.81 |
| Capt. | 833 | 357 | 0.19 | 0.31 | 0.38 | 0.47 | 0.27 | 0.30 | 0.43 | 0.52 | **0.61** | 0.39 |
| Capt.* | 601 | 249 | 0.21 | 0.34 | 0.42 | 0.51 | 0.30 | 0.32 | 0.47 | 0.56 | **0.66** | 0.41 |
| All | 7249 | 2420 | 0.19 | 0.31 | 0.36 | 0.44 | 0.26 | 0.32 | 0.48 | 0.54 | **0.62** | 0.41 |

Table 2: Retrieval performance by modality – Scenario A (2420 original chunks)

| Source | Queries | Chunks | HNSW | | | | | Hybrid search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | k1 | k3 | k5 | k10 | MRR | k1 | k3 | k5 | **k10** | MRR |
| ASR | 5700 | 5700 | 0.22 | 0.36 | 0.41 | 0.49 | 0.30 | 0.32 | 0.50 | 0.57 | **0.64** | 0.43 |
| OCR | 716 | 716 | 0.60 | 0.77 | 0.82 | 0.88 | 0.69 | 0.71 | 0.89 | 0.91 | **0.95** | 0.80 |
| Capt. | 833 | 833 | 0.24 | 0.35 | 0.42 | 0.51 | 0.31 | 0.34 | 0.46 | 0.52 | **0.63** | 0.42 |
| Capt.* | 601 | 601 | 0.26 | 0.38 | 0.46 | 0.55 | 0.34 | 0.34 | 0.49 | 0.55 | **0.67** | 0.43 |
| All | 7249 | 7249 | 0.25 | 0.38 | 0.44 | 0.51 | 0.33 | 0.34 | 0.51 | 0.57 | **0.65** | 0.44 |

Table 3: Retrieval performance by modality – Scenario B (7249 chunks from the AQ framework)

### 5.1. Future Work

Several directions for future work can further enhance and generalize the proposed framework:

- *Evaluation on different domains:* Extending the evaluation beyond Holocaust-related data to other cultural heritage collections or entirely different domains.
- *Cross-modal query capabilities:* Enabling retrieval not only via textual queries but also through queries based on images or audio segments. For example, a user could search for video segments similar to a given sound recording, or find documents semantically related to a sample image.
- *Enhanced evaluation metrics:* Incorporating more sophisticated metrics capturing partial relevance or cross-modal semantic similarity, especially for collections where multiple relevant segments may exist for a single query.

These extensions would further improve the usability and applicability of the framework in digital humanities contexts, providing researchers with flexible and semantically aware access to diverse multimodal archives.

## 6. Acknowledgements

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2025-011.

## 7. Bibliography

Arolsen Archives. Arolsen archives: International center on nazi persecution. Online archive of more than 30 million historical documents related to Nazi persecution.

Artifex Software, Inc. 2026. Pymupdf. Python bindings for the MuPDF PDF toolkit. Version X.Y.Z.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Martin Bulin, Jan Svec, Pavel Ircing, Adam Fremund, and Filip Polak. 2025. Exploring oral history archives using state-of-the-art artificial intelligence methods. *Acta Informatica Pragensia*, 14(2):207–214.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

FFmpeg Developers. 2026. Ffmpeg multimedia framework. Version 7.1 (or latest version).

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Wen-tau Yih, Yann Dauphin, and Wilker Yala. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

| Rank | Score | Top-5 chunks for query *What was located approximately twenty miles from Berlin?* |
|---|---|---|
| 1 | 0.3644 | But worst of all was in Berlin. I didn't get this in Kongsberg. I don't think they were organized. Yet that way Berlin they had |
| **2** | **0.3519** | **Is father. Her father was born in a little town about twenty miles from Berlin and known now and and they had properties a lot of proves. They had a so call ia** |
| 3 | 0.3501 | have intellectually known something like that was coming. It was inevitable even even the most assimilated would be at some point hit with this. How far it would go? Of course, nobody knew at that time, but this was the first time it really came home to him. There was his sign his name be smerched by these vermin. And what approximate date was that? Thirty four or five, fairly shortly after the Nazis came to power? In thirty three, it was at that time that he met. And now go on to answer your question why did we leave Konigsberg and went to Berlin? |
| 4 | 0.3466 | Her father was born in a little town about twenty miles from Berlin and known now and and they had properties a lot of proves. They had a so call iagod Rita good. It's a huge farm. There were ten thousand actors of land. We're talking about twenty five thousand acres. They. The father had a lot of comments and industry. He was selling coal for heating in Berlin. The grandfather had a raising cows, horses, pigs and supplying meat to Berlin and a dirty all kind of dirty cheese, milk or whatever. Then they had a factory, they made a food for horses and he used to raise horses for the Wehrmacht. |
| 5 | 0.3450 | That was the reason we moved to Berlin. And can you describe briefly what your home was like in Berlin and what the area that you lived in was like? Well, I lived in Bilmastorf, which is West Berlin. We lived in a typical Berlin four story building and I have a picture of it. We had, I don't know, ten twelve room apartment. There were two apartments to each floor, an elevator landing and very conveniently, the other people who lived on the same floor as we did were also Jewish, which was wonderful. |

Table 4: Top-5 retrieved chunks using the *(ii) Hybrid search approach* (see Section 3.3)


Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of SIGIR*.

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based speech recognition models for oral history archives in english, german, and czech. In *Proc. Interspeech 2023*, pages 201–205.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Yu. A. Malkov and D. A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs.

OpenAI. 2024. GPT-4o. `https://openai.com`. Large multimodal language model.

Pavel Pecina, Petra Hoffmannová, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. 2007. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, pages 674–686. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jan Švec, Luboš Šmídl, and Jan Lehečka. 2023. Transformer-based encoder-encoder architecture for spoken term detection. In *Pattern Recognition*, pages 346–357, Cham. Springer Nature Switzerland.

USC Shoah Foundation. Visual history archive. Digital archive of video testimonies and associated metadata, maintained by USC Shoah Foundation.

Jan Švec, Martin Bulín, Adam Frémund, and Filip Polák. 2024. Asking questions framework for oral history archives. In *Advances in Information Retrieval – 46th European Conference on Information Retrieval, ECIR 2024, Proceedings, Part III*, volume 14610 of *Lecture Notes in Computer Science*, pages 167–180. Springer.

# Automatic Transcription of Holocaust Testimonies in Yiddish: Orthographic Comparison and Cross-Domain Validation

**Isaac L. Bleaman**
University of California, Berkeley
Department of Linguistics and Center for Jewish Studies
bleaman@berkeley.edu

**Abstract**

The digital processing of Holocaust testimony interviews is essential for the long-term preservation and accessibility of survivors' narratives. However, automatic speech recognition (ASR) for Yiddish—the primary language of most Holocaust victims and survivors—remains underdeveloped. This paper introduces the first ASR system for European Yiddish, focused on the Northeastern ("Lithuanian") dialect and trained and evaluated on testimony interviews from the Corpus of Spoken Yiddish in Europe (42 hours of speech segments from 60 survivors). A systematic comparison of CTC-based ASR models using transcripts with different orthographic representations reveals that a Hebrew-based phonemic system with precomposed Unicode is optimal, achieving a mean word error rate (WER) of 37.96% compared to 59.40% WER for romanized Yiddish and 99.67% WER (catastrophic failure) for standard Yiddish spelled with decomposed Unicode. Cross-domain testing on Yiddish audiobooks provides additional support for a phonemic representation (27.07% WER, 6.56% CER). Together, the results suggest that automatic transcription developed from oral Holocaust testimonies can support further technological innovation in service of Yiddish-speaking communities.

**Keywords:** ASR, Yiddish, Holocaust testimonies, orthographic normalization

## 1. Introduction

Audio- and video-recorded testimony interviews are unparalleled resources for understanding the Holocaust through the firsthand accounts of survivors. The USC Shoah Foundation Visual History Archive holds tens of thousands of digitized interviews with Holocaust survivors, which were collected mostly in the 1990s in locations all around the world and delivered in dozens of languages. Automatic speech recognition (ASR) software has been developed to make the content of many of these testimonies searchable and viewable as subtitles in video players. However, these technological advances have not benefited all languages in equal measure. No such ASR capability exists for Yiddish, the primary language of most Eastern European Jewish communities destroyed in the Holocaust (Birnbaum, 2016, 42), as well as a significant number of recorded survivor testimonies. This constitutes a major barrier for the accessibility of Yiddish-language testimonies, which affects not only historians of the Holocaust but also Yiddish linguists and language learners.

Developing ASR for Yiddish-language testimonies presents both linguistic and technical challenges. Although the language continues to be used in Jewish communities around the world, most of the dialects once spoken across the pre-Holocaust European heartland are now severely endangered, and in some cases underdocumented. Furthermore, Yiddish-speaking survivors lived in highly multilingual environments, both before the Holocaust and in their post-war countries of resettlement. This makes for oral testimonies that are highly complex, in which survivors routinely engage in code-switching (with languages as diverse as Hebrew and Hungarian) as well as dialect mixing. An effective ASR system would therefore need to handle multiple Yiddish dialects and extensive language mixing.

Beyond these challenges, Yiddish presents a great deal of orthographic complexity. While it is traditionally written in a Hebrew-based orthography, Yiddish can also be transliterated into other alphabets (Latin, Cyrillic, etc.). Within Hebrew script, there is a convention in which all words are spelled phonemically unless they come from the so-called "Semitic component" (Hebrew- and Aramaic-origin words), in which case they are spelled according to the norms of those languages (Jacobs, 2005, 48). Additionally, numerous orthographies (not all standardized) have been in use in different times and places, and in today's digital texts, there are also competing Unicode normalization forms. The interaction between orthographic representations and modern neural ASR architectures remains unexplored for Yiddish, yet this choice has significant downstream consequences for model training and performance.

This paper presents a proof-of-concept ASR system for Northeastern Yiddish, also known as *Litvish* 'Lithuanian' Yiddish, a cluster of dialects spoken across a territory that includes present-day Lithuania, Latvia, Belarus, northeastern Poland, and northern and eastern Ukraine (Weinreich, 1963, 337; Jacobs, 2005, 65). Northeastern Yiddish was chosen because standard Yiddish



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 20–28
11 May 2026. 

spelling (and standard romanization) is more-or-less phonemically transparent for this dialect. Using audio recordings and transcripts from the Corpus of Spoken Yiddish in Europe (CSYE; Bleaman and Nove, 2025), we investigate which orthographic representations enable effective ASR training. More specifically, this work makes the following contributions:

1. We systematically compare three orthographies for representing Yiddish speech during ASR training and evaluation: romanization (ROM); a standardized Hebrew-based script, with phonemic spellings of all words in decomposed Unicode (STD); and a Hebrew-based representation of Yiddish phonemes in precomposed Unicode (PHON). Our experiments demonstrate that the PHON system—which can readily be back-transformed into a more human-readable standard Yiddish spelling—achieves a much lower word error rate (WER) than the ROM system.

2. We identify a critical incompatibility between STD, which uses Hebrew letters and combining diacritics, and Connectionist Temporal Classification (CTC)-based ASR training. The STD approach fails catastrophically, producing unintelligible output. Given that decomposed Unicode is standard for digital Yiddish text today, this finding has immediate implications for corpus development, and it also extends to other languages that use combining characters to capture phonemic distinctions.

3. We validate the robustness of an ASR system trained on Holocaust testimony interviews through a cross-domain evaluation of Yiddish audiobooks, using a dataset compiled for text-to-speech (TTS) applications (Webber et al., 2022; Bleaman et al., 2023). Despite being trained on spontaneous speech, our ASR system generalizes to read speech as indicated by WER.

4. We provide resources to support future work: a trained ASR model, orthographic preprocessing utilities, and an interactive demo.

The results of this project demonstrate that effective ASR for Yiddish-language testimonies is achievable but highly dependent on specific design choices, which have consequences for other technologies developed for Yiddish-speaking communities. The remainder of this paper is organized as follows: Section 2 describes the testimonies and datasets used for model training and evaluation. Section 3 details the orthographic representations chosen for training. Section 4 presents the model architecture, training configuration, and evaluation methodology. Section 5 reports results and discusses findings and implications. Section 6 concludes with limitations and future directions. Brief statements about data availability and ethical considerations are provided before the reference lists.

## 2. Data

### 2.1. The Speech Corpus

The primary data for this project come from the Corpus of Spoken Yiddish in Europe (CSYE; Bleaman and Nove, 2025), a collection of manually transcribed Holocaust survivor testimonies in Yiddish sourced from the USC Shoah Foundation Visual History Archive (VHA). At the time of model training, the corpus contained interviews with 60 speakers of Northeastern Yiddish. Like other testimonies in the VHA, these interviews in Yiddish were conducted by trained volunteers in locations all around the world, and generally proceed chronologically as survivors recount their personal and family histories before, during, and after World War II. In addition to the complexities outlined above related to dialect and language mixing, the conversational nature of these interviews—including overlapping speech between survivor and interviewer, disfluencies including filled pauses, and moments of emotional intensity—presents important ASR challenges that are not typical for read or scripted speech.

The CSYE includes downloadable audio files extracted from VHA video files (digitized video cassettes). These were converted to 16kHz mono WAV format. Transcripts in the CSYE are annotated as *reviewed* or *unreviewed*, reflecting whether or not the transcript for a particular video cassette was reviewed by a member of the CSYE team other than the original transcriber. We included both reviewed and unreviewed segments in our training and testing to maximize coverage.

Aside from speaker diarization, which is the result of a machine learning algorithm and manual correction, all of the transcripts in the CSYE were produced by hand by Yiddish-speaking team members trained in the CSYE transcription conventions. The survivor and interviewer are transcribed on separate time-aligned text tiers in ELAN files (Max Planck Institute for Psycholinguistics, 2021). Transcription conventions are based on standard YIVO transliteration, an orthographic representation in a Latin character set widely used in the Yiddish scholarly community (YIVO, 1999; Bleaman, 2019). CSYE conventions instruct transcribers to faithfully transcribe dialectal vocabulary items, but not to modify spellings to reflect historical sound changes that predictably differenti-

ate the dialects. For example, the written form ‹beygl› 'bagel(s)' can represent either /beɪgl/ in Northeastern and Southeastern Yiddish or /baɪgl/ in Central Yiddish.[1] Because the corpus was designed (in part) to support research in sociophonetics, the transcripts include faithful representations of partial words, filled pauses, and other disfluencies—elements that are often omitted from ASR output. More information on CSYE transcription methodology is documented in Bleaman and Nove (2025).

### 2.2. Data Preprocessing

We segmented the audio and transcripts into short phrase-level chunks, based on the segmentation already present in the CSYE. Only speech from the survivors, not the interviewers, was included in our dataset. Because the current project was envisioned to be a proof-of-concept for a Yiddish ASR system, we applied filtering to remove the following speech segments:

- **Overlapping speech:** Segments produced by the survivor that overlapped with segments produced by the interviewer
- **Unclear or misheard words:** Segments containing *UNK* (convention for words that were unintelligible to the transcriber) or angle brackets (convention for uncertain transcriptions)
- **Partial words:** Segments containing any word strings ending in a hyphen (convention for partial words)
- **Fillers:** Segments containing one or more predefined filled pauses (*uh*, *uhm*, *ehm*, etc.)
- **Borrowings:** Segments containing words marked as borrowings (those with 2+ adjacent uppercase letters)
- **Short segments:** Segments shorter than 0.5 seconds

Of the 114,092 total speech segments from Northeastern Yiddish-speaking survivors, 63,346 segments (55.5%) remained after these filtering steps. This corresponds to 42.21 total hours of isolated speech segments.

We then applied several orthography-specific text preprocessing steps. These included whitespace normalization, punctuation removal, replacing remaining hyphens (those used in compounds) with spaces, and various orthography-specific character transformations, which are detailed in Section 3.

[1] In this paper, angle brackets are used to represent orthographic forms. Slashes represent phonemic forms.

Finally, we partitioned speakers (*not* segments or tape transcripts) into a training set (70%: 42 speakers), a validation set (10%: 6 speakers), and a test set (20%: 12 speakers) using a fixed random seed. This ensures that test speakers are completely unseen during model training and can be used to evaluate how well the ASR system generalizes to new voices. A fixed random seed ensures that the same speaker partition is used across all three orthographic representations for a fair comparison.

### 2.3. Cross-Domain Corpus

For cross-domain validation of an ASR system trained on spontaneous conversational speech, we used the Reading Electronic Yiddish Documents (REYD) corpus of audiobook narrations, which was assembled for a project to create a text-to-speech (TTS) dataset and system for Yiddish (Webber et al., 2022; Bleaman et al., 2023). The dataset consists of short audio segments matched with text files from readings of Yiddish literature, which were taken from two different public repositories: the Yiddish Book Center's Sami Rohr Library of Recorded Yiddish Books, originally recorded in the 1980s and 1990s in Montreal, and the "World of Yiddish" webpage, recorded in the early 2000s at the University of Haifa. From the dataset, we used recordings from the speakers labeled *lit1* and *lit2* (two Northeastern Yiddish-speaking narrators: Sara Blacher-Retter and Leib Rubinov) and the set of utterances labeled as *yivo_respelled* (those written in a Hebrew-based script, with all words spelled phonemically). The entire subcorpus for these two narrators was used for cross-domain testing. For each orthographic model, we transformed the REYD reference texts using the same preprocessing pipeline applied to the CSYE data.

Table 1 provides summary statistics for all datasets used in this study.

## 3. Orthographic Representations

As mentioned above, Yiddish can be written using multiple orthographic systems, and even within a single system, users can apply various encoding and normalization choices. For ASR development, the choice of orthographic representation constrains the model's output vocabulary and plays an important role in training. In this project, we systematically compare three approaches, which involve two different scripts (Latin-based vs. Hebrew-based) and Unicode normalization strategies, as well as other Yiddish-specific choices that are elaborated on in this section.

| Dataset | Speakers | Hours | Segments | Domain |
|---|---|---|---|---|
| CSYE training | 42 | 30.83 | 46,432 | Testimony (conversational speech) |
| CSYE validation | 6 | 2.80 | 3,803 | Testimony (conversational speech) |
| CSYE test | 12 | 8.58 | 13,111 | Testimony (conversational speech) |
| REYD | 2 | 5.32 | 3,632* | Audiobooks (read speech) |

Table 1: Corpus statistics after preprocessing and data partitioning. All CSYE orthographic representations use identical speaker partitions. *REYD segment counts vary somewhat by orthography, due to the respelling of reference texts and filtering rules.

### 3.1. Romanization (ROM)

The CSYE transcripts are originally produced in romanized Yiddish adapted from YIVO conventions for transliteration, and this orthographic representation serves as the baseline for our experimentation. After the preprocessing and filtering steps outlined above, we convert the remaining transcribed segments to lowercase. The resulting vocabulary contains 24 characters: 21 letters (‹a›–‹z› excluding ‹c j q w x›), word boundary marker, and standard special tokens for padding and unknown characters.

While the ROM system is phonemically transparent, it is not a one-to-one mapping of grapheme to phoneme; many Latin letter combinations correspond to a single phoneme. For example, ‹tog› ‘day’ corresponds to /tog/, but diphthongs and certain consonants are represented by multiple characters each, e.g., ‹boykh› ‘stomach’ corresponds to /bo͡ɪχ/.

### 3.2. Standard Hebrew-Based with Decomposed Unicode (STD)

To create Hebrew-script representations, we automatically respelled the original romanized transcripts using the `detransliterate()` function from the `yiddish` library (Bleaman, 2024). This uses rule-based pattern matching to convert standard YIVO transliteration into the Hebrew-alphabetic script, without correcting the spelling of words of Semitic origin (i.e., these are spelled phonemically). This output is then fed into the `replace_with_decomposed()` function to represent vowel marks and other *nekudes* (“pointing”) as combining diacritics with preceding letter graphemes. Further, the argument `vov_yud=True` is specified to produce a few Yiddish-specific ligatures (‹וו וי יי›). With the exception of the phonemic spelling of Semitic-origin words, the output of all of these steps reflects how standard Yiddish is typically encoded in most digital documents today.

While a Hebrew-based orthography addresses *some* of the many-to-one character-to-phoneme mappings of the ROM system—e.g., the consonant /χ/ corresponds to ‹kh› in ROM but to the singleton grapheme *khof* ‹כ› in STD—the use of combining diacritics means that an even larger number of sounds are represented by multiple Unicode characters. For example, the consonant /f/ becomes ‹פּ›, a two-character sequence consisting of a plain *fey* ‹פ› followed by the *rofe* diacritic. Additionally, a silent *alef* ‹א› is required in many vowel-initial words, and five consonant phonemes have special letter forms (distinct Unicode allographs) when they appear in word-final position. For example, the word /o͡ɪχ/ ‘also’ (romanized as ‹oykh›) is represented in STD as ‹אויך›, which begins with a silent *alef* and ends with the word-final allograph of *khof* ‹כ›.

The vocabulary contains 36 characters: Hebrew base letters (excluding ‹ח› and ‹ת›, which only appear in Semitic spellings), the combining diacritics used in standard Yiddish (those seen here: ‹אַ אָ יִ פּ שׂ›), word-final allographs (‹ך ם ן ף ץ›), Yiddish ligatures ‹וו וי יי›, word boundary, and standard special tokens.

### 3.3. Phonemic Hebrew with Precomposed Unicode (PHON)

In anticipation of the problems that might arise from the use of decomposed Unicode with a CTC-based ASR system, we created a Hebrew-based phonemic representation in precomposed Unicode characters. This differs from STD in the following ways:

- All letters with combining diacritics, e.g., ‹אַ› (U+05D0 for *alef* plus U+05B7 for *pasekh*), are replaced with precomposed equivalents from the “Alphabetic Presentation Forms” Unicode block of ligatures, e.g., ‹אַ› (U+FB2E).
- Silent *alef* letters are removed throughout.
- Word-final allographs are replaced with their nonfinal forms.
- The consonant /j/ and the vowel /i/, which are both (usually) represented in STD by the letter *yud* ‹י›, are distinguished in PHON: ‹י› for the consonant and ‹יִ› (U+FB1D) for the vowel.

- The vowel /u/, which is represented in STD either as ‹ו› or as ‹וּ› depending on context, is consistently represented in PHON as ‹וּ› (U+FB35).

The result of these modifications is a system of one-to-one mappings between Unicode characters and phonemes. While this differs from the way Yiddish is customarily represented in digital text, it can readily be converted back to a familiar, human-readable format by performing "roundtrip" transliteration, i.e., by first applying the `yiddish` library's `transliterate()` function and then the `detransliterate()` function.

The vocabulary for PHON contains 30 characters: several Hebrew base letters (‹ב ג ד ה ז ט י כ ל מ נ ס ע צ ק ר ש›), Yiddish ligatures (‹וו וי יי›), Hebrew letters precomposed with diacritics (‹אַ אָ אּ בּ וּ יִ פּ פֿ›), word boundary, and standard special tokens.

A comparison of orthographic representations for an example Yiddish utterance from the CSYE is provided in Table 2.

## 4. Experimental Setup

### 4.1. Model Architecture

We built upon Wav2Vec-BERT 2.0 (Seamless Communication et al., 2023), a large-scale self-supervised speech encoder developed by Meta, which was pretrained on over 4.5 million hours of unlabeled multilingual audio. The model uses convolutional acoustic feature extraction followed by a conformer-based encoder, producing high-dimensional speech embeddings that generalize across languages and downstream tasks. To adapt the model to Yiddish, we followed the fine-tuning procedure of Lacombe (2024), which covers dataset preparation, CTC-based vocabulary construction, processor setup, and CTC-based fine-tuning, with a convolutional adapter layer added.

The base model contains approximately 600 million parameters. We initialized from the publicly available *facebook/w2v-bert-2.0* checkpoint and fine-tuned all parameters on the CSYE dataset described above. The CTC training objective assumes a monotonic alignment between acoustic frames and output characters, a constraint that becomes critical for evaluating different orthographic representations (see Section 5).

### 4.2. Training Configuration

All models were trained with identical hyperparameters (Table 3), adapted from Lacombe (2024) and held constant across all three orthographies. No data augmentation (e.g., noise injection, speed perturbation) was applied during fine-tuning. We fine-tuned for up to 10 epochs with early stopping based on validation WER. Models were evaluated every 300 steps, with training stopping if no improvement occurred for 3 consecutive evaluations. Training used mixed-precision (FP16) and gradient checkpointing to reduce memory consumption. Total training time ranged from 1.3 to 1.7 hours on a Tesla T4 GPU.

### 4.3. Evaluation Metrics

We report word error rate (WER) and character error rate (CER) on the held-out test set. WER measures the edit distance between predicted and reference word sequences normalized by the number of reference words:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Reference Words}} \tag{1}$$

CER applies the same metric at the character level. Unlike WER, which disproportionately penalizes misplaced whitespace, CER provides a more forgiving measure of accuracy by treating word-boundary errors as single-character edits rather than complete word-level mismatches. While WER is the standard metric, and the one used for evaluation during training, we report both in Section 5.

Initial experiments with a single random seed (42) across all three orthographies identified PHON as best-performing. We therefore trained PHON with four additional random seeds (43–46) to quantify initialization variance and establish statistical confidence. ROM (the baseline orthography) and STD (which failed to produce intelligible output; see below) were trained just once.

Therefore, for the PHON models, we report mean ± standard deviation across the five random seeds, but report a single value for ROM and STD. Cross-domain evaluation using the REYD audiobook corpus reports the same metrics, where the reference texts undergo the same orthographic transformations defined above in Section 3.

## 5. Results and Discussion

### 5.1. In-Domain Performance (CSYE)

Table 4 presents our main results on the CSYE test set. The phonemic Hebrew representation (PHON) achieved the best performance with a mean 37.96% WER, representing a 36.1% relative improvement over the romanization baseline (59.40% WER). This substantial improvement is consistent with the hypothesis that one-to-one mapping between phonemes and graphemes optimizes CTC training for Yiddish ASR.

| Orthography | Alphabet | Example |
|---|---|---|
| Original (from the CSYE) | Latin | mayn familye-nomen iz Dimantshteyn |
| ROM (romanized) | Latin | mayn familye nomen iz dimantshteyn |
| STD (standard Hebrew-based) | Hebrew, decomposed | מײַן פֿאַמיליע נאָמען איז דימאַנטשטיין |
| PHON (phonemic Hebrew-based) | Hebrew, precomposed | מײַנ פֿאַמִיליע נאָמען יִז דִימאַנטשטיינ |

Table 2: An utterance from the Corpus of Spoken Yiddish in Europe, as represented in the training data for each orthographic system after preprocessing. The utterance comes from the testimony of Holocaust survivor Aizik Dimantstein (1996).

| Parameter | Value |
|---|---|
| Base model | w2v-BERT 2.0 |
| Optimizer | AdamW |
| Effective batch size | 32 |
| Learning rate | $5x10^{-5}$ |
| LR scheduler | Cosine |
| Warmup steps | 1,000 |
| Max epochs | 10 |
| Evaluation frequency | Every 300 steps |
| Early stopping patience | 3 evaluations |
| Precision | FP16 |

Table 3: Training hyperparameters. All models use identical speaker-based data splits. Five PHON models were trained with different random seeds; ROM and STD each trained with a single random seed.

| Model | WER (%) | CER (%) |
|---|---|---|
| ROM | 59.40 | 18.73 |
| STD | 99.67 | 93.21 |
| PHON | 37.96 ± 0.77 | 13.39 ± 0.46 |

Table 4: Test set performance on CSYE (13,111 segments from 12 unseen speakers). PHON results show mean and standard deviation across five random seeds.

The five PHON models showed consistent performance across random seeds, with WER ranging from 37.22% (seed 44) to 39.14% (seed 45). The small standard deviation indicates that our results are robust to initialization variance. All five seeds achieve substantial improvements over the ROM baseline.

The ROM model established a baseline to demonstrate that ASR for Yiddish testimonies is feasible even with the default Latin-script representations from the CSYE. However, the Hebrew orthography with decomposed Unicode (STD) catastrophically failed, with 99.67% WER and 93.21% CER—producing text largely consisting of isolated diacritics, repeated characters, and empty strings.

This failure presumably stems from the use of a decomposed Unicode character set for Yiddish. CTC assumes a roughly monotonic alignment between acoustic frames and output tokens, but decomposition often splits single phonemes like /a/ into multiple code points (a letter plus a combining diacritic). In such cases, the model must predict multiple sequential tokens for essentially the same acoustic span. Other factors, such as the use of silent *alef*s or multiple allographs to represent the same phoneme (e.g., word-initial and -medial ‹כ› vs. word-final ‹ך›), could increase variability in the target sequence, but they do not fundamentally contradict the temporal alignment assumptions in the same way as decomposition.

| Model | WER (%) | CER (%) |
|---|---|---|
| ROM | 51.84 | 10.53 |
| STD | 98.98 | 84.17 |
| PHON | 27.07 ± 2.99 | 6.56 ± 0.69 |

Table 5: Cross-domain performance on REYD audiobooks (2 speakers, 5.32 hours). PHON results show mean ± standard deviation across five random seeds.

### 5.2. Cross-Domain Performance (REYD)

Table 5 summarizes the results of applying the ASR models trained on transcribed Holocaust testimonies to the REYD audiobook corpus, and Table 6 provides example ASR outputs. Remarkably, the PHON models achieve better performance on the REYD dataset (27.07% WER) than on the CSYE test set (37.96% WER), demonstrating robust cross-domain generalization. This improvement likely reflects the slower, more careful speaking style of audiobook narration compared to the spontaneous speech of testimony interviews. Another relevant factor may be the reduced background noise variability in the audiobooks, which were recorded in a studio environment rather than in the speakers' homes. In any event, the cross-domain improvement suggests that the model has learned generalizable acoustic phonetic patterns of Northeastern Yiddish rather than testimony-specific characteristics.

The ROM baseline also improved on REYD (51.84% WER vs. 59.40% on the CSYE), while STD's failure remained consistent across domains (98.98% WER on REYD and 99.67% on CSYE). For example, the STD model produced 537 completely empty predictions out of 3,632 REYD utter-

ances, further confirming that its failure is systematic rather than dataset-specific.

The strong cross-domain performance suggests that PHON-based models can be applied to diverse Yiddish audio collections beyond Holocaust testimonies. The WER on REYD approaches the performance levels where ASR can become practically useful for searching and information extraction—all the more so if the output is manually corrected.

## 6. Conclusion, Limitations, and Future Directions

Holocaust testimony archives hold thousands of hours of interviews in Yiddish, yet these recordings remain largely unsearchable and inaccessible for large-scale analysis. We address this barrier by developing the first automatic speech recognition system for Northeastern Yiddish, trained on transcribed survivor testimonies from the Corpus of Spoken Yiddish in Europe.

Our phonemic Hebrew orthography (PHON) achieves a mean WER of 37.96% on conversational testimony speech, a large improvement over a romanized baseline. Critically, we identified that decomposed Unicode—commonly used in digital Yiddish text—fails in CTC-based ASR, which underscores the importance of normalization for both corpus creation and downstream applications. This finding extends beyond Yiddish to other languages that use combining diacritics for phonemic distinctions reflected in spelling.

Several important limitations should be noted. Our data preprocessing removed a large portion of the original transcribed speech segments through aggressive filtering for speaker overlap, disfluencies, borrowings, and code-switches. While this filtering was done to maximize the success of ASR training, it means our models are optimized for clean, single-speaker utterances rather than naturalistic testimony speech. Interviews with Holocaust survivors frequently contain overlapping speech between the survivor and the interviewer, false starts and hesitations, and language mixing; it is not yet known how an ASR system that includes such segments would perform in training or evaluation, and thus whether it would be suitable for production deployment in archival settings.

Additionally, we trained exclusively on Northeastern Yiddish, represented by 60 of the currently available 158 speakers in the CSYE. A natural extension of the current project would strive for inclusion of all three broad dialects of Eastern Yiddish: Northeastern ("Lithuanian"), Central ("Polish"), and Southeastern ("Ukrainian"). Improved dialect coverage would likely make this ASR system much more useful across archives of Yiddish-language testimonies, and other collections of Yiddish speech. One potential future direction would be to combine dialect identification with dialect-specific ASR: an initial classifier would first identify the speaker's regional variety and then route the audio to a specialized model trained on that dialect. This pipeline would provide a unified interface for a "whole language" ASR model for Yiddish.

Several other technical improvements could enhance performance. Integrating language models trained on text corpora—a standard approach in production systems—could substantially improve ASR performance for Yiddish. Even for the best phonemic models tested on clean audiobook data, 27.07% WER remains too high for production transcription without a significant amount of correction. Repositories such as the Yiddish Book Center's digital library (Yiddish Book Center, 2022) provide large-scale text data suitable for training language models aimed at correcting phonetically plausible but lexically invalid outputs. Additionally, data augmentation techniques such as noise injection could improve robustness to different datasets and recording conditions.

## 7. Acknowledgments

I gratefully acknowledge the USC Shoah Foundation – The Institute for Visual History and Education for its support of this project.

Thank you to Jacob J. Webber for sharing a tutorial that helped me debug my training code, and to Elan Rosenfeld for comments on the paper.

This material is based upon work supported by the National Science Foundation under Award No. BCS-2142797.

## 8. Data Availability and Ethical Considerations

The transcripts and audio recordings from the Corpus of Spoken Yiddish in Europe (CSYE) are available to the public online. Users of the corpus must abide by the USC Shoah Foundation Terms of Use as well as the CSYE Terms of Use, both of which are provided in the CSYE User Guide. The dataset compiled for the Reading Electronic Yiddish Documents (REYD) text-to-speech project, a subset of which was used for cross-domain testing, is also available online. See the Language Resource References section below for more details.

A phonemic ASR model is released at https://huggingface.co/ibleaman/w2v-bert-2.0-yiddish-northeastern for non-commercial research and educational purposes only. The model card also includes links to notebooks containing orthographic preprocessing functions and an interactive ASR demo.

| | |
|---|---|
| **Original utterance:** | זי הייסט: "די רוסישע רעוואָלוציע אין קייזערלעכן הויף" |
| **Translation:** | It [the new play] is called: “The Russian Revolution in the Imperial Court” |
| **ROM** | |
| *Reference:* | zi heyst di rusishe revolutsye in keyzerlekhn hoyf |
| *Prediction:* | zi heystdi rusisherevolutsiyein keyzerlekh un hoyf |
| **STD** | |
| *Reference:* | זי הייסט די רוסישע רעוואָלוציע אין קייזערלעכן הויף |
| *Prediction:* | ע א |
| **PHON** | |
| *Reference:* | זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ |
| **PHON (seed 42)** | |
| *Prediction:* | זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן האָויפֿ |
| **PHON (seed 43)** | |
| *Prediction:* | זִי הייסט דִי רוּסִישע רעוואָלוּצע יִן קייזערלעכן הויפֿ |
| **PHON (seed 44)** | |
| *Prediction:* | זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ |
| **PHON (seed 45)** | |
| *Prediction:* | זיי הייסטדִי רוּסִישע רעוואָלוּציע יִן קייזערלעכ ן הויפֿ |
| **PHON (seed 46)** | |
| *Prediction:* | זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ |

Table 6: Predictions for an example utterance from the REYD test set across all models and seeds. ROM primarily shows word boundary errors, STD produces unintelligible output, and PHON models achieve accurate transcription for seeds 44 and 46 with others showing minor errors.

The survivor testimonies used in this project were sourced from the USC Shoah Foundation VHA and licensed for inclusion in the CSYE. Due to the sensitive nature of testimony data, all CSYE transcripts are produced by hand with utmost care to ensure the texts are accurate and reliable. Users of the ASR model should be aware of its performance limitations and verify the accuracy of all generated transcripts against the original audio.

## 9. Bibliographical References

## References


Solomon A. Birnbaum. 2016. *Yiddish: A Survey and a Grammar*, 2nd edition. University of Toronto Press, Toronto. Originally published in 1979.

Isaac L. Bleaman. 2019. Guidelines for Yiddish in bibliographies: A supplement to YIVO transliteration. *In geveb*.

Isaac L. Bleaman. 2024. `yiddish`: A Python library for processing Yiddish text.

Aizik Dimantstein. 1996. Interview 20327. *USC Shoah Foundation Visual History Archive*. USC Shoah Foundation. Accessed April 1, 2026.

Neil G. Jacobs. 2005. *Yiddish: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Yoach Lacombe. 2024. Fine-tune w2v2-BERT for low-resource ASR with Transformers. Hugging Face, Community Blog & Articles.

Max Planck Institute for Psycholinguistics. 2021. ELAN [computer program].

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. arXiv.

Uriel Weinreich. 1963. Four riddles in bilingual dialectology. In *American Contributions to the Fifth International Congress of Slavists, Sofia, September 1963*, volume 1: *Linguistic Contributions*, pages 335–359. Mouton, The Hague.

Yiddish Book Center. 2022. Steven Spielberg digital Yiddish library.

YIVO (Yidisher visnshaftlekher institut). 1999. *Der eynheytlekher yidisher oysleyg [The Standardized Yiddish Orthography]*, 6th edition. YIVO Institute for Jewish Research and League for Yiddish, New York.

## 10. Language Resource References

Isaac L. Bleaman and Chaya R. Nove. 2025. The Corpus of Spoken Yiddish in Europe: Goals, methods, and applications. *Language Documentation & Conservation*, 19:142–157. Corpus available online: https://www.yiddishcorpus.org/csye.

Isaac L. Bleaman, Jacob J. Webber, and Samuel K. Lo. 2023. Speech synthesis in the “mother tongue”: Designing, training, and evaluating a text-to-speech system for Yiddish. *Journal of Jewish Languages*, 11(1):15–43. Dataset available at: https://github.com/REYD-TTS.

Jacob J. Webber, Samuel K. Lo, and Isaac L. Bleaman. 2022. REYD – The first Yiddish text-to-speech dataset and system. In *Proceedings of Interspeech 2022*. Dataset available at: https://github.com/REYD-TTS.

# From Consensus to Split Decisions: ABC-Stratified Sentiment in Holocaust Oral Histories

**Daban Q. Jaff**
Erfurt University, Erfurt, Germany
daban.hamad_ameen@uni-erfurt.de

**Abstract**

Polarity detection becomes substantially more challenging under domain shift, particularly in heterogeneous, long-form narratives with complex discourse structure, such as Holocaust oral histories. This paper presents a corpus-scale diagnostic study of off-the-shelf sentiment classifiers on long-form Holocaust oral histories, using three pretrained transformer-based polarity classifiers on a corpus of 107,305 utterances and 579,013 sentences. After assembling model outputs, we introduce an agreement-based stability taxonomy (ABC) to stratify inter-model output stability. We report pairwise percent agreement, Cohen's $\kappa$, Fleiss' $\kappa$, and row-normalized confusion matrices to localize systematic disagreement. As an auxiliary descriptive signal, a T5-based emotion classifier is applied to stratified samples from each agreement stratum to compare emotion distributions across strata. The combination of multi-model label triangulation and the ABC taxonomy provides a cautious, operational framework for characterizing where and how sentiment models diverge in sensitive historical narratives. Inter-model agreement is low to moderate overall and is driven primarily by boundary decisions around neutrality.

**Keywords**: Holocaust oral history, sentiment analysis, model disagreement, agreement, emotion

## 1. Introduction

Sentiment analysis (SA) focuses on identifying evaluative meanings, such as polarity or emotional states (Cambria et al., 2017). It is typically framed as a component of opinion mining, where the goal is to extract attitudes toward specific entities or events (Pang and Lee, 2008; Liu, 2012). Methodologically, SA has evolved from lexicon-based (Taboada et al., 2011) and rule-based (Hutto and Gilbert, 2014) to machine learning (Turney, 2002) and modern transformer architectures (Vaswani et al., 2017; Devlin et al., 2019). The contemporary models leverage deep contextual representations, though their effectiveness remains heavily dependent on the chosen unit of analysis, especially on whether it is a single sentence or an entire document (Pang and Lee, 2008; Liu, 2012).

A major obstacle for applying off-the-shelf SA systems to Holocaust oral histories is domain shift: models trained on different genres (e.g., product reviews or Twitter) face a changed input distribution when applied to long-form historical narratives, which can alter label propensities (Blitzer et al., 2007; Glorot et al., 2011; Pan and Yang, 2010). In Holocaust oral histories, evaluative meaning is often expressed indirectly (through description of the lived experience, stance taking, or moral framing), distributed across multiple sentences, and confounded by reported speech and the verbal reconstruction of experience over time. These characteristics can reduce the density of explicit sentiment cues and make polarity judgments less stable.

This paper studies the resulting phenomenon of model disagreement in Holocaust oral histories. We run three pretrained sentiment models and quantify how strongly they disagree at both sentence and utterance levels. The central methodological goal is neither to identify a single best model nor to estimate sentiment accuracy against human-annotated ground truth, but rather to harness the heterogeneous knowledge and inductive biases of these three systems simultaneously. None of the models were fine-tuned on trauma-related discourse. By treating each classifier as an independent knowledge source shaped by its training distribution, the pipeline is designed to expose genuine domain-shift behavior. Moreover, model confidences are not calibrated and are not directly comparable across architectures or training regimes; they are used here only as within-model heuristics and descriptive proxies.

## 2. Related Work

SA is known to degrade under domain shift and in long-form narrative settings with domain-specific linguistic phenomena. Early work shows substantial performance drops when sentiment models are transferred across domains (Blitzer et al., 2007), and transfer-learning surveys attribute this to distribution mismatch between source and target data (Pan and Yang, 2010). Domain variation in vocabulary and expression is therefore a core obstacle for opinion mining (Liu, 2012), motivating domain-adaptation approaches that jointly model diverse domains (Barnes et al., 2018).

Sentiment inference also depends on task definition and textual structure. Unit choice matters because document-level sentiment is not simply an average of sentence sentiments (Pang and Lee,



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 29–36
11 May 2026. 

2008), and aggregation interacts with how opinions are expressed across discourse (Liu, 2012; Kraus and Feuerriegel, 2019). Moreover, subjectivity and attribution can confound polarity when evaluations are embedded in reported or narrative speech (Wilson and Wiebe, 2005; Wiebe et al., 2005). To address these issues in practice, ensembling polarities across models is commonly used to combine complementary sentiment systems, including confidence averaging, stacking, and neural ensembles (Hagen et al., 2015; Troncy et al., 2017; Rouvier, 2017).

There is emerging computational work on sentiment, emotion, and text classification in oral-history interviews. Recent examples include neural sentiment analysis on Holocaust interviews (Blanke et al., 2020), emotion recognition in German oral histories (Gref et al., 2022), geographic emotion modeling of Holocaust testimonies (Ezeani et al., 2024), emotion annotation in the ACT UP Oral History Project (Pessanha et al., 2025), and LLM-based classification of Japanese-American incarceration narratives (Chen et al., 2024; Cherukuri et al., 2025).

Unlike prior works, this study focuses on systematic inter-model disagreement in Holocaust oral histories. To do so, we introduce an ABC agreement taxonomy (with an A split used for polarity-specific analyses) and analyze agreement/divergence across sentence- and utterance-level predictions.

## 3. Method: Triangulation and ABC Taxonomy

We employ three off-the-shelf pretrained transformer sentiment classifiers: SiEBERT (Hartmann et al., 2023), CardiffNLP Twitter-RoBERTa (Barbieri et al., 2020), and NLPTown (nlptown/bert-base-multilingual-uncased-sentiment). The models are deliberately selected to capture the complementary knowledge encoded in models trained under markedly different regimes: general web text, Twitter-style conversational language, and multilingual product reviews. Moreover, SiEBERT was included intentionally despite its binary label space because its forced polarity decisions make unanimous agreement more conservative and make disagreement with the Neutral-capable models analytically useful under domain shift.

Each utterance $u$ is segmented into sentences using NLTK's `sent_tokenize` (punkt-based). Only minimal normalization (whitespace cleanup) is applied prior to segmentation. Label harmonization follows the upstream pipelines: NLPTown's 1–5 star ratings are mapped to three-way polarity (1–2: Negative, 3: Neutral, 4–5: Positive) (confidence reflects certainty in the winning star rating).

| Level | Model | Neg. (%) | Neu. (%) | Pos. (%) |
|---|---|---|---|---|
| *Utterance* 107,305 | | | | |
| | NLPTown | 48.2 | 19.2 | 32.6 |
| | CardiffNLP | 11.3 | 80.4 | 8.3 |
| | SiEBERT | 54.2 | — | 45.8 |
| *Sentence* 579,013 | | | | |
| | NLPTown | 45.1 | 24.0 | 31.0 |
| | CardiffNLP | 21.2 | 69.3 | 9.4 |
| | SiEBERT | 53.9 | — | 46.1 |

Table 1: Polarity distributions across models and granularities.

For each sentence, each model outputs a label and an associated confidence score (the model's predicted probability for that label). For each model, we additionally compute an utterance-level aggregated label by assigning each polarity label $\ell$ the score

$$s(\ell) = \frac{n_\ell}{N} \, \mathrm{meanConf}(\ell),$$

where $n_\ell$ is the number of sentence-level outputs assigned label $\ell$ and $N$ is the total number of sentences in the utterance. We then select the label $\arg\max_\ell s(\ell)$. Scripts and analysis materials are publicly available.[1]

Table 1 summarizes the marginal polarity distributions produced by each model at sentence and utterance levels. These distributions reveal substantial label-propensity differences across models. This motivates the agreement diagnostics and ABC stratification introduced below.

### 3.1. Triangulation

After obtaining sentence-level labels and confidence scores from all three models, and deriving one aggregated utterance-level label per model, we perform cross-model triangulation at two granularities (Table 2), as follows.

**Sentence level.** A consensus label is obtained by majority vote across the three models. If at least two models agree, that label is selected. In the case of a true three-way split (one Negative, one Neutral, one Positive), the label with the highest model-reported confidence is selected.

**Utterance level.** For each model separately, we first use the utterance-level aggregation procedure defined above to obtain one aggregated polarity label from that model's sentence-level outputs. Cross-model triangulation at the utterance

[1] https://github.com/dabjaff/ABC-Stratified-Sentiment-in-Holocaust-Oral-Histories.

| Level | Metric | Neg. (%) | Neu. (%) | Pos. (%) |
|---|---|---|---|---|
| *Utterance* $107{,}305$ | | | | |
| | Count | 49,133 | 18,162 | 40,009 |
| | Percentage | 45.8% | 16.9% | 37.3% |
| *Sentence* $579{,}013$ | | | | |
| | Count | 260,727 | 113,237 | 205,049 |
| | Percentage | 45.0% | 19.6% | 35.4% |

Table 2: Triangulated polarity distribution across granularities.

level then applies majority vote over these three model-specific aggregated labels (equivalently represented as $-1, 0, +1$ for analysis); sentence-level labels and confidence scores are not consulted directly at this stage. In rare triangulation edge cases requiring deterministic tie resolution (95 sentences; 16 utterances), SiEBERT is used as a fallback to ensure reproducible label assignment, not as a claim of superior validity.

#### 3.1.1. Stability Stratification: ABC Taxonomy

While triangulation produces an ensemble label at both granularities, it does not by itself indicate label stability, i.e., the degree of inter-model agreement or disagreement associated with that label. We therefore introduce the ABC taxonomy as a diagnostic stratification framework that tags the outputs of cross-model triangulation by inter-model agreement stability. In this framework, each category represents a different level of consensus across the three-model ensemble, and each sentence and utterance is assigned to one of three agreement categories (see Table 3).

- **Category A (Full Agreement)**: All three models assign the exact same polarity, and the shared polarity is either Positive or Negative, because SiEBERT does not produce a Neutral class.
- **Category B (Partial Agreement)**: Exactly two models agree on the label. This includes (i) cases where the agreement involves Neutral (from CardiffNLP or NLPTown) and (ii) cases where at least two models agree on Positive or Negative.
- **Category C (Maximal Conflict)**: The three models produce three distinct labels (one Negative, one Neutral, one Positive).

Because SiEBERT cannot emit Neutral, Category A should be interpreted as a conservative unanimity subset for non-neutral polarity only, not as a general high-reliability subset over the full three-way label space.

| Level | Cat. | Count ($n$) | Share (%) |
|---|---|---|---|
| *U* ($N = 107{,}305$) | | | |
| | $A_{-1}$ | 8,786 | 8.2 |
| | $A_{+1}$ | 6,873 | 6.4 |
| | B | 73,037 | 68.1 |
| | C | 18,609 | 17.3 |
| *S* ($N = 579{,}013$) | | | |
| | $A_{-1}$ | 85,372 | 14.7 |
| | $A_{+1}$ | 39,119 | 6.8 |
| | B | 365,243 | 63.1 |
| | C | 89,279 | 15.4 |

Table 3: ABC taxonomy prevalence by granularity.

### 3.2. Kappa-based Agreement Diagnostics

To quantify agreement under model variation and complement the discrete agreement strata (A/B/C), we report standard inter-rater reliability diagnostics by treating the three sentiment models as raters and each sentence/utterance as an item (Table 4).

| Pair | Agr | $\kappa$ | $N_{\neq 0}$ | $\text{Agr}_{\neq 0}$ | $\kappa_{\neq 0}$ |
|---|---|---|---|---|---|
| *Utterances* ($N{=}107{,}305$) | | | | | |
| SiEBERT–CardiffNLP | 17.8 | 0.088 | 21,045 | 90.9 | 0.816 |
| SiEBERT–NLPTown | 61.7 | 0.350 | 86,746 | 76.3 | 0.518 |
| CardiffNLP–NLPTown | 32.3 | 0.114 | 18,649 | 88.5 | 0.767 |
| *Sentences* ($N{=}579{,}013$) | | | | | |
| SiEBERT–CardiffNLP | 28.0 | 0.144 | 177,588 | 91.2 | 0.801 |
| SiEBERT–NLPTown | 57.5 | 0.308 | 440,159 | 75.6 | 0.504 |
| CardiffNLP–NLPTown | 42.1 | 0.184 | 150,755 | 87.5 | 0.719 |

Table 4: Pairwise agreement (Agr.) and $\kappa$ for 3-way and polarity-only ($N_{\neq 0}$) subsets.

For each model pair, we compute percent agreement (**Agr.**) and Cohen's $\kappa$ on the shared three-way label space (Negative/Neutral/Positive). Because SiEBERT is binary while CardiffNLP and NLPTown are three-class, Neutral-inclusive agreement and confusion patterns are not directly comparable across all model pairs. We therefore interpret Neutral-boundary effects primarily in the pair where both models can emit Neutral (CardiffNLP and NLPTown). In addition, we exclude any unit labeled Neutral by either model in a given pair and recompute agreement and $\kappa_{\neq 0}$ over {Negative, Positive}.

Finally, we compute Fleiss' $\kappa$ (three raters) on the three-way space and on the polarity-only subset to summarize overall agreement, and we produce row-normalized confusion matrices for each classifier pair to localize which labels drive disagreement.

### 3.3. Auxiliary Emotion Profiling

As an auxiliary descriptive signal, we apply a T5-based emotion classifier (`mrm8488/t5-base-`

| Level | $N$ (3-way) | Fleiss' $\kappa$ | $N_{\neq 0}$ | Fleiss' $\kappa_{\neq 0}$ |
|---|---|---|---|---|
| Utterance | 107,305 | 0.0535 | 18,649 | 0.7835 |
| Sentence | 579,013 | 0.1287 | 150,755 | 0.7398 |

Table 5: Overall three-model agreement (Fleiss' $\kappa$) on the full three-way label space and on the polarity-only subset ($N_{\neq 0}$), where units labeled NEUTRAL by CARDIFFNLP or NLPTOWN are excluded.

`finetuned-emotion`; Raffel et al., 2020) to assess whether ABC strata exhibit coherent affective profiles. We use T5 as a descriptive probe because it predicts discrete emotions within a different model family/objective (text-to-text generation), reducing the risk of simply reproducing the same polarity decision boundary. Like the sentiment models, it remains out-of-domain for Holocaust testimony discourse, so its outputs are interpreted descriptively only.

Because Category A splits into two polarity-consistent subsets, we define four groups for affective triangulation: $A_{+1}$ (full tri-model agreement on POSITIVE), $A_{-1}$ (full tri-model agreement on NEGATIVE), and Categories B and C. We randomly sample $2{,}000$ utterances ($500$ per group) and $4{,}000$ sentences ($1{,}000$ per group), restricting inputs to 10–350 words. For each group, we compute (i) emotion-label distributions at sentence and utterance levels and (ii) mean confidence for the predicted emotion label (reported as a relative proxy, not a calibrated probability). We then compare these profiles across groups to assess whether agreement strata align with distinct affective signatures.

### 3.4. Data

The pipeline is applied to CORHOH (Jaff, 2025) (see Table 6), and only survivor utterances are analyzed (107,305 utterances, segmented into 579,013 sentences).

## 4. Results

### 4.1. Model-wise Polarity Distributions

Before turning to agreement metrics, Table 1 shows that the three models produce sharply different marginal polarity distributions across both granularities, indicating strong label-propensity mismatch under domain shift. In particular, CARDIFFNLP is strongly NEUTRAL-dominant, NLPTOWN is substantially more polar, and SIEBERT is strictly binary; these model-specific output profiles motivate the inter-model diagnostics reported next.

| Category | Attribute | Count | % |
|---|---|---|---|
| **Gender** | Female | 270 | 54.0 |
| | Male | 230 | 46.0 |
| **Birth cohort** | 1890s–1910s | 120 | 24.0 |
| | 1920s | 320 | 64.0 |
| | 1930s | 60 | 12.0 |
| **Top birthplaces** | Poland | 181 | 36.2 |
| | Germany | 115 | 23.0 |
| | Other (23 loc.) | 204 | 40.8 |
| **Migration era** | 1930s–1940s | 273 | 54.6 |
| | 1950s | 111 | 22.2 |
| | Other/Unknown | 116 | 23.2 |

Table 6: Corpus demographics and background variables (*N=500*).

### 4.2. Inter-model Classification

Inter-model classification is examined using pairwise $\kappa$-based diagnostics (Table 4), overall three-model agreement via Fleiss' $\kappa$ (Table 5), and row-normalized confusion matrices (Table 7).

**Sentence-level (N=579,013)**

| | Negative | Neutral | Positive |
|---|---|---|---|
| Negative | 35.1 | 63.3 | 1.7 |
| Neutral | 12.2 | 80.7 | 7.1 |
| Positive | 8.1 | 69.4 | 22.5 |

**Utterance-level (N=107,305)**

| | Negative | Neutral | Positive |
|---|---|---|---|
| Negative | 18.1 | 80.5 | 1.3 |
| Neutral | 6.5 | 88.3 | 5.2 |
| Positive | 4.1 | 75.5 | 20.4 |

Table 7: Row-normalized confusion matrices (rows: NLPTown, columns: CardiffNLP). Values are percentages (rounded).

Standard agreement statistics contextualize inter-model classification. On the full three-way label space (NEGATIVE/NEUTRAL/POSITIVE), pairwise percent agreement and Cohen's $\kappa$ are low to moderate (Table 4), and overall three-model agreement measured by Fleiss' $\kappa$ is low for both granularities (Table 5). When Neutral labels are included, Fleiss' $\kappa$ is 0.1287 at sentence level and 0.0535 at utterance level, confirming that inter-model disagreement is dominated by boundary decisions around neutrality rather than outright polarity reversal. When Neutral cases are excluded (polarity-only subset), Fleiss' $\kappa$ rises sharply to 0.7398 for sentences and 0.7835 for utterances (Table 5), indicating that the models align much more strongly once the task is reduced to Positive-vs.-Negative polarity.

To localize the disagreement mechanism identified above, we highlight the NLPTOWN–CARDIFFNLP row-normalized confusion matrix (NLPTown rows, CardiffNLP columns), because

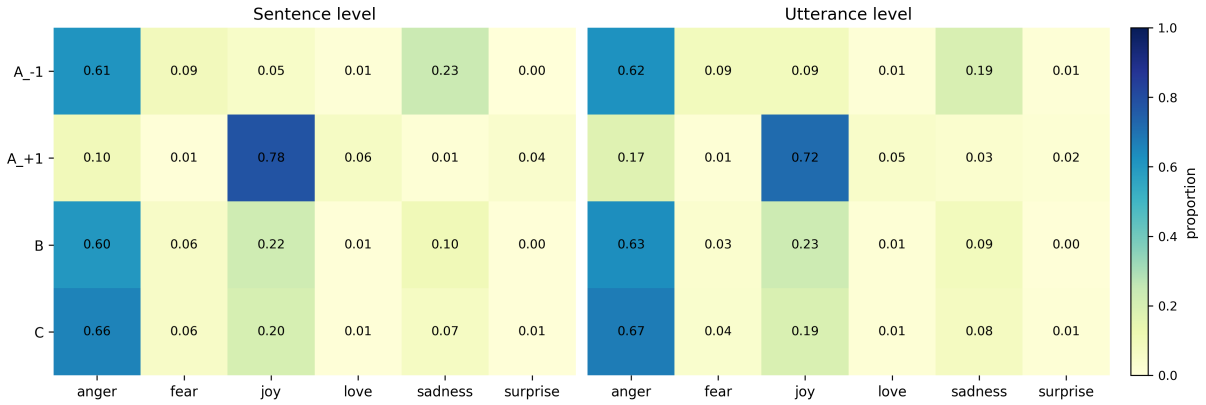


Figure 1: T5 emotion distributions (within-group percentages) across analysis groups ($A_{-1}$, $A_{+1}$, B, C) at sentence (left) and utterance (right) levels.

this pair directly exposes Neutral-boundary behavior between the two three-class models. When NLPTown predicts Positive, CardiffNLP predicts Neutral in 69.4% of sentences and 75.5% of utterances; likewise, NLPTown Negative maps to CardiffNLP Neutral in 63.3% (sentences) and 80.5% (utterances) (Table 7). Even when NLPTown predicts Neutral, CardiffNLP remains Neutral in 80.7% of sentences and 88.3% of utterances (Table 7). Conversely, NLPTown's polarity predictions frequently map to CardiffNLP's Neutral label, directly showing that disagreement is concentrated at the Neutral boundary rather than in systematic Positive/Negative reversal.

## 4.3. Stratification

Agreement patterns are further summarized using the ABC strata (Table 3). Unanimous polarity agreement (Category A) is more common for sentences than utterances, whereas Category B dominates at both granularities and Category C remains non-trivial, indicating persistent disagreement under aggregation. Full agreement ($A_{-1}$+$A_{+1}$) covers 21.5% of sentences but only 14.6% of utterances, while B remains the majority at both levels (Table 3). This makes the ABC taxonomy a practical agreement-based stability stratification for sentiment outputs in Holocaust oral histories: Category A isolates a conservative high-consensus subset suitable for polarity-stratified sampling when higher inter-model stability is desired. Accordingly, $A_{+1}$ and $A_{-1}$ can be used as conservative polarity-specific subsets for downstream analysis, while Categories B and C capture the dominant disagreement region. Because A is polarity-skewed ($A_{-1} > A_{+1}$), polarity-stratified sampling from A should preserve this split explicitly rather than treating A as a single homogeneous "high-agreement" set.

### 4.3.1. Descriptive Emotion Profiles

To profile the ABC agreement strata, we apply a T5-based emotion classifier at both sentence and utterance levels. T5 is also out-of-domain in this setting; it is used only descriptively (not as validation or ground truth) to assess whether the strata exhibit coherent external affective profiles (Figures 1–2).

### 4.3.2. Affective Distribution Patterns

The emotion heatmaps (Figure 1) show polarity-consistent affective profiles in the high-agreement strata. $A_{+1}$ is dominated by joy (78% sentence; 72% utterance), while $A_{-1}$ is dominated by anger (61–62%) with sadness as a substantial secondary emotion (19–23%). In contrast, B and C display more blended profiles (still anger-forward, with anger at 60–67%, but with larger secondary shares of joy at 19–23% and sadness at 7–10%), consistent with affective heterogeneity that may contribute to cross-model polarity disagreement.

### 4.3.3. T5 Confidence Proxy

The certainty heatmaps (Figure 2) partially mirror these patterns: $A_{+1}$ shows the highest certainty for joy (0.88 sentence; 0.83 utterance), while $A_{-1}$ shows relatively high certainty for anger (0.73 sentence; 0.74 utterance) and sadness (0.76 at both granularities). In B and C, certainty is generally less concentrated and varies more across labels, consistent with affective ambiguity in long-form oral-history discourse, although some sparse label cells show high confidence (e.g., B–surprise at the sentence level).

# 5. Conclusion

This study shows that sentiment-classifier disagreement in Holocaust oral histories is not merely a technical nuisance but an analytically meaningful

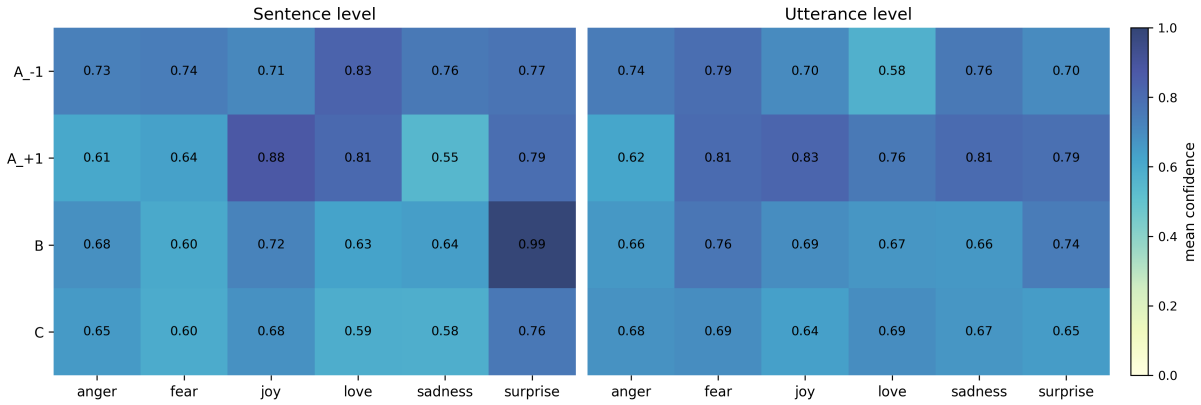


Figure 2: T5 mean confidence (uncalibrated certainty proxy) across analysis groups ($A_{-1}$, $A_{+1}$, B, C) at sentence (left) and utterance (right) levels.

signal of domain-shift sensitivity. Rather than converging on a single sentiment profile, off-the-shelf sentiment models produce different polarity distributions at both sentence and utterance levels, with disagreement concentrated especially around the NEUTRAL boundary. However, the present study is diagnostic rather than interpretive.

Triangulation and ABC provide an operational map of model behavior under domain shift: complemented by a T5-based descriptive affective probe, the framework identifies a conservative non-neutral consensus subset for downstream analysis and broader disagreement regions that can be flagged, filtered, or analyzed separately in future work.

Furthermore, these results provide a principled starting point for future work by indicating where future efforts could focus. The utterance-level aggregation rule is an operational heuristic combining within-model label frequency and confidence; alternative aggregation rules such as unweighted majority vote are left for future work. Future extensions may include domain-adaptive fine-tuning to improve polarity detection in Holocaust oral histories, introducing human-annotated evaluation subsets, and extending the ABC framework to other sensitive oral-history corpora.

## 6. Ethics Statement

This work analyzes publicly available Holocaust oral histories with respect for the survivors, their families, and the historical record. All analyses are strictly descriptive and exploratory. We do not claim that sentiment or emotion labels reflect ground-truth psychological states, nor do we present them as clinical or therapeutic interpretations. The models were used off-the-shelf (without fine-tuning on this corpus) to examine domain-shift behavior in a sensitive historical setting. Our goal is analytical: to identify where and why current NLP tools diverge on Holocaust oral histories.

## 7. Acknowledgements

I gratefully acknowledge the support of the ***Deutscher Akademischer Austauschdienst (DAAD)*** through a PhD research grant (Grant No. 57645448) for my doctoral studies at **Erfurt University** (Host: **Language and Its Structure**, Prof. Dr. Beate Hampe). I am also grateful to Beate Hampe for reading the manuscript and providing valuable comments. I thank the anonymous reviewers for their valuable comments. Last but certainly not least, this work was carried out using **Prince**, the computational resource of the **Language and Its Structure** professorship, for which I am grateful.

## 8. References

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tobias Blanke, Michael Bryant, and Mark Hedges. 2020. Understanding memories of the

Holocaust—a new approach to neural networks in the digital humanities. *Digital Scholarship in the Humanities*, 35(1):17–33.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Haihua Chen, Jeonghyun (Annie) Kim, Jiangping Chen, and Aisa Sakata. 2024. Demystifying oral history with natural language processing and data analytics: a case study of the Densho digital collection. *The Electronic Library*, 42(4):643–663.

Komala Subramanyam Cherukuri, Pranav Abishai Moses, Aisa Sakata, Jiangping Chen, and Haihua Chen. 2025. Large language models for oral history understanding with text classification and sentiment analysis. arXiv preprint arXiv:2508.06729.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ian N. Gregory, Tim Cole, Erik Steiner, and Zephyr Frank. 2024. The geography of 'fear', 'sadness', 'anger' and 'joy': Exploring the emotional landscapes in the Holocaust survivors' testimonies. In *Proceedings of the Seventh Workshop on Narrative Extraction From Texts (Text2Story 2024)*, volume 3671 of *CEUR Workshop Proceedings*, pages 93–103.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 513–520. Omnipress.

Michael Gref, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke, and Joachim Köhler. 2022. A study on the ambiguity in human annotation of German oral history interviews for perceived emotion recognition and sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2022–2031, Marseille, France. European Language Resources Association.

Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. 2015. Twitter sentiment detection via ensemble classification using averaged confidence scores. In *Advances in Information Retrieval (ECIR 2015)*, volume 9022 of *Lecture Notes in Computer Science*, pages 741–754. Springer, Cham.

Jochen Hartmann, Mark Heitmann, Christina Siebert, and Bram Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–97.

C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Daban Q. Jaff. 2025. Corhoh: Text corpus of holocaust oral histories. *Data in Brief*, 59:111426.

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:332–343.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Francisca Pessanha, Ian Padovani, Justus van Klaveren, Heysem Kaya, Almila Akdag, and Judith Masthoff. 2025. Listening to oral history: Emotion annotation and recognition in the ACT UP oral history project. In *SUMAC '25: Proceedings of the 7th International Workshop on analySis, Understanding and proMotion of heritAge Contents*, pages 41–50, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena,

Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mickael Rouvier. 2017. LIA at SemEval-2017 task 4: An ensemble of neural networks for sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 760–765, Vancouver, Canada. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Raphaël Troncy, Enrico Palumbo, Efstratios Sygkounas, and Giuseppe Rizzo. 2017. SentiME++ at SemEval-2017 task 4: Stacking state-of-the-art classifiers to enhance sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 648–652, Vancouver, Canada. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

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, pages 5998–6008.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan. Association for Computational Linguistics.

# EHRI Annotator: A Web-Based Tool for Named Entity Recognition and Linking in Holocaust-Related Texts

**Maria Dermentzi**[1,2]
[1]EHRI-CZ, Prague, Czechia
[2]Toolbox 21 Single Member PC, Kavala, Greece
maria@toolbox21.com

**Abstract**

This paper presents the EHRI Annotator, a web-based tool for multilingual named entity recognition (NER) and entity linking (EL) in Holocaust-related texts. The tool was developed to support services provided by the European Holocaust Research Infrastructure (EHRI), primarily the digital scholarly editions published by EHRI (EHRI Online Editions) by streamlining the process of detecting named entities in documents and linking them to their unique identifiers in EHRI and third-party controlled vocabularies and gazetteers. The EHRI Annotator builds upon previous work on domain-specific NER, taking it a step further to support multilingual EL. The tool adopts a dual entity linking architecture that uses a different matching approach depending on the type of the named entity. It performs semantic matching for entities to be linked to EHRI vocabularies and authority sets which are modestly sized, and string-matching-based retrieval for locations to be linked to the extensive GeoNames gazetteer using a domain-specific relevance weighting. A preliminary evaluation on 264 entities from a manually annotated dataset of Holocaust testimonies in three languages (English, German, Hungarian) yields an Accuracy@5 of 77.7% when it comes to the linking component of the tool. User testing confirms the tool's usability but also highlights areas for improvement.

**Keywords:** named entity recognition, entity linking, Holocaust studies, digital humanities, cultural heritage, multilingual NLP, digital editions

## 1. Introduction

A core service of the European Holocaust Research Infrastructure (EHRI)[1] is the publication of *EHRI Online Editions*[2], which are digital scholarly editions of thematically curated documents. Since 2018, EHRI has supported the publication of seven[3] EHRI Online Editions. The documents included in an EHRI Online Edition primarily include Holocaust testimonies, as well as diplomatic reports and correspondence hosted by different archival institutions around the world. To be included in an EHRI Online Edition, these documents are manually annotated with Extensible Markup Language (XML) according to the Text Encoding Initiative (TEI) P5 guidelines. Specifically, subject matter experts affiliated with EHRI partner institutions enrich the documents with semantic annotations to highlight and give context about the people, locations, organizations, and topics mentioned in them. Where applicable, these annotations also include links to unique identifiers in the EHRI vocabularies [4] and authority sets[5], as well as to GeoNames[6] according to the annotation guidelines created by EHRI[7]. This manual annotation process is extremely resource-intensive because it requires not only close reading of the documents by domain experts but also strong familiarity with large knowledge bases used for linking the named entities found in texts with their associated unique identifiers. At the same time, the result of this annotation process is very useful because it produces documents interlinked with common access points based on semantic similarities regardless of the source collection, enabling research on a specific topic using diverse and often transnational material.

This paper presents the EHRI Annotator[8], a web-based tool that was primarily built to streamline the process of named entity recognition (NER) and entity linking (EL) for Holocaust-related texts being prepared for publication as EHRI Online Editions[9]. The tool builds on previous work on domain-specific NER (Dermentzi and Scheithauer, 2024), which included fine-tuning a multilingual language model for Holocaust-related entity recognition (the EHRI-NER model[10]) using a dataset compiled from the EHRI Online Editions, making it a fit-for-purpose NER model because it has *"learned"* and *"inherited"* the annotation patterns and conventions followed

[1]EHRI project website. Accessed 2/25/2026.
[2]EHRI Online Editions webpage. Accessed 2/25/2026.
[3]At the time of writing on February 25th, 2026.
[4]EHRI Vocabularies. Accessed 2/25/2026.
[5]EHRI Authority Sets. Accessed 2/25/2026.
[6]GeoNames website. Accessed 2/25/2026.
[7]TEI encoding and annotation documentation page. Accessed 2/25/2026.
[8]EHRI Annotator website. Accessed 2/25/2026.
[9]The tool is publicly available as a beta service. The source code is not publicly released at this time.
[10]Available on Hugging Face.



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 37–46
11 May 2026. 

by EHRI Online Edition editors. In the EHRI Annotator, the EHRI-NER model is deployed as part of the NER component of the tool's pipeline, which is the first stage following the user's input of a text. The other major stage of the pipeline is the linking component which is triggered when the user decides that a match should be attempted for any of the entities detected by the NER component. For EL, EHRI Annotator employs a hybrid lexical and machine learning-based strategy to retrieve and suggest to the user the top five potential matches for the named entities that have been detected and deemed linkable by the user. Although the primary aim of the tool is to be a user-friendly platform that facilitates the annotation process when preparing new EHRI Online Editions, the tool can be easily modified and extended to support more use cases, such as metadata enrichment for enhancing the catalog of an archival institution or an aggregator like the EHRI Portal with more interoperable links according to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson et al., 2016), making the documents more findable and accessible and interlinking dispersed sources across institutions and languages. It can also be useful for enhancing a knowledge graph like the one described by García-González and Bryant (2023).

This paper's contributions include: a) the description of the EHRI Annotator pipeline; b) the hybrid architecture used for entity linking which employs a different strategy per entity type for more efficient candidate retrieval; c) preliminary evaluation results of entity linking accuracy on a manually annotated dataset of Holocaust testimonies; d) findings from user testing sessions. The NER model used in the EHRI Annotator is unchanged from Dermentzi and Scheithauer (2024).

## 2. Related Work

This work forms part of a broader effort to offer reliable named entity recognition and linking services in ways that support metadata enrichment of Holocaust-related archival material for the purposes of indexing and information retrieval but also for the contextualization of this material within an international landscape of dispersed, multilingual archival resources. Previous work (Dermentzi and Scheithauer, 2024) described why this is useful for EHRI and its services while also making the first step towards offering a multilingual NER model for the Holocaust domain. Having access to a reliable enough NER model was key to progressing towards finding an EL approach that is reasonably efficient and accurate for our use case. While in recent years there has been a lot of focus on NER and EL for historical documents and this research topic has been addressed as the target of shared tasks (Ehrmann et al., 2022b, 2020, 2022a), previous linking approaches typically target Wikidata as the knowledge base to link to, whereas domain-specific vocabularies like the ones used to link entities in the EHRI Online Editions are not prioritized as much. For a comprehensive survey of NER and entity disambiguation in historical documents, see Ehrmann et al. (2023).

The named entity linking approach followed in this paper was inspired by the work of Arora and Dell (2024) on the LinkTransformer package. LinkTransformer treats record linkage as a text retrieval task, where entities are encoded into dense vector representations using a transformer language model and then cosine similarity over these embeddings is being measured to retrieve the nearest neighbor in the target knowledge base. This approach supports multilingual matching without translation, as multilingual models such as LaBSE (Feng et al., 2022) map texts in different languages into a shared embedding space. Before developing the EHRI Annotator, the author experimented with the LinkTransformer package and found its approach to be an effective strategy for linking against the EHRI controlled vocabularies, which contain approximately 13,000 entries. However, the same was not true for linking against the GeoNames gazetteer, where the sheer volume of toponyms and aliases makes embedding-based retrieval impractical in terms of indexing cost and inference times. As described in Arora et al. (2024), the GeoNames gazetteer is so large and comprehensive that toponym disambiguation can be very accurate using string matching methods alone.

Following Arora et al.'s (2024; 2024) insights, for EHRI entities, the EHRI Annotator adopts the dense retrieval paradigm using LaBSE embeddings indexed in a vector database, while for geographic entities, we employ text-based retrieval with domain-specific relevance weighting. This dual-strategy design is explained by the differences among the target knowledge bases. The EHRI vocabularies and authority sets contain limited alias and translation coverage with entries appearing mostly in English under their preferred name, making semantic matching essential for cross-lingual recall. GeoNames, conversely, is rich in translations and alternative names across many languages, making string-based matching both sufficient and more scalable.

## 3. System Architecture

The EHRI Annotator (screenshot in Figure 1 below) consists of three components: a web-based user interface for document input and annotation verification and exporting; an NER processing backend that segments input texts and runs inference

using the EHRI-NER model via an asynchronous task queue; and an EL microservice powered by a vector database (Qdrant[11]). Having the NER and EL components decoupled, with the EL service operating as an independent application programming interface (API), was a deliberate decision to allow greater flexibility on how each component is maintained, scaled, and deployed, anticipating the need to eventually expand coverage and allow other EHRI services to query the EL API independently for metadata enrichment outside the EHRI Annotator.

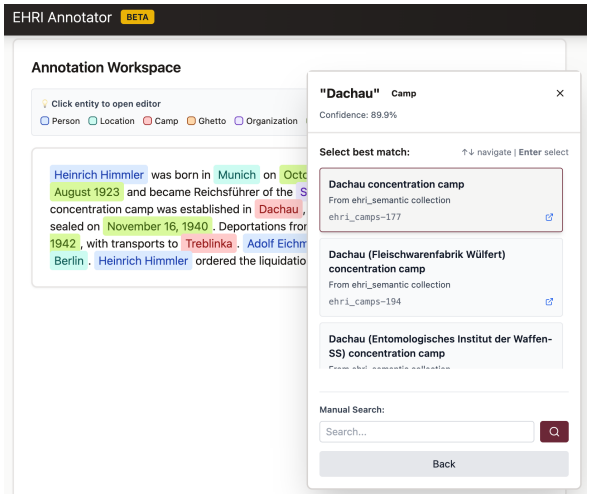


Figure 1: Screenshot of the entity linking process within the EHRI Annotator.

Upon entering the website, the user is prompted to input some text to a text box and click on a *"Process Text"* button which triggers the NER component of the application. As mentioned in the previous sections, for its NER component the EHRI Annotator currently relies on the EHRI-NER model, a detailed description of which can be found in Dermentzi and Scheithauer (2024). This model was developed by fine-tuning XLM-RoBERTa-Large for NER of six entity types, namely *PERSON*, *ORGANIZATION*, *LOCATION*, *CAMP*, *GHETTO*, and *DATE*. The NER backend splits the text into smaller chunks (if the text is too long for the model's context window to handle) and runs the model to return detected entities which are then displayed highlighted in the document. The user can then review each prediction to accept, reject or adjust its boundaries or label through an editing panel. For each entity detected (apart from *DATE* entities, where this is not applicable), the user can trigger the entity linking process by clicking on the *Link* button, which sends the query to the linking microservice and returns a ranked list of the top five candidates exceeding a certain threshold. The user can then select the correct match or perform a manual search against the index if no suitable candidates were returned. Any edits can be propagated to all mentions of the same entity with the same spelling within the same input text. Once all entities have been verified and linked, the user can export the annotated document as a TEI P5 XML file to process further with a text editor and prepare for publication. For linked entities with geographic coordinates (these could be *CAMP*, *GHETTO*, or *LOCATION* entities), linked places are additionally displayed on an interactive map.

Once the EL process of an entity is triggered by the user, the retrieval strategy depends on the entity's assigned label. This filtering strategy was inspired by what Arora and Dell (2024) describe as *"blocking"* in their paper and is used to make the tool as fast and efficient as possible. Therefore, *PERSON* entities get matched against the *EHRI Personalities* authority set, *ORGANIZATION* entities get matched against the *EHRI Corporate Bodies* authority set, *CAMP* entities against the *EHRI Camps* vocabulary, *GHETTO* entities against the *EHRI Ghettos* vocabulary and *LOCATION* entities against *GeoNames*.

Specifically, the linking component of the EHRI Annotator is supported by two Qdrant *"collections"*. The first collection concerns entities from the EHRI vocabularies and authority sets, which are indexed for semantic matching. These entities are encoded with LaBSE (Feng et al., 2022) into 768-dimensional vectors for approximate nearest neighbor search. Candidate retrieval returns the top 2,000 entries by cosine similarity, which are then re-scored using a function that combines semantic similarity, string similarity (a multi-stage scoring function from exact match through to fuzzy matching), context-aware scoring that incorporates the text of one neighboring entity to the left and one to the right of the target mention to aid disambiguation (e.g., the presence of the entity "Riga" in the context near the entity "Gestapo" helps rank "Gestapo Riga" higher in the candidate list compared to other Gestapo-related entities in the relevant EHRI authority set), and a multi-alias boost that rewards entities with multiple matching name variants. Exact string matches receive a much higher score to guarantee first-rank placement. Character-level fuzzy matching using the `rapidfuzz` library (Bachmann, 2025) with a similarity threshold of 0.7 is applied during candidate re-ranking for EHRI entities to handle minor spelling variations and Optical Character Recognition (OCR) errors common in archival texts. This multi-stage linking strategy was refined through extensive trial and error since there was no domain-specific dataset available for training a supervised ranking model at the time of building the EHRI Annotator. Semantic matching is

[11] GitHub repository for Qdrant vector search engine.

essential for matching against EHRI vocabularies, which have limited multilingual coverage. For example, matching *"malou pevnost"*, meaning small fortress in Czech, to the German equivalent *"Kleine Festung"* in the EHRI Camps dataset is only possible through semantic matching because the EHRI Camps vocabulary only lists the German name.

The second Qdrant collection concerns the GeoNames gazetteer, which contains over 13 million entries in total. Embedding every entry and its alias names would be prohibitively expensive in terms of storage and inference time. Instead, taking into account the conclusion reached by Arora et al. (2024) that toponym disambiguation against a comprehensive gazetteer can be highly accurate using non-neural methods, location entities are matched using text-based retrieval only. To enable this within Qdrant, which requires a vector for every entry, GeoNames entries are indexed with placeholder *"dummy"* zero vectors and rely exclusively on Qdrant's built-in text index of each entry's multilingual alias array for candidate retrieval. Moreover, we do not index the entire GeoNames dataset but rather filter it down to approximately 7.8 million entries based on a curated set of 102 GeoNames feature codes (e.g., populated places, historical sites, camps, railway stations, religious sites. The full list is provided in Appendix A.). Again, the selection of these feature codes was refined through iterative trial and error. When a good location match is not returned although it exists within the GeoNames dataset, the author examines whether a new feature code should be considered. Another concern was how to handle morphological variation in highly inflected languages. The solution to this was to lemmatize queries before matching using `Stanza` (Qi et al., 2020) with models trained on Universal Dependencies (Zeman et al., 2023) for Czech (cs), Polish (pl), Slovak (sk), German (de), Hungarian (hu), Russian (ru), Ukrainian (uk), Lithuanian (lt), Belarusian (be), Greek (el), and Hebrew (he). Candidate locations are ranked using a relevance score pre-computed during indexing that combines feature code importance weighted by domain relevance (e.g., camps and historical sites are weighted higher than generic buildings), country priority weights reflecting Holocaust and WWII geography (e.g., Poland, Germany, and Austria receive the highest weights), and a population factor. The final ranking for each candidate is determined by the product of this relevance weight and a text match quality score derived from comparing the query against the entry's primary name and aliases. The full feature code and country weight configuration as it currently stands is detailed in Appendix A.

The system architecture is designed to retrieve plausible candidates efficiently while leaving the final disambiguation decision up to the domain expert annotating the document. Verified annotations are exported as TEI P5 XML with `ref` attributes pointing to EHRI Portal Unique Resource Identifiers (URIs) or GeoNames URIs, pre-annotated for further processing as part of the Online Editions publication pipeline.

## 4. Evaluation

A preliminary evaluation of the EHRI Annotator has taken place both qualitatively through user testing sessions with Holocaust researchers who were asked to fill in feedback forms but also quantitatively through an automatic evaluation of entity linking accuracy against a gold-standard dataset which is still under active curation.

### 4.1. Dataset

The evaluation dataset was produced during the first EHRI–CLARIN Datathon, held on 26-27 February 2025 in Budapest, Hungary, co-organized by EHRI, CLARIN, ELTE University Research Center for Computational Social Science, and the Leibniz Institute for the History and Culture of Eastern Europe. During the event, participants annotated Holocaust testimonies using the INCEpTION platform (Klie et al., 2018). The source documents were provided by the Hungarian Jewish Museum and Archives and the Wiener Holocaust Library. *PERSON*, *ORGANIZATION*, *CAMP*, and *GHETTO* entities were linked to EHRI vocabularies and authority sets, while *LOCATION* entities were linked to Wikidata. At the time of writing, 40 (32 English, 7 German, 1 Hungarian) of the 140 documents processed during the event have been curated to ensure correct selection of URIs and consistent application of the annotation guidelines shared during the event. The full dataset and a detailed description of the annotation process will be part of a future publication once the curation process has been fully completed.

### 4.2. Entity Linking Evaluation

To evaluate entity linking, Wikidata identifiers for *LOCATION* entities were mapped to GeoNames via Wikidata property P1566. There were 30 location entities annotated with Wikidata identifiers that lacked a GeoNames mapping via property P1566. For non-location entities, EHRI vocabulary identifiers were compared directly. Entities annotated only with Wikidata identifiers for which no corresponding EHRI or GeoNames mapping exists were excluded. Entities were deduplicated by spelling, entity type, and document language, since the main evaluation (see Table 1 below) queries the linking service without document context and identical

queries produce identical candidate rankings. After deduplication and exclusions, 264 entities were retained for evaluation. Accuracy@1 (correct entity ranked first), Accuracy@5 (correct entity in top five), and Mean Reciprocal Rank (MRR) over the top 10 retrieved candidates are presented in Table 1 below. The results in Table 1 aggregate entities from all curated documents across all three languages.

Table 1: Entity linking performance by type.

| Type | N | Acc@1 | Acc@5 | MRR |
|---|---|---|---|---|
| LOCATION | 156 | 42.3% | 73.7% | 0.556 |
| PERSON | 30 | 76.7% | 100.0% | 0.883 |
| ORGANIZATION | 45 | 62.2% | 71.1% | 0.654 |
| CAMP | 23 | 78.3% | 87.0% | 0.811 |
| GHETTO | 10 | 80.0% | 80.0% | 0.800 |
| **Overall** | **264** | **54.2%** | **77.7%** | **0.641** |

The overall Accuracy@5 of 77.7% means that the correct knowledge base entry appears among the top five candidates in most cases. Since the EHRI Annotator is taking a human-in-the-loop approach in its design, this metric is useful because it shows that the system retrieves a good set of top five candidates for the domain expert to choose from. *LOCATION* entities show the lowest Accuracy@1 (42.3%) while Accuracy@5 remains reasonably high at 73.7% given that there are many locations with the same name which makes the ranking of the results harder. It is worth noting that the system returned no candidate matches for 23 out of 156 *LOCATION*-type entities under evaluation. These entities generally concern poor OCR or spelling mistakes (e.g., *"Heidelterg"* for Heidelberg, *"Theresienstad"* for Theresienstadt, *"Shtirotava"* for Škirotava). *ORGANIZATION* entities show the lowest Accuracy@5 (71.1%).

An ablation study was conducted on all ***non-location*** entities in the evaluation dataset (N=108) to estimate the contribution of the context-aware scoring described in Section 3. For each entity, two requests were sent to the linking service: one with context derived from one neighboring entity to the left and one neighboring entity to the right (matching the deployed service's behavior); and one request without context. Context-aware scoring improved Accuracy@1 from 71.3% to 74.1% and Accuracy@5 from 83.3% to 84.3%, with MRR increasing from 0.764 to 0.782.

However, this quantitative evaluation of the tool is preliminary while the gold standard dataset is under curation. In particular, results for *PERSON* (N=30), *CAMP* (N=23), and *GHETTO* (N=10) entities should be interpreted with caution given the small sample sizes. Nevertheless, this small-scale quantitative evaluation, taken into account together with the user feedback expressed under the qualitative evaluation described below, shows that this tool can already be useful in supporting Online Edition Editors to annotate new documents.

### 4.3. User Evaluation

The tool was evaluated by 11 users after two hands-on testing sessions, a workshop organized by EHRI-CZ in Prague and an EHRI webinar. Participants tested the tool on texts in English, German, Czech, Slovak, and Italian. They were then asked to complete a feedback form covering overall usability, quality of NER and EL predictions, shortcomings, and things to improve. All 11 participants rated the overall experience as either *Excellent* (8) or *Good* (3). Eight participants found the interface *Very easy* to use and three rated it *Mostly easy*; 10 of 11 described the layout as clean and easy to navigate.

NER accuracy was rated *Very accurate* by six participants and *Mostly accurate* by five. Participants noted several recurring NER errors, namely nationality adjectives misclassified as locations (e.g., *"British"*), incomplete entity boundaries requiring manual correction (e.g., the text reads *"of the Swedish Red Cross"* but the model detects *"Swedish"* as a separate *LOCATION* entity and *Red Cross* as its own *ORGANIZATION* entity. This is a problem because if we try to link *"Red Cross"* directly, we get presented with the wrong match. The correct entity here would be Swedish Red Cross as one *ORGANIZATION* entity.), and difficulty with retrieving matches for misspelled names (e.g., *"Krakoff"* for *"Kraków"*). One participant noted that morphological inflection in certain languages caused recognition failures.

Entity linking quality was rated *Very good* (correct match almost always ranked first) by eight participants and *Mostly good* (correct match usually in the first few candidates) by three, verifying qualitatively the results of the quantitative evaluation (albeit limited). When it came to more constructive feedback, participants requested expansion of the knowledge base to include vocabularies that cover additional historical entities such as the Slovak State or the Protectorate. Additionally, participants requested additional export formats including CSV, JSON, and spreadsheets. Editing features were generally well-received, though boundary editing was rated the most difficult task (*Easy*: 5, *OK*: 6), suggesting room for interface improvement. Six participants exported TEI XML; of these, four rated the export quality as *Excellent* and two as *Good - Needed minor tweaks*.

## 5. Discussion and Conclusion

The EHRI Annotator is a work in progress requiring continuous refinement. The evaluation presented here is preliminary, the dataset covers only 40 of the 140 documents from the datathon and sample sizes for several entity types are small. Nevertheless, several limitations of the current system emerged from both the quantitative and the qualitative evaluation. Disambiguating *LOCATION* entities is very challenging when dealing with such a large gazetteer like GeoNames. The system fails to retrieve candidates for non-standard spellings, while embedding this resource to enable semantic matching is prohibitive in terms of resources needed. The main limitation of the current EL approach when it comes to *LOCATION* entities is that the way Qdrant has been set up for the GeoNames collection (zero vectors and reliance on text-based retrieval) limits candidate retrieval to exact token matching, meaning that even small variations in spelling or OCR errors can lead to zero candidates even before any ranking can take place. While Qdrant remains a very practical solution in terms of keeping the infrastructure as simple and easy to maintain as possible given that we are already using it for the vector database, migrating to a search engine with rich full text search features such as Elasticsearch could help address these retrieval failures.

Another observation that can be made is that common errors of the EHRI-NER model that have been documented in previous work (Dermentzi and Scheithauer, 2024) can spill over to the linking stage as noted in the user evaluation reported here. It was also noted that the EHRI vocabularies themselves are not as comprehensive as users would like them to be, with unlinked entities often reflecting gaps in the authority sets rather than system failures. User feedback expressed the need to expand vocabulary coverage to include additional historical entities.

Given the sensitivity of Holocaust materials, all models used by the EHRI Annotator are self-hosted on a dedicated server in the European Union (EU), with no data transferred to third-party APIs. Original texts are discarded after NER inference and only entity-level data appears in system logs. This feature is essential for working with archival institutions which are often bound by strict data protection and ethical obligations regarding the materials in their custody. However, it also imposes practical constraints. Since galleries, libraries, archives, and museum (GLAM) institutions typically lack the infrastructure and resources to deploy and maintain computationally intensive models, this limits the adoption of state-of-the-art (SOTA) approaches that rely on large language models (LLMs) or cloud-based third-party API services. The architecture described in this paper was designed with these constraints in mind, favoring lightweight models that can be self-hosted over more powerful but externally controlled alternatives.

Current development priorities include support for adding new entities directly through the interface (currently possible only by editing the TEI XML output), linking to the EHRI Terms vocabulary which could be combined with previous work on automated subject indexing (Dermentzi et al., 2025), and additional import and export formats to support use cases beyond the EHRI Online Editions, such as archival metadata enrichment and geographic visualization. Longer-term goals include allowing users to connect custom vocabularies for domain-specific linking.

Another area of future work is experimenting with other NER and EL approaches. In Section 3, we described the three-component architecture of the EHRI Annotator. While the current system is based on the EHRI-NER model, the modularity of the service allows for experimentation with alternative models and techniques. Comparative evaluation against such approaches is planned for future work once the full evaluation dataset is available. For example, testing how open-source LLMs or other architectures like GLiNER (Zaratiana et al., 2024) perform against the EHRI-NER model can inform the choice of model for the NER and EL components before the tool moves from prototype to production. The fully curated dataset could also be used to train custom ranking models.

In conclusion, the EHRI Annotator demonstrates that a hybrid entity linking architecture which combines semantic retrieval for smaller vocabularies with text-based retrieval and domain-specific weighting for large gazetteers can provide effective candidate retrieval in a human-in-the-loop setting even when dealing with very large knowledge bases, as long as the larger knowledge bases are comprehensive enough in terms of alternative names and translations of the entries. The tool is currently deployed as a working prototype but it has already received positive reception. The hope is that it contributes to making Holocaust-related archival material more accessible, interoperable, and discoverable across institutions and languages.

## 6. Acknowledgements

The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

## 7. Bibliographical References

Abhishek Arora and Melissa Dell. 2024. LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 221–231, Bangkok, Thailand. Association for Computational Linguistics.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. 2024. Contrastive Entity Coreference and Disambiguation for Historical Texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186, Miami, Florida, USA. Association for Computational Linguistics.

Max Bachmann. 2025. rapidfuzz/rapidfuzz: Release 3.13.0.

Maria Dermentzi, Mike Bryant, Fabio Rovigo, and Herminio García-González. 2025. Multilingual Automated Subject Indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. *DH Benelux Journal*, 7(Breaking Silos, Connecting Data: Advancing Integration and Collaboration in Digital Humanities).

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

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

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

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022b. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. ArXiv:2007.01852.

Herminio García-González and Mike Bryant. 2023. The Holocaust Archival Material Knowledge Graph. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web – ISWC 2023*, volume 14266, pages 362–379. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Place: Santa Fe, USA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ArXiv:2003.07082 [cs].

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun

Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Francisco Costa, Marine Courtin, Mihaela Cristescu, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl,

Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Taksum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. Universal dependencies 2.12. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

## A. GeoNames Index Configuration

The GeoNames gazetteer (13.4 million entries) was filtered to 7.8 million entries by retaining only locations the feature code of which is included in the list below. Each indexed entry receives a precomputed relevance score combining a feature

code importance weight (Table 2), a country priority weight (Table 3), and a population factor.

Table 2: The 102 GeoNames feature codes retained for indexing, sorted alphabetically, with importance weights (W). Codes not listed are excluded from indexing.

| Code | W | Code | W | Code | W |
|---|---|---|---|---|---|
| ADM1 | 0.8 | FRST | 0.6 | PPLA5 | 0.6 |
| ADM1H | 0.8 | FT | 0.6 | PPLC | 0.9 |
| ADM2 | 0.7 | GRVE | 0.7 | PPLCH | 0.9 |
| ADM2H | 0.7 | HBR | 0.6 | PPLF | 0.6 |
| ADM3 | 0.6 | HSP | 0.6 | PPLH | 0.6 |
| ADM3H | 0.6 | HSPD | 0.5 | PPLL | 0.6 |
| ADM4 | 0.5 | HSTS | 0.8 | PPLQ | 0.7 |
| ADM4H | 0.5 | HTL | 0.5 | PPLS | 0.6 |
| ADM5 | 0.4 | INDS | 0.5 | PPLW | 0.7 |
| ADM5H | 0.4 | INSM | 0.6 | PPLX | 0.6 |
| ADMD | 0.3 | ISL | 0.8 | PRN | 0.7 |
| ADMDH | 0.3 | ISLS | 0.7 | QUAY | 0.6 |
| AIRB | 0.6 | LIBR | 0.6 | RGN | 0.9 |
| AIRQ | 0.5 | MFG | 0.5 | RGNE | 0.9 |
| BAY | 0.5 | MILB | 0.6 | RGNH | 0.9 |
| BAYS | 0.5 | MKT | 0.6 | RR | 0.5 |
| BDG | 0.5 | ML | 0.5 | RSTN | 0.6 |
| BNK | 0.6 | MN | 0.6 | RSTP | 0.6 |
| BRKS | 0.7 | MNMT | 0.7 | RVN | 0.8 |
| BTL | 0.7 | MSTY | 0.6 | SCH | 0.6 |
| CAVE | 0.6 | MUS | 0.6 | SCHC | 0.6 |
| CH | 0.6 | NVB | 0.6 | SEA | 0.7 |
| CMP | 0.9 | PCL | 1.0 | SQR | 0.7 |
| CMPLA | 0.9 | PCLD | 0.9 | STNB | 0.6 |
| CMPQ | 0.8 | PCLF | 0.9 | STNR | 0.6 |
| CMPRF | 0.7 | PCLH | 1.0 | STRT | 0.6 |
| CMTY | 0.7 | PCLI | 1.0 | SYG | 0.8 |
| CSTL | 0.6 | PCLS | 0.9 | THTR | 0.6 |
| CVNT | 0.6 | PIER | 0.6 | TMB | 0.7 |
| DCK | 0.6 | PPL | 0.6 | TMPL | 0.7 |
| DCKB | 0.6 | PPLA | 0.8 | TNL | 0.5 |
| DIP | 0.6 | PPLA2 | 0.7 | UNIV | 0.6 |
| EST | 0.5 | PPLA3 | 0.7 | WHRF | 0.6 |
| FRM | 0.5 | PPLA4 | 0.6 | WRCK | 0.5 |

Table 3: Country priority weights. All countries not listed receive a default weight of 0.3.

| Weight | Countries |
|---|---|
| 1.0 | DE, PL, CZ, SK, AT, HU, NL, BE, FR |
| 0.9 | GB, IT, RO, BG, HR, GR, IL, PS |
| 0.8 | RU, UA, BY, LT, LV, EE, LU, MT, YU, CS |
| 0.7 | NO, DK, SE, FI, CH, ES, PT, TR, RS, AL, SI, BA, CY, GI, US |
| 0.6 | CA, AU, NZ, ZA, BR, AR, CU, MA, DZ, TN, SY, LB, IQ, IE, ME, MK, MD, XK, HK |
| 0.5 | JP, CN, PH, ID, SG, MM, TH, VN, MY, KR, TW, LY, EG, IR, MX, UY, CL, BO, DO, IS, IN |
| 0.3 | All other countries (default) |

# The Shape of Testimony: A Scalable Framework for Oral History Archive Comparison

**Itamar Trainin, Renana Keydar, Amit Pinchevski**
Hebrew University of Jerusalem
{itamar.trainin, renana.keydar, amitpi}@mail.huji.ac.il

**Abstract**

Researchers in Holocaust studies have often distinguished between two styles of oral survivor testimony: the USC Shoah Foundation's interviews tend to follow a structured, interviewer-guided format, whereas the Yale Fortunoff Video Archive generally favors a more free-form, open-ended style. This distinction has influenced both scholarly research and the development of later archives. In this study, we critically examine that claim by conducting a large-scale computational analysis of more than 1,600 testimonies from both collections. Leveraging discourse segmentation, topic modeling, and large language model (LLM) based analysis, we quantify the "structuredness" level of testimonies through topic coherence, interviewer–survivor dynamics, and the distribution of question types. Our results generally corroborate the structural differences identified in earlier research, while also revealing significant overlaps between the collections, both within individual interviews and across common narrative patterns. This complicates the simple "structured vs. free-form" dichotomy often applied to these oral histories. Beyond revisiting a foundational claim in Holocaust studies, our work provides a scalable, replicable framework for comparative corpus analysis. As a proof of concept, it suggests broader applications for digital oral history, narrative analysis, and the design of citizen-science annotation platforms.

**Keywords:** Holocaust Studies, Oral History Language Resources, Computational Archive Comparison, Large language models (LLMs)

## 1. Introduction

A common claim in the field of Holocaust Studies suggests that the two main oral history video testimony archives, the Yale Fortunoff Video Archive for Holocaust Testimonies (Henceforth: Yale or Fortunoff) and the USC Shoah Foundation (Henceforth: USC or Shoah Foundation), present two opposite styles of interviews. It is said that the interview practices by Yale tend to be loose and open-ended, whereas USC follows a stricter format. Several studies have ascertained this difference by exploring the institutional agenda and policy as well as through close readings of selected testimonies (Wieviorka, 2006; Shenker, 2015; Pollin-Galay, 2018) and recently also by a small-scale empirical study (Presner et al., 2024).

Yet to empirically evaluate these assumptions, a large-scale comparative analysis is required, which is the purpose of this paper. This analysis calls for advanced computational tools, such as those developed recently in the digital humanities and computational history, for examining large narrative corpora. To date, applying these tools to oral testimonies has been limited due to the interpretive challenge and ethical complexities involved.

In what follows, we seek to ascertain the reported difference between the two testimony styles by means of a computational comparison of over 1600 Holocaust survivor testimonies drawn from the USC and Yale archives. Using large language models (LLMs), dialogical segmentation, topic modeling, and question classification, we develop a reproducible pipeline to measure "structuredness" across interviews. We examine topical coherence, interviewer-survivor dynamics, and the evolution of question types along interviews to assess whether and how these archives actually diverge.

Our findings confirm that USC testimonies tend to follow a more guided and structured format, especially in the early parts of the interview. Yale testimonies, by contrast, display greater topical fluidity, earlier emergence of core themes, and a higher proportion of open-ended questions. However, our analysis also uncovers significant convergences across both collections in later stages of the interviews, as well as similar narrative arcs, which seem to follow the chronology of Holocaust experience.

This article thus makes a dual contribution: (1) it offers a data-driven reassessment of a central historiographical claim within Holocaust testimony scholarship; (2) it introduces a scalable framework for computational comparison of large oral history corpora. By demonstrating how complex, ethically sensitive narratives can be analyzed computationally without compromising their integrity and interpretive richness, we open a pathway for further digital humanities work in trauma studies, oral history, and memory research.

## 2. Oral history archives of the Holocaust

The Yale Fortunoff Video Archive for Holocaust Testimonies and the USC Shoah Foundation Visual



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 47–58
11 May 2026. 

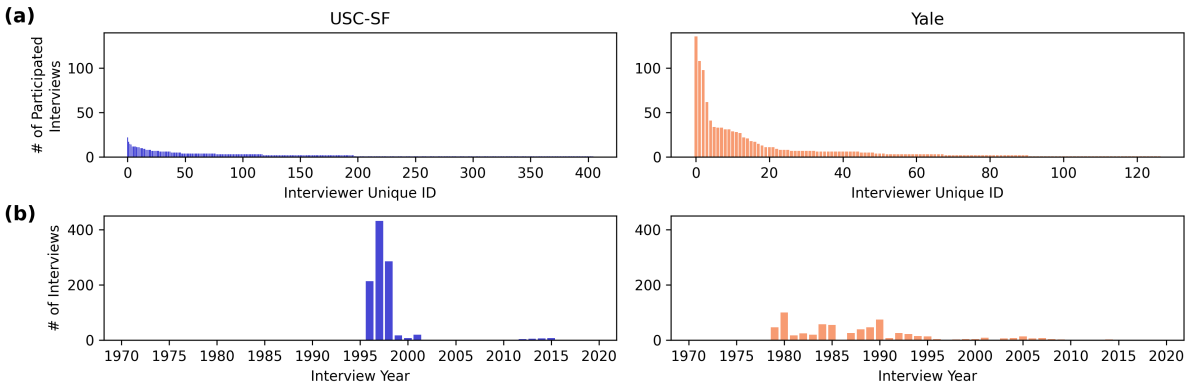


Figure 1: (a) Number of interviews conducted (individually or jointly) by each interviewer in each archive. (b) Annual distribution of testimonies recorded across both archives, computed over the subset of testimonies available for this study.

History Archive are the two main and largest collections of video testimonies of Holocaust survivors. The Fortunoff Archive began in 1979 as a local initiative in New Haven, Connecticut, later relocating to Yale University (Hartman, 1995). Holding more than 4,400 testimonies collected in the US, Europe, and Israel, its mission is to collect, preserve, and share the stories of those who were there. Being the first of its kind, the archive set the standard for projects to come (Felman and Laub, 1992a; Hartman, 1996). The USC archive was founded in 1994 by Steven Spielberg with the goal of recording 50,000 testimonies, for which hundreds of interviewers were trained in multiple countries (Smith, 2022a). It has since surpassed that goal, and in addition to Holocaust testimonies, has expanded to record victim narratives of other genocides (including Armenian, Rwandan, Cambodian) (Assmann, 2011; Smith, 2022a). Both archives were initially based on videotape technology and, over time, came to embrace digital technology, while incorporating and developing new methods to access and engage with testimonies (Assmann, 2011; Shandler, 2022).

At its core, an oral history interview is a process in which an interviewer and an interviewee spend extended time together engaged in storytelling and attentive listening. It is a collaborative act of constructing a narrative. Unlike other forms of qualitative interviewing (such as in-depth interviews), oral history interviews tend to be less narrowly focused in their topical organization (Hesse-Biber and Leavy, 2005). As with other oral history initiatives, each of the Holocaust archives has developed its own interview methodology.

From the outset, interviews recorded at Yale followed a clear code of conduct: putting the witnesses at the center and allowing them to relate their story freely with minimal guidance from interviewers. The interviewers' main role was to provide survivors with a supportive listening companion during the interviews, all of which took place on the Yale campus (Felman and Laub, 1992a). Conversely, USC interviews take place at the witness's home and are typically preceded by a preparation meeting. USC employs a large number of interviewers in many countries, who all go through formal training and follow a set of elaborate and formulated guidelines (USC Shoah Foundation, 2021) (Smith, 2022b)). The Yale archive has relied on a relatively small group of interviewers, mostly academics, some with psychological training. As will be shown below, despite these disparities, interviews recorded by both archives tend to unfold chronologically, beginning with life before the war, through early signs of change, antisemitism and persecution, experiences during the war (ghetto, captivity, forced labor, extermination), end of war, and finally life after the war. All interviews examined here were transcribed manually by their respective archive, together with speaker identification and speaker-side segmentation.

It is hard to imagine contemporary Holocaust history and memory without the prevalence of survivors' testimonies. The two archives have contributed significantly to the legitimacy of personal narratives as a source for the study of recent history, particularly of traumatic events (Friedländer, 1993; Hartman, 1996; Shandler, 2017). By their nature, testimonies are based on personal memory whose accuracy might vary, but what they nevertheless express is something beyond factual knowledge about past events. It is the relating of the personal experience of how it felt, which expresses the human aspect often missing from grand historical accounts (Felman and Laub, 1992b; Langer, 1993). As such, these narratives of pain and loss demand care and respect while retaining the singularity and integrity of each voice. Over the years, numerous audiovisual interviews with survivors have been recorded. The practical impossibility of watching tens of thousands of such testimonies presents an unprece-

dented challenge for a morally informed study of recent history. It is precisely here that carefully designed digital tools can enable new, previously unattainable modes of engagement with testimony (Keydar, 2020, 2022). Recent advances in natural language processing and large language models have opened new algorithmic ways of engaging with thousands of Holocaust testimonies at scale (Blanke et al., 2019; Naron and Toth, 2020; Ezeani et al., 2024; Presner et al., 2024; Shizgal et al., 2025; Keydar et al., 2026). However, to the best of our knowledge, there has so far been no attempt to harness these tools for large-scale comparative analysis across oral history archives.

## 3. Data and Methodology

This study analyzes 1,668 Holocaust survivor testimonies: 1,000 from the USC Shoah Foundation archive and 668 from the Yale Fortunoff Archive. Each interview was manually transcribed, annotated with speaker labels, and segmented into question–answer (Q/A) pairs, as outlined below. Due to the sensitive nature of these collections, transcripts are accessible only by direct request to the respective institutions. All testimonies in this study were conducted in English.

### 3.1. Date of testimonies

The two archives developed and followed highly different collection strategies. The Fortunoff Video Archive, established in 1979, has followed an open-ended acquisition policy, resulting in interviews collected over more than four decades. In contrast, the USC Shoah Foundation conducted a large-scale collection effort primarily between 1994 and 1999. However, within our available dataset, the testimonies from both archives are concentrated towards the initial years of operation. Fig. 1(b) shows the distribution of testimonies per year in the subset of testimonies available for this study.

### 3.2. Interview Length

USC interviews average 23,396 words[1] ($\sigma = 10{,}397$), while Fortunoff interviews average 13,622 words ($\sigma = 7{,}649$). Although both archives use similar methodologies, their protocols differ. One notable outcome is the substantial difference in length: USC interviews are, on average, 1.7 times longer than those from Fortunoff. This results in a greater quantity of content and potentially richer detail in the USC interviews. To mitigate the influence of this disparity, our quantitative metrics normalize for testimony length.

[1] A word defined as any whitespace-separated token.

### 3.3. Distribution of interviewers

| | USC | | Yale | |
|---|---|---|---|---|
| Top | Name | # | Name | # |
| 1 | Lorrie Fein | 22 | Vlock Laurel | 136 |
| 2 | Esther Finder | 17 | Laub Dori | 108 |
| 3 | Joanna Buchan | 15 | Kline Dana | 98 |
| 4 | Reuben Zylberszpic | 14 | Rudof Weiner | 62 |
| 5 | Joseph Huttler | 12 | Millen Susan | 41 |
| 6 | Dina Cohen | 12 | Frances Cohen | 34 |
| 7 | Ruth Meyer | 12 | Langer Lawrence | 33 |
| 8 | Zepporah Glass | 12 | Herz Moss | 33 |
| 9 | Florence Shuster | 11 | Strochlic Kathy | 31 |
| 10 | Yvonne Walter | 11 | Katz Helen | 31 |

Table 1: Top 10 participating interviewers and corresponding number of interviews conducted by the interviewer either as an individual or as part of a joint interview.

Fig. 1(a) shows the number of interviews each interviewer conducted in each archive. The Fortunoff corpus engaged a small cohort of interviewers, some of whom conducted over 100 interviews each, reflecting an approach that values a more personal interviewing style. In contrast, no individual interviewer in the USC corpus conducted more than 22 interviews (see top 10 participating interviewers in table 1). This distributed practice aligns with USC archive's emphasis on standardized interview protocols, which prioritize methodological consistency over personal interviewing style.

### 3.4. Methodology

Each testimony was segmented into chronological subunits using two complementary strategies, designed to capture both micro- and macro-level structures.

At the micro-level, we employed a topic-oriented segmentation: each Q/A pair was treated as a discrete topical unit, under the assumption (following prior work, (Ifergan et al., 2024)) that individual exchanges typically cover single topics. Short Q/A pairs were merged with adjacent exchanges when appropriate to maintain contextual coherence. This segmentation allowed us to examine the dialogic aspects of interviewer–survivor interaction while retaining the structure of the narrative.

For macro-level analysis, we divided each testimony into $k = 15$ equal-length chronological segments based on prior work ((Ifergan et al., 2024)) and manual validation, enabling a normalized temporal comparison across testimonies of varying lengths and numbers of Q/A pairs. From each segment, we extracted a range of features such as topical diversity, segmental coherence, answer and question lengths, and question type classification using custom pipelines and GPT-based classification.

Thus, in this study, we quantify the differences in the "structuredness" across collections, emphasizing the topical and interviewer–survivor dialogic dynamics. Nonetheless, additional dimensions that may confound a comparative analysis between Yale and USC interviews are discussed in §3.1 - §3.3.

# 4. Findings

## 4.1. Topical Sequence Analysis

A central dimension of "structuredness" in oral history interviews is the sequencing of topics across the testimonial timeline. Prior scholarship (e.g., (Piper et al., 2021; Wagner et al., 2022; Ranade et al., 2022; Ifergan et al., 2024; Wagner et al., 2025; Shizgal et al., 2025; Trainin and Abend, 2025)) suggests that highly structured interviews should exhibit a more predictable topical order, narrower chronological coverage per topic, and reduced thematic diversity across segments, whereas free-form interviews are expected to display looser sequencing and greater thematic fluidity.

To evaluate this hypothesis, we apply a computational topic-analysis pipeline modeled on, yet distinct from, earlier manual or semi-supervised approaches. We build upon the framework introduced in (Ifergan et al., 2024), but instead of employing topics generated by LDA (Blei et al., 2003) or BERTopic (Grootendorst, 2022) followed by expansive manual topic labeling and heuristic clustering thresholds, we introduce an alternative LLM-based strategy inspired by (Trainin and Abend, 2025). In doing so, we provide a scalable and fully automatic methodology for identifying emergent topical patterns across thousands of segments.

Our pipeline unfolds in two stages. First, an LLM (ChatGPT) generates a concise descriptive label for every Q/A pair, which serves as the micro-level unit of analysis (see Prompt 7.1). In the second stage (see Prompt 7.2), we prompt the LLM to identify the Top-K recurring topics for each chronological segment. This staged approach captures both the local dynamics of interviewer–survivor exchanges and the broader macro-level narrative structure. Fig. 7 in the appendix presents an illustration of this pipeline.

To validate our methodology, we calculated a Topic Coverage score (Trainin and Abend, 2025) for each inferred topic using a random sample of 50 testimonies. Additionally, we compared the USC topics produced by our pipeline to those derived in (Ifergan et al., 2024) and found strong structural convergence, indicating that the LLM-based approach is robust despite its limited interpretability.

Table 2 visualizes the top three topics per chronological segment. Recurring topics are manually color-coded across segments using a consistent palette, enabling readers to trace the persistence or transformation of themes over time. Coverage values for each topic are displayed alongside the labels.

Across both corpora, we observe a clear chronological trajectory: testimonies move from prewar life, to the onset of persecution, to deportation and camp experiences, and ultimately to liberation and postwar recovery. One salient structural difference, however, is the USC archive's recurring concluding segment dedicated to discussing personal photographs, labeled by our method as "Family Memories". This photo segment appears consistently in USC interviews but is largely absent from the Yale corpus, marking a distinct divergence in institutional interviewing protocols.

Despite the overall similarity of the macro-historical arc, the two corpora diverge in topical behavior and narrative dynamics. USC testimonies display a more predictable and segmented topical progression, with clearer boundaries between phases, an effect aligned with the archive's structured interview guidelines. Yale testimonies, in contrast, exhibit greater thematic fluidity: topics begin earlier, overlap more frequently, and persist across multiple segments, producing a narrative rhythm that is less tightly scaffolded by interviewer intervention.

Patterns of topical diversity likewise differ. In Yale interviews, the dominant topics in early segments tend to be introspective and affective, with broader diversification emerging only by the third topic. USC interviews, by contrast, show greater topic variation as early as the second and third topics, yielding a more multidimensional narrative distribution. These differences reflect not only interviewing style but also the interactional norms cultivated by each institution.

Taken together, the results complicate the binary distinction between "structured" and "free-form" interviews. Both archives follow a broadly linear historical logic, yet they diverge in how topics arise, persist, and shift across the testimonial timeline. The Yale corpus tends toward emotional continuity and reflective narration, while the USC corpus exhibits a more segmented and guided progression. The combination of LLM-based topic extraction, cross-method validation (LDA, BERTopic), and explicit visualization reveals that "structuredness" in Holocaust testimony is not merely a function of protocol but a dynamic interplay between institutional design, interviewer choices, and survivor narrative agency.

## 4.2. Question-Answer Dynamics

To measure question and answer dynamics, we performed a simple word count using the `NLTK` (Bird and Loper, 2004) library. We then plotted changes

| USC-SF | | | |
|---|---|---|---|
| **Seg.** | **Topic 1** | **Topic 2** | **Topic 3** |
| 1 | Childhood Memories (0.82) | Family Heritage (0.87) | Jewish Identity (0.67) |
| 2 | Childhood Memories (0.88) | Experiences of Antisemitism (0.48) | Family Dynamics (0.76) |
| 3 | Childhood Memories (0.86) | Experiences of Antisemitism (0.61) | Life in the Ghetto (0.36) |
| 4 | Life in the Ghetto (0.45) | Family Separation (0.45) | Survival Strategies (0.69) |
| 5 | Life in the Ghetto (0.48) | Family Separation (0.52) | Survival Strategies (0.74) |
| 6 | Life in the Ghetto (0.51) | Survival Strategies (0.76) | Family Separation (0.53) |
| 7 | Family Separation (0.51) | Survival in Hiding (0.56) | Life in the Ghetto (0.52) |
| 8 | Survival in Hiding (0.56) | Family Separation (0.51) | Experiences in Auschwitz (0.33) |
| 9 | Survival in Hiding (0.58) | Life in Concentration Camps (0.52) | Family Separation (0.47) |
| 10 | Survival in Hiding (0.56) | Family Separation (0.46) | Post-War Resilience (0.56) |
| 11 | Survival in Hiding (0.52) | Family Loss and Reunion (0.40) | Post-War Identity Struggles (0.54) |
| 12 | Survival and Resilience (0.88) | Family Loss and Reunion (0.46) | Post-War Identity Struggles (0.64) |
| 13 | Family Loss in the Holocaust (0.44) | Survival and Resilience (0.85) | Post-War Identity Struggles (0.70) |
| 14 | Family Memories (0.76) | Survival Reflections (0.79) | Holocaust Legacy (0.69) |
| 15 | Family Memories (0.82) | Legacy of Survival (0.78) | Holocaust Remembrance (0.63) |
| **Yale** | | | |
| **Seg.** | **Topic 1** | **Topic 2** | **Topic 3** |
| 1 | Childhood Memories Before the War (0.76) | Jewish Identity and Anti-Semitism (0.80) | Family Life Before the Holocaust (0.72) |
| 2 | Survival in the Ghetto (0.55) | Family Separation (0.50) | Experiences of Anti-Semitism (0.76) |
| 3 | Survival in the Ghetto (0.57) | Family Separation (0.53) | Hiding from Persecution (0.54) |
| 4 | Survival in the Ghetto (0.59) | Family Separation Trauma (0.60) | Experiences in Auschwitz (0.28) |
| 5 | Survival in the Ghetto (0.64) | Family Separation (0.49) | Experiences in Auschwitz (0.32) |
| 6 | Survival Strategies in Hiding (0.58) | Family Separation and Loss (0.61) | Experiences in Concentration Camps (0.60) |
| 7 | Survival in Hiding (0.54) | Family Separation (0.48) | Life in the Ghetto (0.50) |
| 8 | Survival in Hiding (0.53) | Family Separation (0.52) | Life in Concentration Camps (0.58) |
| 9 | Survival in Concentration Camps (0.60) | Life in the Ghetto (0.47) | Family Separation and Loss (0.55) |
| 10 | Survival in Hiding (0.56) | Family Loss and Trauma (0.68) | Post-War Resilience (0.57) |
| 11 | Survival in Hiding (0.56) | Family Loss in the Holocaust (0.42) | Experiences in Concentration Camps (0.67) |
| 12 | Survival Strategies (0.77) | Family Loss (0.43) | Post-War Resilience (0.64) |
| 13 | Survival in Concentration Camps (0.53) | Family Loss and Trauma (0.71) | Post-War Resilience (0.65) |
| 14 | Survival and Trauma (0.89) | Family Loss and Memory (0.68) | Identity After Liberation (0.73) |
| 15 | Survival and Memory (0.74) | Family Loss and Resilience (0.55) | Holocaust Education and Remembrance (0.53) |

Table 2: Top three topics per chronological segment. Recurring topics are manually color-coded across segments using a consistent palette showing the transformation of themes over time. Coverage values for each topic are displayed alongside the labels.

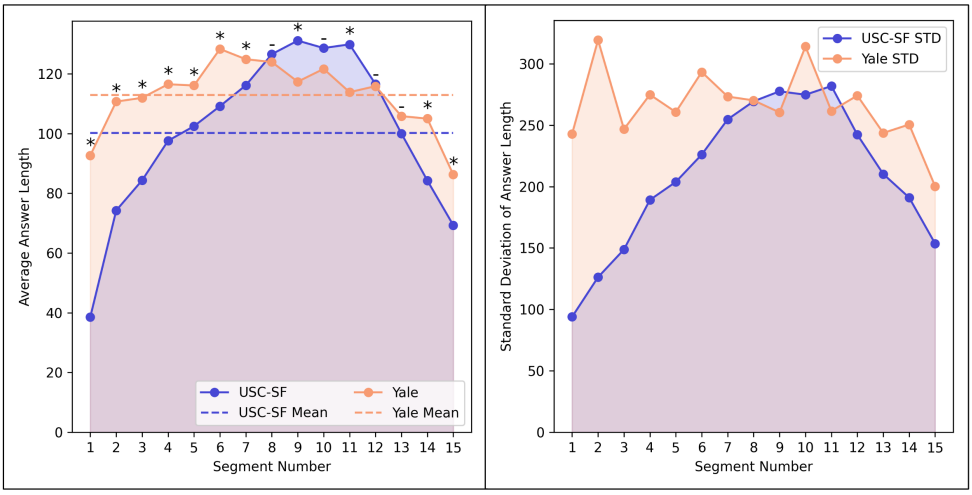


Figure 2: Mean (left) and SD (right) of answer length over time for USC and Yale archives. Statistically significant ($t$-test) differences are marked with an asterisk (*).

in average word count across testimony segments for each corpus, reporting both the mean and variance. A two-sample $t$-test was conducted for each segment to assess whether differences between the corpora were statistically significant (indicated by an asterisk). The overall mean length was plotted as a dashed reference line. See Figures 2, 3.

**Answer Length**: Yale survivors tended to provide longer answers in the early segments of their testimonies, whereas responses in the USC corpus gradually lengthened over time. Approximately between segments 1–8, Yale testimonies exhibited longer responses on average. Around segments 8–11, this pattern reversed, with USC answers becoming longer. Toward the end of the testimonies, both collections showed a relative decline in answer length, although Yale participants generally maintained slightly longer responses.

Examining the standard deviation (SD) of an-

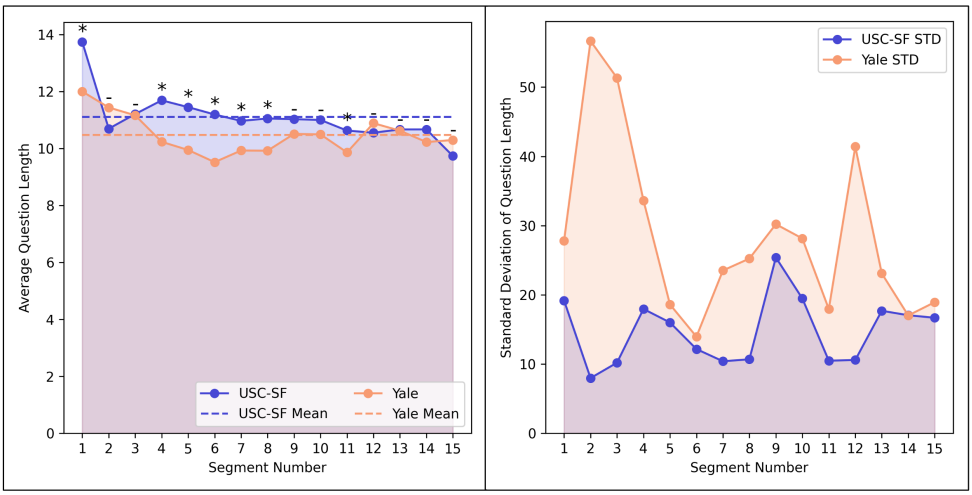


Figure 3: Mean (left) and SD (right) of question length over time for USC and Yale archives. Statistically significant ($t$-test) differences are marked with an asterisk (*).

swer length reveals the degree of variability across testimonies. Higher SD values indicate greater fluctuation in response length, which may reflect differences in survivors' narrative styles, interviewer intervention, or emotional pacing. In this context, Yale testimonies exhibit higher variability, suggesting that survivors alternated between extended reflections and shorter responses, while USC testimonies show more consistent pacing and structure.

**Question Length and Frequency**: USC interviewers tended to ask longer and more frequent questions, particularly during the early portions of the testimonies. This pattern aligns with the more guided and segmented structure of USC interviews. In contrast, Yale interviewers posed shorter and less predictable questions. The high variability in question length within the Yale corpus (as reflected in SD values) is noteworthy, given the relatively small pool of interviewers, suggesting considerable stylistic diversity and less standardized interviewing practice.

**Convergence**: Both archives displayed a reduction in answer length toward the final segments, a trend likely associated with standardized closing routines or reflective conclusions. Statistical testing confirmed that differences in answer length between the two corpora were significant in the early interview stages but diminished progressively toward the end.

Taken together, the analysis of question and answer length reveals distinct narrative and methodological dynamics within the two archives. The longer early responses in the Yale corpus suggest a more open and reflective interview style, one that allows survivors to elaborate freely and set the narrative pace. In contrast, the USC interviews begin with shorter, more structured exchanges but expand over time, indicating an interviewer-driven scaffolding that gradually gives way to survivor-led narration. The higher variability in Yale question and answer lengths reflects more dialogical spontaneity, where interviewers adapt responsively to the survivor's storytelling, whereas the more uniform USC patterns demonstrate adherence to a formalized interview guide. The convergence observed in later segments, marked by shorter responses and reduced differences between the corpora, points to the influence of shared testimonial conventions and closing rituals that shape the end of interviews across institutions. Overall, these patterns highlight that interview structure is not static but evolves through the testimony, balancing institutional design with survivor agency and emotional rhythm.

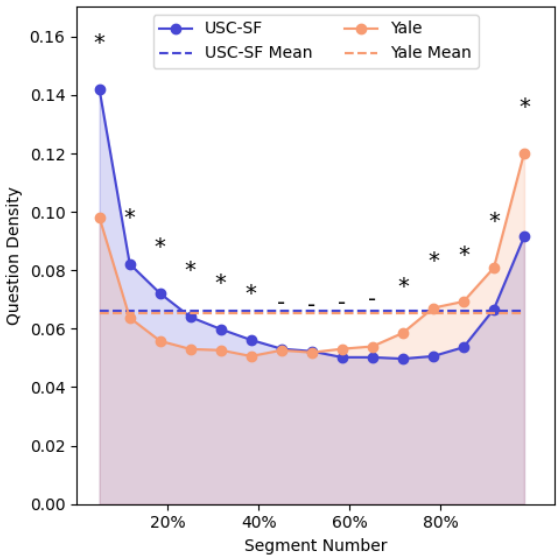


Figure 4: Intervention Density for USC and Yale archives. Statistically significant ($t$-test) differences are marked with an asterisk (*).

### 4.3. Intervention Density

To assess the density of interviewer interventions, we measured the duration of uninterrupted survivor speech. Lacking timestamp data, we used word count as a proxy for duration by calculating the proportion of uninterrupted survivor words relative to the total testimony length.

To examine how this measure changes over time, we divided each interview into 15 equal segments based on the cumulative word count. This segmentation, therefore, represents proportional divisions of the testimony length rather than the micro-level segmentation used in the Q/A-based analysis. For each corpus, we computed the average intervention density per segment across all testimonies. This word-based segmentation allows us to capture the relative distribution of interviewer interventions throughout the testimonies, independent of total interview length, and provides a comparable temporal framework across the two archives. The results, presented in Fig. 4, display the average uninterrupted speech density for each segment.

The analysis of intervention density provides further insight into the interactional dynamics that differentiate the two archives. Longer uninterrupted segments of survivor speech indicate greater narrative autonomy, whereas higher intervention density suggests tighter interviewer control or increased dialogical guidance. In this context, the Yale testimonies tend to feature longer uninterrupted stretches, reflecting a more survivor-centered approach that privileges emotional flow and narrative continuity. The USC interviews, by contrast, exhibit a denser pattern of interventions, especially in the earlier portions of the testimonies, which is consistent with a structured interview format emphasizing chronological order. However, in both collections, there is a gradual reduction in intervention density as testimonies progress, suggesting perhaps increased confidence on the survivors' part or greater flexibility on the interviewers' part. This suggests that narrative authority is negotiated throughout the testimonial encounter, resulting in an evolving dynamic between institutional control and individual agency in the making of oral history.

### 4.4. Question Types

Building on a framework proposed in ((Mittelstadt et al., 2016), Fig. 3.6), we hypothesize that interviewing style affects the distribution of question types–specifically "why," "what," "where," "who," "when," "how," and "other". Structured interviews are expected to rely more heavily on factual and directive questions (e.g., "what," "when," "who"), whereas less structured or conversational interviews are likely to include a broader range of open-ended or ambiguous questions (classified as "other").

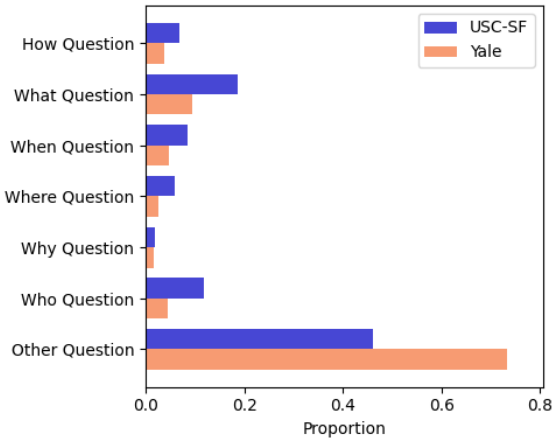


Figure 5: Distribution of question types for USC and Yale archives.

To test this hypothesis, we extracted all interviewer questions from both testimony corpora and employed an LLM to classify each into one of the seven categories (see Prompt 7.3 in the Appendix). Model predictions were validated through manual review of a random subset of 50 examples per category (3,500 questions in total). The overall and segment-based distributions of question types are presented in Figures 5 and 6.

The LLM-based classification revealed clear stylistic differences between the two archives. Yale interviewers used a higher proportion of "Other" questions, exhibiting a more flexible and exploratory interviewing style that allows survivors to guide the conversation. In contrast, USC interviewers relied predominantly on "what," "who," and "when" questions, particularly in the early segments, consistent with a more structured and documentary-oriented approach. That said, USC interviews show a gradual increase in "Other" question types over time, which might indicate that as rapport developed, the exchanges became more dialogic and open-ended.

Overall, the distribution of question types offers insight into institutional interviewing practices and the dynamics of testimonial co-construction. The USC format prioritizes a more formal interview structure, limiting opportunities for narrative divergence or emotional elaboration. The Yale format fosters greater conversational variability and affective depth, which allows far greater survivor agency in shaping the narrative, but for this reason is more dependent on survivors' compliance and initiative. These findings illustrate how question type mediates the balance between institutional protocol and personal voice, revealing how methodological design and interpersonal rapport together shape the structure and tone of Holocaust testimony.

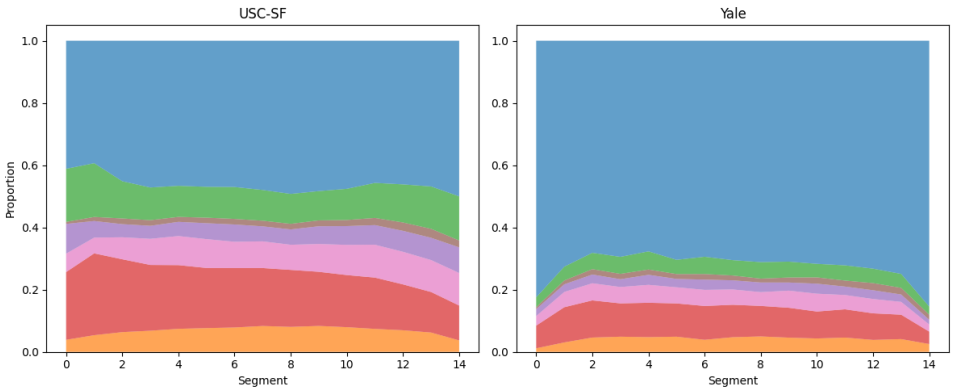


Figure 6: Distribution of question types over time for USC and Yale archives.

## 5. Toward a Scalable Framework for Oral History Comparison

Our analysis confirms that USC interviews are, overall, more structured than those in the Yale archive, particularly in question types, topical flow, and interviewer interventions. The USC corpus displays clearer segmental boundaries and a more predictable topical progression, whereas Yale testimonies reveal greater thematic fluidity, variability in questioning style, and more open-ended conversational turns. As the interview unfolds, however, these contrasts diminish as both interview styles move toward similar narrative rhythms shaped by survivor agency, historical chronology, and the affective dynamics of testimony. Ultimately, despite institutional differences, what becomes apparent is that in both collections the dialogical exchange between interviewer and survivor meaningfully shapes the unfolding of testimony alongside the structural template. This finding underscores that "structuredness" in oral history is not an institutional constant but a dynamic property of interaction, emerging from the interplay between methodological design, interpersonal trust, and the survivor's narrative agency.

Beyond these findings, the study contributes a concrete methodological framework for comparative oral history analysis. The proposed pipeline introduces a sequence of computational strategies designed specifically for dialogic and large-scale testimonial data.

First, it develops a segmentation method that extracts coherent question-answer trajectories from long interviews, enabling alignment across archives that differ in format and length. This segmentation serves as the foundation for cross-corpus comparison, allowing structural and thematic patterns to be analyzed at a shared level of granularity.

Second, the study advances a novel approach to LLM-based topic extraction. Working with oral history presents unique challenges, including contextual overflow, diffuse topic boundaries, and highly personal narrative structures. To address these, the pipeline employs a staged prompting strategy inspired by map-reduce logic: local prompts generate micro-level topics for each segment, and a secondary synthesis step aggregates these into common themes across the corpus. This approach mitigates the limitations of unsupervised topic modeling and allows for interpretable, reproducible results grounded in qualitative meaning.

Third, the framework integrates iterative evaluation and refinement. Through experimentation with classification strategies, segmentation thresholds, and prompt design, the study develops an ontology of best practices for LLM-assisted oral history analysis. The resulting workflow balances automation and interpretation, enabling scalable comparison while preserving the discursive and emotional complexity of survivor narratives.

Taken together, these contributions establish a reproducible and extensible model for comparing oral history archives at scale. Rather than treating computational methods as an abstraction from humanistic analysis, the pipeline operationalizes interpretive categories such as "structuredness, topical coherence, and interviewer style into measurable and comparable features. It allows distinct institutional collections to be analyzed within a shared framework, revealing how methodological design, interview dynamics, and survivor agency jointly shape the production of testimony.

More broadly, this study demonstrates how digital humanities can move from isolated case studies toward comparative infrastructures that connect archives and traditions of memory work. The pipeline developed here enables not only new forms of analysis but also new questions about the ethics, scale, and dialogical nature of historical testimony. It shows that computational methods, when used critically, can deepen rather than diminish the interpretive work of the humanities, thus illuminating the ways in which stories are told, recorded, and remembered across time and institutional boundaries.

## 6. Ethics Statement and Limitations

Throughout the study, the OpenAI-GPT suite of models was used, with manual validation of segment samples conducted to confirm model consistency. Furthermore, all data access was obtained with institutional permission and handled strictly in accordance with ethical standards for trauma data. Future work will involve integrating this pipeline into an open-access annotation platform for cross-disciplinary use.

## References


Aleida Assmann. 2011. *Cultural memory and Western civilization: Functions, media, archives*. Cambridge University Press.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. 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. *Digital Scholarship in the Humanities*, 35(1):17–33.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ifeanyi Ezeani, Paul Rayson, Ian N. Gregory, Timothy Cole, Emily Steiner, and Zachary Frank. 2024. The geography of 'fear', 'sadness', 'anger' and 'joy': Exploring the emotional landscapes in the holocaust survivors' testimonies. In *Proceedings of Text2Story@ECIR*, pages 93–103.

Shoshana Felman and Dori Laub. 1992a. *Testimony: Crises of witnessing in literature, psychoanalysis, and history*. Taylor & Francis.

Shoshana Felman and Dori Laub. 1992b. *Testimony: Crises of Witnessing in Literature, Psychoanalysis and History*. Routledge, London / New York.

Saul Friedländer. 1993. *Memory, History, and the Extermination of the Jews of Europe*. Indiana University Press, Bloomington.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Geoffrey H Hartman. 1995. Learning from survivors: The yale testimony project. *Holocaust and Genocide Studies*, 9(2):192–207.

Geoffrey H Hartman. 1996. *The longest shadow: In the aftermath of the Holocaust*. Indiana University Press.

Sharlene Nagy Hesse-Biber and Patricia Leavy. 2005. *The Practice of Qualitative Research*. SAGE Publications, Thousand Oaks, CA. Also published in 2006 edition.

Maxim Ifergan, Omri Abend, Renana Keydar, and Amit Pinchevski. 2024. Identifying narrative patterns and outliers in holocaust testimonies using topic modeling. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024*, pages 44–52.

Renana Keydar. 2020. Listening from afar: An algorithmic analysis of testimonies from the international criminal courts. *University of Illinois Journal of Law, Technology & Policy*, 2020(1):55–83.

Renana Keydar. 2022. Changing the lens on survivor testimony: Topic modeling the eichmann trial. *Jewish Studies Quarterly*, 29(4):412–435.

Renana Keydar, Amit Pinchevski, Maxim Ifergan, and Omri Abend. 2026. The testimony of the multitude: Toward a computational model of listening to holocaust testimony. *Holocaust and Genocide Studies*, 40(2).

Lawrence L Langer. 1993. *Holocaust testimonies: The ruins of memory*. Yale University Press.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Stephen Naron and Gabor Mihaly Toth. 2020. *Let Them Speak: An Effort to Reconnect Communities of Survivors in a Digital Archive*, pages 71–94. Springer International Publishing, Cham.

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.

Hannah Pollin-Galay. 2018. *Ecologies of Witnessing: Language, Place, and Holocaust Testimony*. Yale University Press, New Haven, CT.

Todd Presner, Anna Bonazzi, Rachel Deblinger, Lizhou Fan, Michelle Lee, Kyle Rosen, and Campbell Yamane. 2024. *Ethics of the Algorithm: Digital Humanities and Holocaust Memory*. Princeton University Press, Princeton, NJ.

Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. 2022. Computational understanding of narratives: A survey. *IEEE Access*, 10:101575–101594.

Jeffrey Shandler. 2017. *Holocaust Memory in the Digital Age: Survivors' Stories and New Media Practices*. Stanford Studies in Jewish History and Culture. Stanford University Press, Stanford, CA.

Jeffrey Shandler. 2022. Digitizing holocaust memories. *Jewish studies in the digital age*, 5:25.

Noah Shenker. 2015. *Reframing Holocaust Testimony*. The Modern Jewish Experience. Indiana University Press, Bloomington, IN.

Esther Shizgal, Eitan Wagner, Renana Keydar, and Omri Abend. 2025. Computational analysis of character development in holocaust testimonies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22721–22745.

Stephen D Smith. 2022a. *The trajectory of Holocaust memory: The crisis of testimony in theory and practice*. Routledge.

Stephen D. Smith. 2022b. *The Trajectory of Holocaust Memory: The Crisis of Testimony in Theory and Practice*, 1 edition. Routledge, London.

Itamar Trainin and Omri Abend. 2025. T5score: A methodology for automatically assessing the quality of llm generated multi-document topic sets. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26347–26375.

USC Shoah Foundation. 2021. Interviewer guidelines. https://sfi.usc.edu/sites/default/files/docfiles/uscsf_interviewer_guidelines_03_2021.pdf. PDF document, accessed December 14, 2025.

Eitan Wagner, Renana Keydar, and Omri Abend. 2025. Unsupervised location mapping for narrative corpora. *arXiv preprint arXiv:2504.05954*.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. *arXiv preprint arXiv:2210.13783*.

Annette Wieviorka. 2006. *The Era of the Witness*. Cornell University Press, Ithaca, NY.

# Appendix

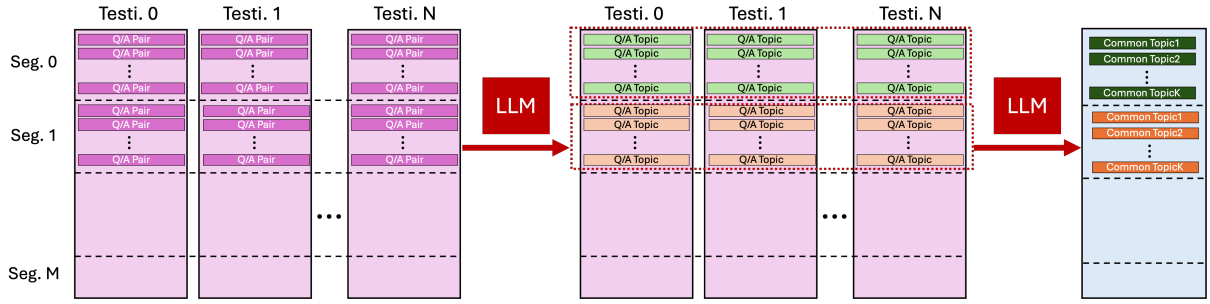


Figure 7: The topical sequence extraction pipeline.

## 7. LLM Prompts

### 7.1. Q/A Topic Naming

```
"""
You are a Holocaust researcher.
You will be presented with a short text snippet from a
conversation between an interviewer and a Holocaust survivor.
'INT' represents the interviewer and 'SUBJECT' or
'<survivor_name>' represents the survivor.
Given the text snippet please generate a short title describing
the most prominent topic in the text.
Make sure that the title is short and limited to a few words.
Make sure that the title is comprehensive, specific,
interpretable, and short.
Make sure that the title captures only a single topic.
Output format:
Title: "<title>"
Reason: "<reason>"
Text snippet:
"{text_snippet}"
"""
```

### 7.2. Common Topics Extraction

```
"""
You are a Holocaust researcher.
You will be presented with a set of titles representing topics
extracted from Holocaust survivor interviews.

Title Set:
{title_set}

Your task is:
- Generate {num_topics} distinct titles that best describe the
most common and prominent titles in set.
- Titles must be concise (maximum of a few words), specific,
interpretable, and distinct.
- Do NOT combine multiple topics into a single title or use
conjunctions like “and”.

Desired output format:
{output_format}
```

```
The common titles are:
1.
"""
```

### 7.3. Question Type Classification

```
"""
You are a Holocaust researcher analyzing survivor testimonies.
The testimonies are transcripts of an oral interviews.
The interviews is composed of speaker sides -- the interviewer
questions and the survivor answers.
A prefix of "INT" identifies an "interviewer" line.
A prefix of "<initials>" identifies a suviror line.
You will be given one question asked by an interviewer during a
conversation with a Holocaust survivor.

Your task is to classify the question into exactly **one** of
the following types, based on its structure and intent:

Question Types:
1. How Question - Asks about a method, process, or manner
2. What Question - Asks for information, descriptions, or
clarifications
3. When Question - Asks about time or timing
4. Where Question - Asks about a place or location
5. Why Question - Asks about cause, motivation, or reason
6. Who Question - Asks about a person or group
7. Other Question - Does not match any of the categories
above or is ambiguous.

Then, write a brief explanation (no more than one sentence)
explaining why you selected that type.

Do not return more than one type.

### Output JSON format:
{{
    "type": "<one of the 7 types above>",
    "explanation": "<one-sentence explanation>"
}}

### Input:
Interviewer Question: "{speaker_line}"
"""
```

# From Oral History to Structured Data: The MalachNER Dataset

**Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián, Pavel Pecina**
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
{bruckner,hofmeisterova,kocian,pecina}@ufal.mff.cuni.cz

**Abstract**

We present MalachNER, a new multilingual dataset for Named Entity Recognition (NER) in oral testimonies of Holocaust survivors. MalachNER has been sourced from different archives and annotated based on comprehensive domain-specific guidelines refined by a collaboration of international experts. Covering 10 European languages, differs significantly from previously released datasets: It is primarily based on noisy, verbatim transcribed speech, rather than on digitized written documents. These transcripts are characterized, among other challenges, by fillers, dialectal speech, and in-line annotations indicating incomprehensible words, which are not commonly encountered in other datasets. However, large volumes of yet unprocessed oral history make such a dataset a necessity. In addition to the description of the dataset and its annotation guidelines, we show with baseline experiments that MalachNER is complementary with previously released data, and the key to training domain-specific language models that generalize well to written and oral testimony alike, achieving state-of-the-art performance on both types of documents.

**Keywords:** Named Entity Recognition, Multilingual, Dataset, Speech, Holocaust Testimony

## 1. Introduction

While Named Entity Recognition (NER) is a well-known language processing task, it is still relatively unexplored in speech (Caubrière et al., 2020) and in the historical domain (Ehrmann et al., 2023). In particular, annotated language resources related to Nazi persecution are scarce (Carter et al., 2022; Anuradha Nanomi Arachchige et al., 2023) despite the existence of enormous archives such as the USC Shoah Foundation's Visual History Archive[1].

These archives are primarily based on oral testimony of Holocaust survivors. The nature of these documents introduces several challenges, since speech transcripts, regardless of whether they have been transcribed manually or automatically, include speech artifacts that are missing in most NER datasets. These artifacts include, for example, filler words and in-line annotations denoting inaudible words. Such noise cannot always be reliably removed automatically, which is amplified by the variety of artifacts increasing with the number of languages, source archives, and transcriptionists. The necessity of a dataset providing these challenges arises from the fact that oral history is an important resource for Holocaust studies (Vrzgulova, 2024), and manually curated transcriptions are often not available (Lehečka et al., 2023).

In the following, we present MalachNER, a multilingual NER dataset of manually transcribed Holocaust survivor testimonies sourced from different archives, annotated by domain experts who are speakers of Croatian, Czech, Danish, Dutch, English, German, Hungarian, Polish, Serbian, and Slovak. Finally, we compare different domain-specific NER models and show that the processing of speech-transcribed documents is challenging for language models trained exclusively on written testimony, and vice versa; but state-of-the-art performance on both types of documents can be achieved by a model fine-tuned simultaneously on written and oral history.

MalachNER is published under a closed license for academic usage on LINDAT[2]. The source code of the NER experiments is available on GitHub[3], and the best model produced by these experiments is available on Hugging Face[4] under the MIT license. The annotation guidelines are included in the dataset repository and additionally mirrored[5] for open access.

## 2. Related Work

The most notable multilingual NER dataset specific to Holocaust testimony is EHRI-NER (Dermentzi and Scheithauer, 2024). The documents in this dataset have been repurposed from the EHRI Online Editions[6], which consist of digitized written documents that were not originally annotated for

[1] https://vha.usc.edu

[2] http://hdl.handle.net/20.500.12800/1-6129

[3] https://github.com/chbridges/malach-ner

[4] https://huggingface.co/ufal/xlm-roberta-ehri-malach-ner

[5] https://ufallab.ms.mff.cuni.cz/~bruckner/htres2026/

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



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 59–65
11 May 2026. 

| Source | Type | Example |
|---|---|---|
| Mlynář (2016) | pause | This was my first sibling. {{....}} You want to continue about my siblings? |
| | uncertain word | Ah... it was in in {{Kazinci}} street... |
| | background noise | My fathers name was {{Video stops for a moment}}, mother was Regina... |
| VHA | incomprehensible | We called them [NON-ENGLISH]. |
| | long pause | And [PAUSES FOR 3 SECONDS] they went to Sweden. |
| | short pause | An ex– a– a– a Pole, a Christian Pole came back from the United States. |
| | uncertain word | I had [? otherwise ?] impression that the relationship was good. |
| FU Berlin | incomprehensible | Now, I think that living in this_. We still went to school. |
| | pauses, comments | (-) And so, (–) in 1944, uh, winter 43 [1943] my mother became very ill. |
| | video description | <end of tape 1> |

Table 1: Examples of noise appearing in different transcripts.

NER, but for Entity Linking. As such, the EHRI-NER also tags generic non-named entities such as "barracks" or "family", given that these entities can be disambiguated within the context of the document they appear in. EHRI-NER extends the common tags Person, Organization, and Location, with the domain-specific tags Ghetto, Camp, and Date. Anuradha Nanomi Arachchige et al. (2023) proposed an iterative hybrid annotation approach for the annotation of Holocaust-specific name entities in digitized English-only Holocaust testimonies using a granular domain-specific entity type ontology, including specific tags such as Warships and Rivers. Ehrmann et al. (2023) point out the arising challenges in historical documents, including different types of noise introduced by digitized written documents and transcribed speech, as well as historical language and entities not accounted for by language models trained on modern text.

## 3. Dataset Description

### 3.1. Data Source and Collection

#### 3.1.1. Testimony archives

Most of the testimonies appearing in MalachNER originate from the Visual History Archive (VHA) and have been manually transcribed and published either by the USC Shoah Foundation itself or by Freie Universität Berlin within the "Zeugen der Shoah" project[7]. Ten languages have been chosen based on the availability of domain experts who speak these languages.

The transcripts taken directly from the VHA include interviews with Mark Verstandig (English, longest VHA interview), Walter Guttmann (Dutch), Ruth Felix (Czech), and Halina Elczewska (Polish). The transcripts sourced from FU Berlin include interviews with Simon Wiesenthal (German, exceptional relevance), Lajos Erdélyi (Hungarian), Softic Sadrudina Gavrankapetanovic (Croatian), and Branislav Ackovic (Serbian). Additionally, the longest transcribed Danish interview, with Rosalin Christensen, has been provided by Mlynář (2016). Most of these testimonies have been selected by a domain expert according to different factors such as the length, relevance to the domain, and the density of named entities. For Croatian and Serbian, no other manual transcripts are available.

Three additional short interviews in Czech, with Karel Blahouš, Maria Kotrbáčková, and Drahomíra Blosgebrová, have been obtained from the United States Holocaust Memorial Museum (USHMM)[8] to match the proportion of Czech with the other languages and increase the variety of speakers. A Slovak interview with Rozália Guttmannová has been provided by the Milan Šimečka Foundation[9].

The dataset contains the transcription of 13 testimonies of approximately 37 hours of speech in total. Each testimony is split into "tapes", where each tape covers approximately 30 minutes of speech. Except for the significantly longer English and German interviews and the slightly smaller Croatian and Serbian data, the languages are proportional in terms of tokens.

#### 3.1.2. Preprocessing

The transcripts mentioned in Section 3.1.1 have varying degrees of noise. The Slovak transcript has been carefully curated by the Milan Šimečka Foundation and does not, in fact, feature any speech artifacts or in-line annotations, and interruptions or inquiries by the interviewer are scarce. The Czech interviews from USHMM are more conversational, but speech artifacts in their transcripts are limited.

The remaining interviews, on the other hand, have been transcribed in much more detail, reproducing the original speech more faithfully, annotating pauses, and containing markers for incomprehensible utterances or background noise, as well as comments. Similar to language-dependent filler words, in-line annotations can switch languages and contain typos. This is particularly extreme in the FU Berlin transcripts, where not all noise can be

[7] https://transcripts.vha.fu-berlin.de

[8] https://collections.ushmm.org

[9] https://nadaciamilanasimecku.sk

removed safely. A comprehensive list of noise examples is given in Table 1. Generally, all noise that can be removed safely has been removed or substituted with simple regular expressions. However, sufficient speech artifacts, such as filler words and repetitions, remain in the text to reproduce speech more accurately and make the task significantly more challenging than EHRI-NER (Dermentzi and Scheithauer, 2024), which is primarily based on digitized written testimony and protocols.

### 3.2. Annotation Guidelines

After preprocessing, domain experts have annotated the testimony transcripts in LabelStudio (Tkachenko et al., 2020-2022) with a combination of the general-domain entity types defined by CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and additional domain-specific entity types defined by EHRI-NER (Dermentzi and Scheithauer, 2024). The resulting dataset contains the following types:

- **PER (Person)** Names of identifiable individuals and families, not including titles.

  E.g., ~~Dr~~ György Neuhauser, Führer, Novákovi (family), Petrův ("Petr's", possessive adjective)

- **ORG (Organization)** Named groups, institutions, or firms.

  E.g., SS, Nazis, Czechoslovak Army, Germans

- **LOC (Location)** Geographic locations, including states, cities, names of temples, synagogues, and rivers.

  E.g., Vienna, Kazinczy Street, Danube, Austro-Hungarian Empire

- **CAMP** Nazi Camps (e.g., concentration camps, extermination camps, transit camps). Used as a more specific subtype of LOC.

  E.g., Birkenau, Auschwitz, Theresienstadt family camp (section of Auschwitz II-Birkenau)

- **GHETTO** Jewish Ghettos. Used as a more specific subtype of LOC.

  E.g., Theresienstadt, Budapest (if referring to one of the ghettos in the city)

- **DATE** Calendar dates and years mentioned in the text. Does not include underspecified times such as "February" or "Monday".

  E.g., 11th February 1940, 13.4.44, '42

- **MISC** Entities that cannot be assigned to any other tag, such as historical events, ethnicities, religious groups, ideologies, or languages.

  E.g., Reichspogromnacht, Mein Kampf, Germans, Jewish, Holocaust, Zyklon B

Additionally, a TERM category for Holocaust-specific terminology such as "gas chambers" has been defined. This category has not been used for annotations, but helped annotators disambiguate between MISC entities and generic terms that do not refer to specific named entities. The EHRI vocabularies[10] have been used as a help to disambiguate LOC, CAMP, and GHETTO.

The annotation guidelines are largely based on guidelines composed for NER and Entity Linking annotations of testimonies from the Wiener Holocaust Library and the Hungarian Jewish Museum and Archives[11] (see Dermentzi and Scheithauer, 2024), but have been adapted to exclusively NER and actively expanded in collaboration with the annotators to resolve as many ambiguities as possible.

### 3.3. Dataset Statistics

10/23 English and 4/7 Czech tapes have been annotated by additional annotators. As suggested by recent literature (Mayhew et al., 2024), we measure the inter-annotator agreement by computing the $F_1$ scores of annotated spans using the `seqeval` framework (Nakayama, 2018). Czech has a high overall agreement of 95%, with perfect agreement on Person and Ghetto entities. The agreement on English is lower (81%), since the additional annotator processed the documents when the guidelines were still in an early state and many ambiguities were unresolved. In particular, the lowest agreements on ORG (68%) and MISC (78%) stem from the ambiguity of terms such as "Germans", which can refer to the German military or ethnicity.

The produced dataset has been tokenized with UDPipe (Straka, 2018), since by default, LabelStudio tokenizes the documents in a non-standard way, without splitting punctuation marks from words and thus tagging them as parts of entity spans. Table 2 shows the distribution of tokens and entities for each language. MalachNER is slightly smaller than EHRI-NER and has a smaller entity density stemming from the noise present in the speech transcripts. The distribution of entity types is similar, although MalachNER has fewer tagged dates and a significant lack of ghettos, as they are rarely mentioned by name. This discrepancy is possibly due to the fact that in EHRI-NER, also non-named entities are tagged. Secondly, different geographically conditioned trajectories of the Holocaust might have this effect.

The annotated dataset is split into training and test splits for each language. These splits have not

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

[11] https://www.ehri-project.eu/call-for-applications-unlocking-holocaust-testimony-ehri-clarin-datathon-workshop/

| Language | Tokens | PER | ORG | LOC | CAMP | GHETTO | DATE | MISC | Total |
|---|---|---|---|---|---|---|---|---|---|
| Croatian | 17,405 | 137 | 60 | 211 | 0 | 3 | 62 | 171 | 644 |
| Czech | 27,751 | 235 | 50 | 263 | 18 | 18 | 44 | 306 | 934 |
| Danish | 26,434 | 24 | 83 | 113 | 102 | 7 | 27 | 159 | 515 |
| Dutch | 23,369 | 157 | 119 | 338 | 80 | 2 | 114 | 188 | 1,028 |
| English | 96,279 | 621 | 381 | 879 | 12 | 0 | 170 | 1,230 | 3,293 |
| German | 75,132 | 558 | 342 | 638 | 123 | 0 | 169 | 695 | 2,525 |
| Hungarian | 25,402 | 205 | 58 | 311 | 27 | 0 | 70 | 327 | 998 |
| Polish | 23,403 | 187 | 76 | 93 | 23 | 15 | 27 | 57 | 478 |
| Serbian | 15,615 | 53 | 47 | 124 | 3 | 0 | 26 | 44 | 297 |
| Slovak | 24,710 | 110 | 105 | 152 | 87 | 3 | 66 | 241 | 764 |
| Total | 355,500 | 2,287 | 1,321 | 3,122 | 475 | 48 | 805 | 3,418 | 11,476 |

Table 2: The number of tokens and annotated entities per language. The last column denotes the total number of entities.

been sampled randomly: First mentions of camps and ghettos commonly appear in the second or third tape of a testimony. Thus, every fifth tape, starting with tape 2, serves as test data, resulting in test ratios between 20% and 30% per language. For instance, from an interview with 23 available tapes, tapes 2, 6, 11, 16, and 21 are selected as the test split, resulting in a 22% test ratio.

## 4. Baseline Experiments

### 4.1. Experimental Setup

Two architectures are considered as baseline models: `XLM-RoBERTa-large` (Conneau et al., 2020) and `XLM-RoBERTa-ehri-ner-all`[12], which is `XLM-RoBERTa-large` fine-tuned on the EHRI-NER dataset (Dermentzi and Scheithauer, 2024). The models are evaluated in two experiments:

1. `XLM-RoBERTa-ehri-ner-all` is evaluated on the EHRI-NER and MalachNER test sets. Then, it fine-tuned further on the MalachNER training set, and re-evaluated on both test sets.
2. `XLM-RoBERTa-large` is fine-tuned from scratch on the training splits of EHRI-NER, MalachNER, and on both datasets at once. Each time, it is evaluated on the test splits of both datasets.
3. The best model fine-tuned on both datasets is further evaluated on each individual language in both datasets.

Hyperparameters are the same as those used by Dermentzi and Scheithauer (2024): Models are trained for 3 epochs using a learning rate of 3e-5 with weight decay at 0.01 and a batch size of 16. In addition to a seed of 42, the training is repeated with seeds 0 and 1234 to report 95% confidence intervals of the mean $F_1$ scores. When fine-tuning on MalachNER, 20% of the sentences in the training set are sampled with a fixed seed of 42 to create a held-out development set. When evaluating on EHRI-NER, the predictions of the additional MISC entity type are removed.

[12] https://huggingface.co/ehri-ner/xlm-roberta-large-ehri-ner-all

### 4.2. Results

For the first two experiments, the 95% confidence intervals of tag-wise and overall $F_1$ scores are reported in Table 3. The $F_1$ scores of the per-language evaluation of the best model are reported in Table 4. The different nature of the two datasets becomes obvious in Table 3: Models trained only on EHRI-NER perform badly on the speech transcripts of MalachNER, whereas models trained only on MalachNER perform badly on EHRI-NER. While the continued fine-tuning of `XLM-RoBERTa-large-ehri-ner` on MalachNER leads to the best models on MalachNER, in particular for the recognition of ghettos, the $F_1$ scores on EHRI-NER drop significantly, indicating the forgetting of earlier learned concepts.

The best possible solution to handle written and oral testimony at once is to combine both datasets and sample from both during fine-tuning: The resulting models achieve $F_1$ scores comparable with the best models fine-tuned only on one dataset for all entity types except for organizations and, in the case of MalachNER, ghettos. However, the confidence intervals suggest that the improvements are not significant in most cases. The lower score for organizations can be explained by the ambiguity with non-organizational groups of people, such as ethnicities and religious groups, which are covered by the added MISC type in MalachNER.

In addition to ORG, the GHETTO tag stands out in the MalachNER experiments with low $F_1$ scores and high variance. Here, the results benefit more from fine-tuning on the EHRI-NER training set than on the MalachNER training set. This can be explained by the lack of ghettos mentioned by name in the sourced testimonies, which was also shown

| | | XLM-R-large-ehri-ner | | XLM-RoBERTa-large | | |
|---|---|---|---|---|---|---|
| | | **Frozen** | **Fine-tuned**$_M$ | **Fine-tuned**$_E$ | **Fine-tuned**$_M$ | **Fine-tuned**$_{EM}$ |
| EHRI-NER | PER | 87.00 | 82.00 ± 0.00 | 86.00 ± 2.48 | *74.33 ± 1.43* | **86.67 ± 1.43** |
| | ORG | 63.00 | 41.33 ± 5.17 | **65.33 ± 1.43** | *33.67 ± 1.43* | 62.33 ± 1.43 |
| | LOC | 82.00 | 67.67 ± 1.43 | **82.00 ± 2.48** | *59.67 ± 3.79* | **82.00 ± 2.48** |
| | CAMP | 70.00 | 62.33 ± 3.79 | **73.33 ± 1.43** | *41.00 ± 13.14* | 72.00 ± 4.30 |
| | GHETTO | 80.00 | 77.67 ± 2.87 | 84.33 ± 2.87 | *76.00 ± 17.39* | **85.00 ± 2.48** |
| | DATE | 84.00 | 66.67 ± 14.34 | **86.67 ± 1.43** | *49.67 ± 23.61* | 81.33 ± 5.17 |
| | Overall | 81.00 | 67.67 ± 3.79 | **82.00 ± 0.00** | *58.00 ± 4.97* | 81.00 ± 2.48 |
| MalachNER | PER | 79.00 | **90.67 ± 1.43** | *83.00 ± 2.48* | **90.67 ± 1.43** | **90.67 ± 1.43** |
| | ORG | 29.00 | **72.67 ± 5.17** | *25.00 ± 2.48* | 72.33 ± 5.17 | 70.00 ± 4.30 |
| | LOC | 71.00 | **90.00 ± 0.00** | *66.67 ± 2.87* | 89.67 ± 1.43 | 89.00 ± 0.00 |
| | CAMP | 66.00 | 74.00 ± 2.48 | *69.33 ± 14.56* | 75.33 ± 7.17 | **77.67 ± 6.25** |
| | GHETTO | 67.00 | 67.00 ± 23.96 | **71.00 ± 6.57** | *55.67 ± 1.43* | 68.00 ± 0.00 |
| | DATE | 50.00 | 83.67 ± 7.99 | *53.33 ± 5.17* | 81.33 ± 6.25 | **84.33 ± 3.79** |
| | MISC | – | – | – | **84.67 ± 1.43** | *83.00 ± 0.00* |
| | Overall | 66.00 | **85.67 ± 1.43** | *64.33 ± 1.43* | 84.67 ± 1.43 | 84.33 ± 1.43 |

Table 3: Mean $F_1$ scores and 95% confidence intervals of different models fine-tuned and evaluated three times on EHRI-NER and MalachNER. The subscripts *E* and *M* denote fine-tuning on EHRI-NER and MalachNER, respectively, whereas *EM* denotes that the model samples from both datasets during fine-tuning. Note that the column **Fine-tuned**$_E$ reproduces the frozen XLM-RoBERTa-large-ehri-ner model with different initial seeds. **Best** and *worst* results of the newly fine-tuned models are marked.

| Test Split | | PER | ORG | LOC | CAMP | GHETTO | DATE | MISC | Overall | Support |
|---|---|---|---|---|---|---|---|---|---|---|
| EHRI-NER | cs | 0.93 | 0.43 | 0.81 | 0.72 | 0.87 | 0.89 | – | 0.82 | 536 |
| | de | 0.85 | 0.61 | 0.82 | 0.79 | 0.87 | 0.80 | – | 0.80 | 802 |
| | en | – | – | 1.00 | – | – | – | *0.00* | 0.67 | 1 |
| | fr | – | – | 1.00 | – | – | – | – | 1.00 | 1 |
| | hu | 0.94 | 0.67 | 0.71 | 0.79 | – | 0.82 | *0.00* | 0.75 | 79 |
| | nl | 1.00 | 0.89 | 1.00 | – | – | – | – | – | 9 |
| | pl | 0.84 | 0.80 | 0.79 | 0.73 | 0.67 | 0.62 | – | 0.77 | 117 |
| | sk | 1.00 | 1.00 | 0.95 | – | – | – | – | 0.97 | 17 |
| | yi | 0.73 | 0.47 | 0.77 | 0.19 | 0.33 | 0.00 | – | 0.71 | 217 |
| MalachNER | cs | 0.87 | 0.82 | 0.64 | 0.96 | 1.00 | 0.77 | 0.92 | 0.86 | 287 |
| | da | 0.86 | 0.43 | 0.90 | 0.90 | 0.00 | 0.88 | 0.66 | 0.74 | 186 |
| | de | 0.93 | 0.72 | 0.87 | 0.76 | – | 0.82 | 0.79 | 0.83 | 554 |
| | en | 0.92 | 0.67 | 0.97 | *0.00* | – | 0.97 | 0.91 | 0.91 | 716 |
| | hr | 0.94 | 1.00 | 1.00 | – | 1.00 | 0.75 | 0.81 | 0.87 | 76 |
| | hu | 0.95 | 0.82 | 0.95 | 1.00 | – | 0.80 | 0.78 | 0.85 | 134 |
| | nl | 0.93 | 0.82 | 0.85 | 0.63 | 0.00 | 0.76 | 0.65 | 0.80 | 288 |
| | pl | 0.85 | 0.60 | 0.65 | – | 0.57 | 0.60 | 0.20 | 0.65 | 90 |
| | sk | 1.00 | 0.98 | 0.81 | 0.70 | 1.00 | 0.92 | 0.88 | 0.89 | 231 |
| | sr | 0.82 | 0.36 | 0.96 | 0.00 | – | – | 0.74 | 0.81 | 71 |

Table 4: $F_1$ scores of the best NER model for all languages in both datasets. The last column shows the total number of annotated entities in the test split. Note that the published EHRI-NER test splits are not representative for all languages. Scores of *0.00* in *italics* indicate that the model predicted an entity type not present in the test split, which is generally the case for the MISC type in EHRI-NER. In all other cases where the score is 0.00, the test split contains only up to 3 instances of the corresponding entity type.

by Dermentzi and Scheithauer (2024) in the EHRI-NER dataset. As can be seen in Table 4, most of the test data lacks this entity type. Special attention should be given to this tag during model selection, and this problem can likely only be solved with additional data containing more annotated named ghettos. The surprisingly low scores for Polish in MalachNER indicate inconsistencies in the annotations either within MalachNER or with the Polish EHRI-NER data.

## 5. Conclusions

We presented MalachNER, a new Named Entity Recognition dataset based on transcribed oral testimonies in 10 languages. Expanding on entity ontologies defined by previous NER datasets, we compiled comprehensive domain-specific annotation guidelines and processed approximately 37 hours of speech with the help of domain experts. Baseline experiments show that MalachNER is complemen-

tary to existing datasets, and a model fine-tuned simultaneously on written and oral history can bridge the gap between these types of documents, achieving comparable state-of-the-art results on both despite the added challenge of noise.

On the other hand, NE-annotated Holocaust-related language resources are still scarce, which is reflected in the models' performance in tagging organizations and ghettos. This can be tackled in the future by acquiring more annotated data, but also by fine-tuning models that were pre-trained on large amounts of unannotated domain-specific data (Brückner et al., 2026). Furthermore, noise removal techniques can be explored to improve the data quality.

The MalachNER dataset, its annotation guidelines, experimental code, and the best model resulting from the experiments are published and can be accessed via the hyperlinks found in the introduction.

## 6. Limitations

While aiming to cover as many languages as possible, the number of languages has been limited by the number of available speakers of these languages. As a result, only one of the ten languages is represented by more than one speaker and thus exhibits more variance in speech. Furthermore, the amount of text per language has been limited by the availability of manual transcripts, as well as the time budget of the annotators, leading to a language imbalance in the created dataset. Due to licensing, the data is published under a closed license.

## 7. Acknowledgements

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

The work described herein has been using tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

This research was partially supported by Charles University, project GA UK No. 380126 and SVV project number 260 821.

## 8. Bibliographical References

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*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Christopher Brückner, Jan Lehečka, Jan Švec, and Pavel Pecina. 2026. Modeling the language of holocaust survivors' testimony with domain-adapted transformers. In *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, Palma de Mallorca, Spain. ELRA.

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*, 37(3):837–858.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France. European Language Resources Association.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Interspeech 2023*, pages 201–205.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Jakub Mlynář. 2016. Pluralita identit v autobiografickém vyprávění československých Židů žijících v zahraničí. *HISTORICKÁ SOCIOLOGIE*, 2016:33–51.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Monika Vrzgulova. 2024. Open forum: Oral history in holocaust research. *Eastern European Holocaust Studies*, 2(1):151–157.

# Emotions In Oral History Interviews: A Multimodal Approach to Holocaust Testimonies

**Nele Mantaj, Vaibhav Agarwal, Inés Matres**
University of Trier, Technical University of Munich, University of Helsinki
s2nemant@uni-trier.de, vaibhav.agarwal@tum.de, ines.matres@helsinki.fi

**Abstract**

Video interviews with Holocaust survivors and witnesses comprise, to date, the most globally distributed and comprehensive oral history documentation. As survivors among us disappear, these sources are increasingly important to understand the impact of the Holocaust and mechanisms to overcome the trauma experienced. While historians often rely on written transcripts, these omit emotional nuances conveyed through audiovisual cues such as facial expressions, pauses, and eye movements. This article outlines the resources, data-preparation steps, and analytical methods used during a 10-day Digital Humanities Hackathon project to examine emotions in Holocaust testimonies, incorporating video, audio, and text. The group aimed to determine whether audiovisual signals offer meaningful emotional or sentimental information beyond transcripts. To achieve this, the group worked with a sample of 10 interviews facilitated by the US Holocaust Memorial Museum (USHMM); which were separated into video, audio, and textual components for machine processing and realigned side-by-side for analysis. This resulting "cookbook" lays out a workflow, resources, and practical entry points for preparing oral history interviews for multimodal emotion and sentiment annotation, or to aid the detection of emotionally significant moments for deeper examination.

**Keywords:** Holocaust, Oral History interviews, Multimodal analysis, Emotion analysis

## 1. Background

Testimonies are a crucial source for understanding the Holocaust, providing first-hand accounts and personal narratives from survivors and witnesses. These narratives offer insight into individual experiences; however, much of the existing analysis has focused primarily on the accounts of events from these testimonies (Waxman, 2012). In order to leverage the large amounts of oral testimony in diverse oral history archives, machine learning (ML) methods based on transcripts have proven accurate in identifying topics (Ifergan et al., 2024), named entities, such as places or events (Anuradha Nanomi Arachchige et al., 2023), or relationships in Holocaust and other oral history testimony (Anuradha et al., 2023; Laato et al., 2025). Our approach is inspired by the turn in historical research that considers emotion as an important lens to examine past phenomena that helps bridge the gap between personal and collective; and between experience and expression (Eustace et al., 2012). While emotional expressions within these accounts have been explored through qualitative methods, a large-scale approach to emotion with oral testimony on the Holocaust has been minimal and only has aimed to improve Automatic Speech Recognition (ASR) methods (Bukreeva et al., 2023).

The main goal of this paper is to improve the preconditions for a more holistic emotional analysis of Holocaust testimony at scale, by laying out the processes involved in tackling multimodal information contained in interviews. After a detailed description of the data used, firstly, we present a workflow for splitting the interview into three components making each signal (transcript, audio and video) fit for machine-processing while interoperable to be analysed side-by-side (section 3). Secondly, we show what insights about emotion and sentiment can be exacted with aid of digital methods from the video and audio in addition to the transcript analysis (section 4); Finally, in sections 4 and 5 we discuss limitations of computational models and vistas for developing ASR and computer vision to support this approach.

## 2. Hackathon setting and data used

The setting of this project was an academic Hackathon in the space of 10 days in Spring 2025 at the University of Helsinki. The group formed by nine MA students from language studies, history, data and computer science and four instructors [1] gained access to a sample of 100 testimonies of Holocaust survivors and witnesses from the United States Holocaust Memorial Museum [2]. The data selection was done by researchers from the CLARIN Network [3] specialists in corpus linguistics. The dataset was balanced in terms of gender, distribution of witnesses and survivors; and one particular criteria was its language diversity, including 30 interviews in Czech, 23 in Polish, 20 in English, 15 in Dutch and 12 in French (Anuradha et al., 2026).

Although the dataset was not created attending to

[1] seeacknowledgements
[2] https://www.ushmm.org/
[3] https://www.clarin.eu/



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 66–73
11 May 2026. 

their emotional content, highly emotional moments were identified by all students in preparation for the Hackathon. This consisted in viewing at least two interviews and suggesting topics to explore. Early on and after consulting with oral history experts among the instructors, the emotional content and heterogenous expressiveness of interviewees were the most remarkable for students and became the target of our project.

While emotional content could be found in any interview, we soon remarked that ways of expressing and verbalizing emotions were highly dependent on the individual rather than being defined by gender, language spoken, or type of testimony. This allowed some freedom when sampling this huge dataset. A final selection of 10 videos was done (see table 1) with the main requirement that they contained full-transcripts and videos. Additionally, these interviews lasted approximately one hour, because video models took substantial time in processing the material. Balance in gender was maintained and the language diversity was respected in selecting an equal number of interviews in English and Polish, as the group included native speakers of both languages to ensure verification of results. A witness bias (7 of the 10) is due to survivor testimonies being in average longer, but for the purpose of this study, this had not much impact, as both long and short interviews could contain strong emotional content, and each individual is unique in expressing or containing their emotions.

| No. | Type | Born | Lang. | Gender |
|---|---|---|---|---|
| 1 | Survivor | 1933, Germany | EN | Female |
| 2 | Witness | 1915, Unknown | EN | Male |
| 3 | Survivor | 1934, today Czech Rep. | EN | Female |
| 4 | Survivor | 1929, Poland | EN | Male |
| 5 | Witness | Unknown, United States | EN | Male |
| 6 | Witness | 1915, today Ukraine | PL | Male |
| 7 | Witness | 1917, Poland | PL | Female |
| 8 | Witness | 1914, Poland | PL | Female |
| 9 | Witness | 1931, Poland | PL | Female |
| 10 | Witness | 1917, Poland | PL | Male |

Note: EN = English, PL = Polish

Table 1: Subset of interviews used in the hackathon

## 3. Workflow

In this section we explain the steps taken to extract the three signals from the interviews in independent layers, and the processes to transform these into machine-readable formats and applying models and results. In doing this, dependencies emerged and some of these processes were done concomitantly, hence while an illustration of the workflow and dependencies is shown in Figure 1 (next page), for clarity we describe them separately.

In addition to more established sentiment and Emotion analysis applied to text (transcripts), paralinguistic features such as speech patterns, silences, changes in voice, hesitations, or gestures, provide cues for emotional intensity and variation in oral history interviews. To account for granularity for these features we choose to test specialised models which prioritized accuracy and input data type over generalisable output produced using LLMs. Adopting this approach also shed light on the pitfalls of the current state of the art models, which are discussed in section 5.

It is also important to note that a full analytical pipeline should aim to model the entire spectrum of paralinguistic features central to oral testimonies; however, given the setting and time constraints associated to a Hackathon project, modeling these features turned out to be infeasible. While it was possible to successfully detect silences and perform diarization to ensure coherent audio utterances and corresponding text transcripts, the pipeline had to rely on computational models which captured these features in hidden layers (for example, using the Wav2vec 2.0 model based on (Wagner et al., 2023) for valence and arousal detection), instead of an analysis catered towards the paralinguistic features, which would have added further nuance to the results from the modalities.

### 3.1. Text

The original 100 piece dataset included heterogeneous text material in addition to the video recordings. These could be transcriptions in PDF format in their original language, a few had additional translations in English, and some had summaries instead of transcripts. This heterogeneity posed several technical challenges: transcripts were stored in non-machine-readable formats, varied greatly in structure, language, and completeness, and some interviews lacked transcripts entirely. Since the interviews were conducted in different languages, this required language-specific processing pipelines and models.

Existing PDF transcripts were first converted into plain TXT files to enable natural language processing. Due to inconsistent formatting, automatic methods were insufficient to reliably identify speakers or dialogue turns. Therefore, the text was automatically segmented into utterances, defined as uninterrupted sequences of speech by a single speaker. Assigning each utterance to a speaker (interviewer or interviewee) required manual annotation. Finally, the annotated transcripts were converted into structured JSON files. Each JSON object corresponded to a single utterance and included metadata such as speaker identification and interview ID, making the data directly compatible with machine learning models.

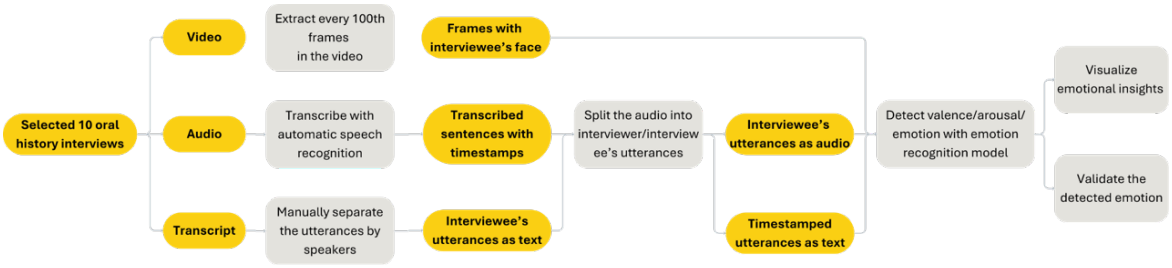


Figure 1: Proposed workflow for multimodal emotion analysis of oral history interviews

For interviews without transcripts, automated speech-to-text transcription was tested. A Whisper-based large language model fine-tuned for French speech recognition was used and produced acceptable transcription quality for French interviews, although manual correction and annotation were still required. In contrast, automated transcription for Polish interviews yielded poor results. This contributed to the decision to adjust the dataset, ultimately relying exclusively on existing Polish and English transcripts, annotated as described above.

All sentiment, emotion, and affective analyses were performed using transformer-based language models, which represent the current standard in natural language processing. Transformers model contextual relationships between all tokens in a text using self-attention mechanisms, enabling them to capture nuanced semantic and emotional information beyond keyword-based approaches. The models used in this workflow were retrieved via the Hugging Face platform.

Sentiment analysis aimed to classify text segments according to their overall polarity (positive, negative, or neutral). Transformer-based sentiment models derived sentiment from contextual embeddings, making them robust to linguistic phenomena such as negation. For English data, the RoBERTa-based model *cardiffnlp/twitter-roberta-base-sentiment*[4] was used, which output predictions for positive, neutral, and negative sentiment. For Polish data, the GPT-2–based model *nie3e/sentiment-polish-gpt2-large*[5] was applied. In addition to positive, negative, and neutral labels, this model output an "ambiguous" category, which was excluded from further analysis to avoid uncertain classifications.

The Emotion analysis expanded sentiment analysis by identifying specific emotional categories. For English transcripts, the model *j-hartmann/emotion-english-roberta-large*[6] was used, targeting seven emotions (anger, disgust, fear, joy, neutral, sadness, and surprise). For Polish data, the model *hplisiecki/polemo_intensity*[7] was applied, which predicted intensity scores for six emotions (happiness, sadness, anger, disgust, fear and pride) but lacked a neutral label. These models output continuous scores rather than single categorical labels.

In addition to categorical emotions, affective states were modelled along the continuous dimensions of valence and arousal. For English data, valence was estimated using the transformer-based model *chrlukas/stories-emotion-c0*[8]. For Polish data, the emotion detection model also provided valence and arousal estimates. Since model outputs differed in scale, all valence scores were linearly transformed to a standardized range between $-1$ and 1 to ensure comparability.

### 3.2. Audio

The task of analyzing the interviewees' speech to track emotional changes over time involved further pre-processing. This required segmenting the speech into coherent parts that contained relevant information about the emotions portrayed in the speech. The pre-processing had to take into consideration the current speaker, the flow of the speech and possibly the topics handled.

In natural speech, emotional states are dynamic, i.e., they fluctuate over time. This is important to note when looking at oral testimonies due to their long duration. When survivors or witnesses recall distinct events, their emotional expression varies across different segments. To capture these nuances, it was necessary to split the continuous audio into coherent segments. Following this, we segmented the speech into utterances, defined as uninterrupted chains of speech that follow the speaker's natural flow. This ensures input for the emotion detection models remaining contextually consistent.

[4] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[5] https://huggingface.co/nie3e/sentiment-polish-gpt2-large

[6] https://huggingface.co/j-hartmann/emotion-english-roberta-large

[7] https://doi.org/10.1007/s11135-025-02116-8

[8] https://huggingface.co/chrlukas/stories-emotion-c0

Furthermore, we were only interested in the interviewee's emotions, so we didn't need to predict the interviewer's emotional expression. This was possible through diarization, which was done using the powerset multi-class segmentation code available in the open-source *pyannote.audio*[9] library to detect the timestamps when the speaker changes in an audio file (Plaquet and Bredin, 2023; Bredin, 2023).

To find the optimal way to split the audio into utterances, we had to select what level of intensity is the cut-off point for a "silence". Some interviews had very long periods of silence that we had to investigate by listening to the recording, while in others, the silence was a part of the natural flow and added value. Some interviews also contained quiet periods where the interview team changed the recording tape, but it could have been possible that some silent periods were marked erroneously. The silences were detected based on Rudolfbyker's code to *split wav*[10] files by silence which takes into account how quiet must the audio be, as well as, for how long must the silence last before noting the cut.

It is also important to note that during diarization, numerous splits were created, which were partly caused by the various non-speaking voices and noises in the interviews. However, given the rich pre-existing transcripts with timestamps and the audio outputs from whisper with timestamps, we were able to overlay the silences detected with the nearest gap in the audio and get utterances that were coherent to the natural speech flow, and aligned with the transcriptions.

The output of the diarization was used together with the utterances to generate the relevant audio files for input in the emotion modeling. Since finally a subset of 10 interviews was selected for qualitative interpretation, all audio and text utterances were checked and corrected manually. The manual check served as an informal verification of the utterances and was performed by two researchers at the hackathon by cross referencing the automated speaker turns and utterances against the original recordings and generated text transcripts. In the few instances where the diarization failed to detect the transition or, if an utterance could be split again, the timestamps and transcript breaks were adjusted manually to ensure perfect alignment. Given the sample size, this manual process was feasible; however, for larger sample sizes in the future, Human-LLM assisted verification could be used.

The emotion modeling for audio was accomplished with large pre-trained models, mainly provided by the Hugging Face model library. For speech emotion recognition (SER) modeling, we used the Wav2vec 2.0 model as implemented and evaluated by Wagner et al. (Wagner et al., 2023). Wav2vec 2.0 is a neural network model relying heavily on transformer architectures with 12 transformer layers. For our analysis, we decided to use the wav2vec model for both English and Polish interviews as the authors of the paper claim the validity of the model for languages other than English. This also kept our results invariant to possible bias introduced by using different models for English and Polish. Training a specific model for Polish analysis was attempted but we observed there to be a lack of labeled Polish data for valence and arousal detection from audio, effectively making the task infeasible given the project's time constraints.

### 3.3. Visual

The video analysis task was to determine the emotions expressed by the survivors and witnesses based on their facial expressions. Rather than replacing traditional text analysis, this visual approach serves as a tool to identify specific 'emotional spikes' that call for deeper qualitative investigation by the researcher. We do not argue for automating the process of understanding oral testimonies, rather to facilitate the researchers to gain insights into potential emotional moments.

In the video analysis pipeline, the videos were analysed frame by frame and not by segmenting them into utterances. While it was necessary for the text and audio inputs to be split into utterances to maintain contextual consistency, facial expression analysis benefited from the higher granularity provided by the individual frames in the video. This is because, firstly, emotions change within seconds; we are interested in these changes or spikes, which would otherwise be lost if averaged across an utterance. Secondly, most model input requirements detect the emotional state based on a single snapshot (i.e., one frame), unlike the text and audio models which require a temporal contextual window. Finally, emotional facial expressions do not necessarily stop when the speaker stops talking. This approach therefore, allows the detection of emotional shifts that may occur across utterances, within a single utterance or during periods of silence.

To process the videos, a first challenge was that they had different frame-per-second rates (30-40 FPS), while it would have been more accurate to extract one per second, we extracted one frame every 100th, which is equivalent to having one snapshot every 2.5-3.3 seconds. In a more dynamic type of video a tighter extraction rate would be advised, however in these videos the interviewee is conven-

[9] https://github.com/pyannote/pyannote-audio

[10] https://gist.github.com/rudolfbyker/8fc0d99ecadad0204813d97fee2c6c06

tionally situated at the centre of the frame and is recorded from a fixed angle. The Emonet [11] emotional detection model, was applied to identify the emotions expressed, and to quantify valence (degree of positivity or negativity) and arousal (level of emotional intensity or excitement). Emonet is a convolutional neural network (CNN) optimized for estimating valence and arousal levels from faces in naturalistic conditions (Toisoul et al., 2021). The model claims to estimate the valence and arousal in a given image with a small margin of error. The model was run on all the extracted frames of the selected interviews to generate frame level predictions for categorical emotion labels (e.g., happy, excited, anger, fear etc.) and valence and arousal scores. In order to maintain feasible processing times, these processes were assisted using a High Performance Computing environment.

To aim at a correct identification of key moments based on the emotion expression of the interviewees, it was important to minimize instances where the interviewer's faces might have been captured. While the camera was focused on the interviewee for the majority of the interview, we included only those frames that detected one face. However, it is still possible that there could be some noise in the dataset that detected the interviewer rather than the interviewee. This resulting dataset contains the emotional trajectory of the testimonies based on the videos. Differences in emotional display across gender and survivor/witness testimonies were analysed using R [12]. By identifying frames with high emotional expression, we can identify specific narratives in the testimony where the visual data provides an unique layer of context, complementing the text and audio workflows.

## 4. Excerpts from the Analysis

To illustrate the main results from the emotion analysis, we extract in this section key moments of two interviews from our dataset (Anuradha et al., 2026). These moments show agreement between two signals (Figure 2) and disagreement (Figure 3) . For layout purposes, the figures are displayed in the following page.

First we highlight an excerpt of the testimony given by Judith Balassa Zucker, born circa 1934 in Czechoslovakia (Figure 2 ) [13]. Zucker survived the Holocaust in her hometown of Krupina, hiding in the mountains with her family and other Jews towards the end of World War II. In the moment shown she recalls the arrival of German soldiers in later stages of the war. The emotional analysis identifies fear, something that a human reader could agree with and would most likely tag the same. Incorporating the emotional analysis overtime for the video, we see an overall agreement (while the frames give more frequent results than the longer utterances in the timeline).

The next excerpt (Figure 3) shows a disagreement between transcript and audiovisual analysis. A peak was detected by the audio and visual models, both concern valence that refers to the positivity of the emotion. In this interview Josefa Anasiewicz, born in 1914, gives testimony of a mass-shooting in her village [14]. When the head-shot from the moment we see the peak we see a smile, according to the transcript, she recalls Jewish Easter holidays and making bread in the immediacy of sad memories about a fire in her house. While the smile validates the result, neither the sentiment or emotional analysis from the transcript identified the fast emotional change.

## 5. Limitations of data and models

There were methodological limitations arising from the data, the hackathon setting and the applied models. Our goal in acknowledging them in detail is to signpost vistas for improving other-than-English and audiovisual models.

Concerning the transcript material was highly heterogeneous, requiring extensive pre-processing and manual standardization. Furthermore, some transformer-based models imposed input length constraints (approximately 500 tokens), which made processing the data at the level of individual utterances necessary. While the selected interviews contained responses short enough to be analysed without further segmentation, longer responses would require splitting, potentially fragmenting semantic and emotional context. While it was not among our goals to make comparisons, the use of different language-specific models for English and Polish sentiment analysis of transcripts limits direct comparability, as the models differ in architecture, label sets, and training data. Finally, as all models were pre-trained and not fine-tuned on interview-specific data, their predictions may not fully capture the nuances. This limitation is further illustrated by the Polish emotion model used in this study, which was reported by its authors to have suffered from corrupted weights in an earlier version, leading to largely random predictions. Although this issue was later corrected, it highlights the broader risk of relying on pre-trained models whose internal limitations or instabilities may not be immediately apparent. If one wishes for comparability across

[11] https://github.com/face-analysis/emonet

[12] https://www.r-project.org/

[13] https://collections.ushmm.org/search/catalog/irn511823

[14] https://collections.ushmm.org/search/catalog/irn507914

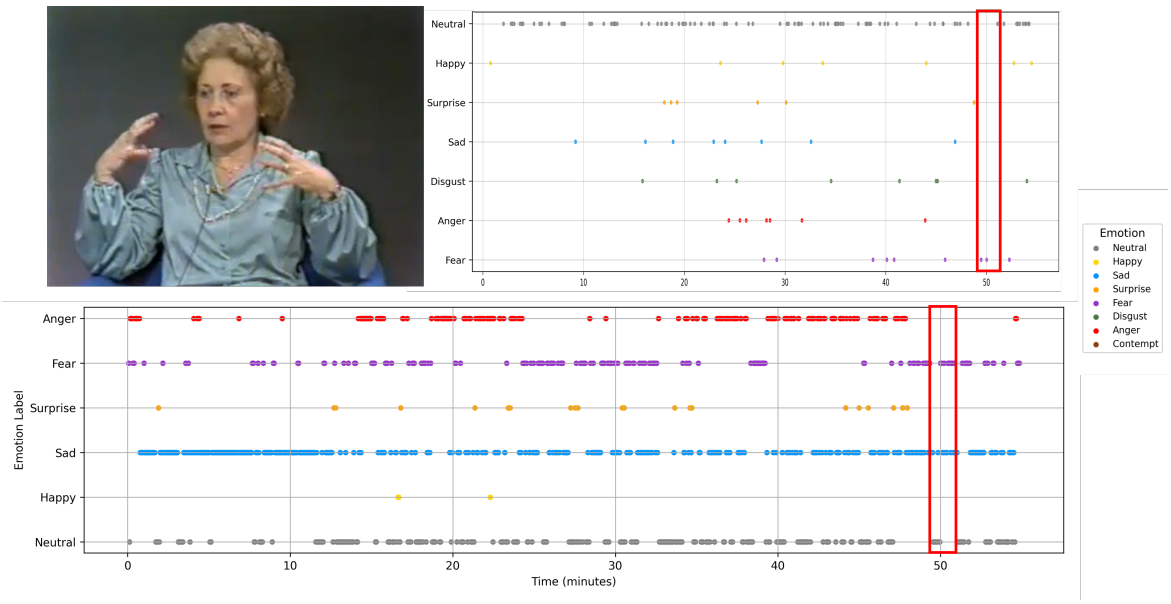


Figure 2: Emotional analysis over time from interview with Judith Zucker (emotion lables for transcript above, for video below): "About 10 o'clock, we got the news that Germans are coming. It was so cold [...] that you wouldn't send a dog out there. The wailing winds, it was unbelievable cold. **We hear the Germans are coming. We got to go.** So we go."

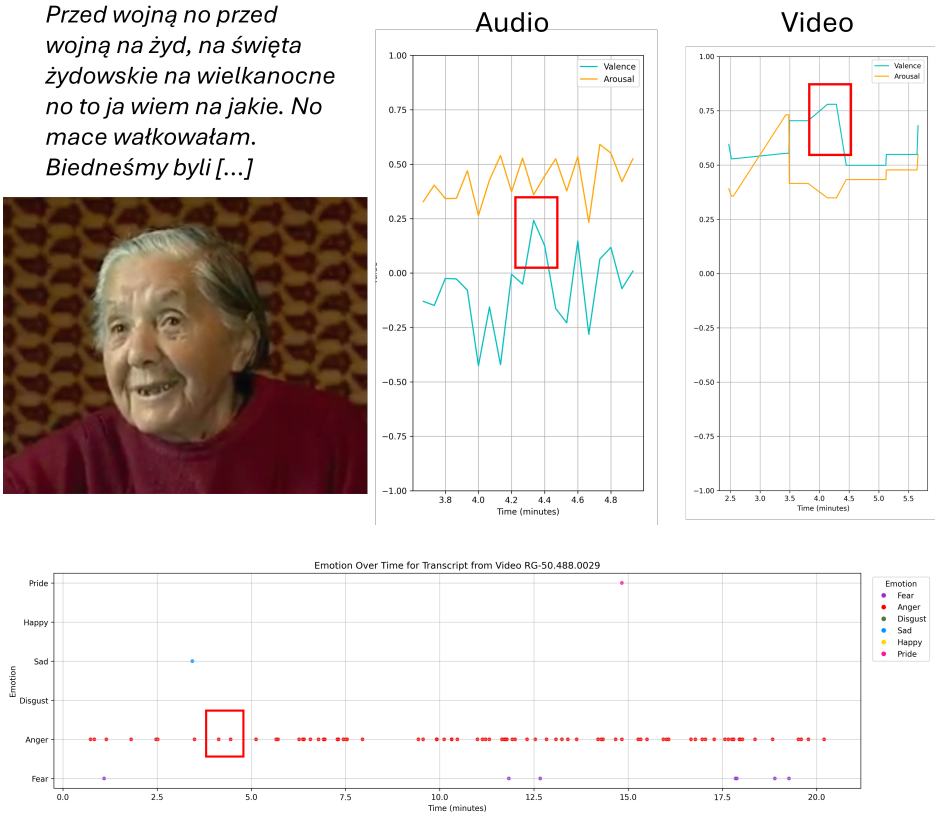


Figure 3: Emotional analysis from interview with Josefa Anasiewicz. Above the valence and arousal peak in the audio and video signals. Below, the emotional analysis based on the transcript, the positive sentiment is not detected. Translation: "Before the war for Jew [sic] for Jewish holidays, for Easter [sic], what do I know which one; I used to roll matzoh [bread]. We were poor [...]"

languages, it would be necessary to build and train a model of their own.

Concerning the audio workflow, a challenge of this approach is the limited reliability of the utterances. As an automatic process, the splitting of the speech might introduce errors. An additional concern remains whether the silences in the speech (context window around silences, breathing patterns, changes in voice etc.) are meaningful for analysis. On the other hand, even if the silent parts of speech contain relevant information about the emotions portrayed, the modeling we used for emotion detection would likely lack the capability for emotion recognition for silent audio. It is also important to note that the pre-trained models are mostly trained in English and with samples from younger adults. This introduces risks of age bias and unavailability of models relevant to different languages. Finally, as noted in the workflow, the diarization and silence detection models were sensitive to the acoustic environment of the interviews. For example, issues such as tape changes, background noise, or long, meaningful silences required manual oversight to ensure that segments were not erroneously discarded.

During the visual analysis, the primary limitation was the age bias in emotion detection. Pretrained facial emotional recognition models, including Emonet, are predominantly trained on datasets of younger individuals. This bias in training data is especially visible in expressions erroneously classified as sadness or anger. Regardless of the speaker's true emotion, changes in physiological features such as drooping eyelids and marionette lines led to these errors. In addition, the environment of the video recordings themselves introduce noise; for instance, camera zooms, lighting changes, changes of video tapes etc., which affect the model's ability to map facial landmarks coherently. Similarly, during instances where the interviewer's face may appear in the frame, the model will predict emotional labels for the interviewer, requiring data cleaning.

Finally, when looking at the textual, audio and visual models, which were developed on different training datasets and architectures, an unbiased comparison between these modalities is impossible. In addition, pre-trained models on the English language performed better when compared to other languages such as Polish or French, highlighting the need for improvement in language-specific or multilingual emotion detection models. It is clear that there is a need to test and develop multilingual and multimodal models, specifically fine-tuned for different use cases, where the training inputs and requirements more closely match the characteristics of Holocaust testimonies.

## 6. Conclusion, applications and vistas for research

An important motivating factor for focusing on emotions considering their verbal and non-verbal expressions, was tied to a fundamental value of having recorded and preserved Holocaust testimony as full-length video interviews: a great deal of emotional communication occurs in the audiovisual dimensions of testimony. The public online catalogue of USHMM alone contains over 14.000 digitized recorded interviews, of which less than a tenth include a transcript. While improvements in ASR, ML and NLP are enabling to ever more accurately transform and translate speech-to-text, the analysis of Holocaust testimony still relies on textual sources. The main contribution of this recipe book is to offer a proof of concept for multimodal large-scale analysis incorporating rich non-verbal emotional information left out of transcripts. Developing multimodal models, or replicating this in a more long-term project that allows to fine-tunine those tried during this project, can help identifying in one quick glance emotionally charged moments from long recordings, and down the line generating labels or metadata to enrich transcripts.

Another aim of this project, was to test a multimodal and partly multilingual approach to Holocaust Testimony. The time invested in preparation of data did not allow to produce a reliable study, but we succeeded in identifying the suitability of models or cues, such as valence and arousal in video and audio being valid indicators of emotionally charged moments. Furthermore we could recognise in models important flaws, such as the deficiency of audio models in other-than-English languages, or the bias of audio and visual models trained with samples of younger population and interpreting overly negative facial expressions of ageing population. Hence, we were able to sign-post vistas to further develop audio-visual models. This points at a related contribution, the preparation of the used dataset which contains rich multilingual interviews that with further refinement can become a benchmark on which existing or new models could be fine tuned or trained (Anuradha et al., 2026).

In zooming into in the workflow that we followed, one last contribution is the itemization of steps that need to be made in order to prepare non-machine readable transcripts and audiovisual recordings for machine-aided analysis. The recipes in this paper can be guide archives that hold oral history interviews and researchers working with audiovisual material, to turn this rich but heterogeneous data into machine readable datasets and benchmarks that allow further development of emotion detection models in other-than-English and ageing populations. Finally, we hope this inspires newcomers to

emotional approaches to Holocaust testimonies or digital humanities researchers to pay attention to the rich emotional information contained in audiovisual recordings.

## 7. Acknowledgements

We want to acknowledge the valuable work of other students of the Oral History group at the DHH2025: Visa Alamännistö, Haruka Buss, Joonatan Huang, Xiaoyue Wang, Rahel Albicker, Muhammad Hassan Qadeer Butt and Ellen Yang; as well as other instructors: Saara Kekki, Yu Wu and particularly Edyta Gawron. We also want to thank the US Holocaust Memorial Museum, Isuri Anuradha and Martin Wayne who facilitated an unrestricted access to the interviews.

## 8. Bibliographical References

Isuri Anuradha, Le An Ha, Ruslan Mitkov, and Vinita Nahar. 2023. Evaluating of Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 117–123, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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*, HIP '23, pages 85–90, New York, NY, USA. Association for Computing Machinery.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Liudmila Bukreeva, Daria Guseva, Mikhail Dolgushin, Vera Evdokimova, and Vasilisa Obotnina. 2023. Emotional Speech Recognition of Holocaust Survivors with Deep Neural Network Models for Russian Language. In Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal, and S. R. Mahadeva Prasanna, editors, *Speech and Computer*, volume 14338, pages 68–76. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Nicole Eustace, Eugenia Lean, Julie Livingston, Jan Plamper, William M. Reddy, and Barbara H. Rosenwein. 2012. *AHR* Conversation: The Historical Study of Emotions. *The American Historical Review*, 117(5):1487–1531.

Maxim Ifergan, Renana Keydar, Omri Abend, and Amit Pinchevski. 2024. Identifying Narrative Patterns and Outliers in Holocaust Testimonies Using Topic Modeling. ArXiv:2405.02650 [cs].

Joonatan Laato, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2025. Extracting Social Connections from Finnish Karelian Refugee Interviews Using LLMs. ArXiv:2502.13566 [cs].

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10645–10659.

Zoë Waxman. 2012. Chapter 7. Transcending History? Methodological Problems in Holocaust Testimony. In Dan Stone, editor, *The Holocaust and Historical Methodology*, pages 143–157. Berghahn Books.

## 9. Data Sources

Isuri Anuradha, Inés Matres, and Yu Wu. 2026. Multilingual dataset of interviews with survivors and witnesses of the holocaust https://doi.org/10.5281/zenodo.18701345.

# Modeling the Language of Holocaust Survivors' Testimony with Domain-Adapted Transformers

**Christopher Brückner**[1], **Jan Lehečka**[2], **Jan Švec**[2], **Pavel Pecina**[1]
[1]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
[2]University of West Bohemia, Department of Cybernetics
Pilsen, Czech Republic
{bruckner,pecina}@ufal.mff.cuni.cz, {honzas,jlehecka}@kky.zcu.cz

**Abstract**

Documents related to the Holocaust increasingly move into the focus of Natural Language Processing research, including the digitization of written text, the automatic transcription of oral archives, and interpretive downstream tasks such as Named Entity Recognition. However, most modern language models are trained primarily on modern text, and thus struggle with historical language, historical entities, and domain-specific terminology. Furthermore, transcribed speech introduces challenges such as transcription errors, noise, filler words, and dialectal speech not often contained in textual datasets. We present XLM-RoBERTa-malach, a text encoder domain-adapted to oral testimonies of Holocaust survivors in seven languages. In addition to descriptions of the data acquisition via Automatic Speech Recognition, data augmentation via Machine Translation, and the continued pretraining of a state-of-the-art multilingual transformer, we evaluate the domain-adapted model on the Named Entity Recognition task. Experiments on this task show superior performance over the general-domain transformer in a multilingual domain-specific setting, including languages not seen during the domain adaptation.

**Keywords:** Language Modeling, Domain Adaptation, Holocaust Testimony, Speech Recognition

## 1. Introduction

With the advent of Transformers, Natural Language Processing (NLP) has made significant improvements in general-domain and domain-specific settings. However, most language models have been trained on modern text and do not generalize well to historical documents, which come with additional challenges, such as orthographic reforms, entity drift, noisy inputs, and a lack of resources (Ehrmann et al., 2023).

While language models domain-adapted to 18th to 20th century text have been shown to outperform general-domain models in the Named Entity Recognition (NER) downstream task in documents from the same time period (Schweter et al., 2022), and NER in Holocaust survivors' testimonies has become of interest (Dermentzi and Scheithauer, 2024), no language model adapted specifically to mid-20th century languages and Holocaust-related terminology does yet exist. With the increasing number of digitized documents, such a model becomes an interesting candidate to assist with the processing of large archives, at potentially better quality than the currently available domain-agnostic solutions.

An additional challenge introduced in this domain comes from the fact that many testimonies, especially in non-English languages, exist only in oral form, which is very different from written documentation and often requires Automatic Speech Recognition (ASR) technologies to make them accessible for further processing (Lehečka et al., 2023).

In this paper, we present XLM-RoBERTa-malach, a multilingual Transformer based on the XLM-RoBERTa architecture (Conneau et al., 2020), domain-adapted to Oral Holocaust Testimony. It is named after the Hebrew word for "angel", or Multilingual Access to Large Spoken ArCHives. The following sections describe the foundations of domain adaptation and Holocaust-related NLP, the acquisition and augmentation of training data via automatic speech recognition and machine translation, the training process, and finally, NER experiments in testimonies showing the outperformance of general-domain models.

While we are not able to publish the dataset itself due to the licensing of the source data, the domain-adapted model is available on Hugging Face[1] under the MIT license. The source code of the NER experiments is available on GitHub[2].

## 2. Related Work

### 2.1. Domain-Specific Language Models

Language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are powerful

[1] https://huggingface.co/ufal/xlm-roberta-malach
[2] https://github.com/chbridges/malach-ner



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 74–83
11 May 2026. 

and, even in the age of Large Language Models, resource-friendly text encoders popular for tasks that do not require language generation. Pretrained on large amounts of text, fine-tuning them on downstream tasks such as Named Entity Recognition (NER) allows them not only to solve the given tasks, but also to adapt their parameters to the language and domain present in the fine-tuning dataset. While the underlying language models have typically been pretrained on domain-agnostic text, it has been shown that models can achieve even better results if they have already been pretrained on text in the same domain as they will eventually be fine-tuned on.

For example, language models are typically pretrained primarily on modern text and do not generalize well to historical documents. Such texts are subject to language change in various dimensions, including changing spelling conventions, words losing or gaining additional meanings, and locations changing their names (Ehrmann et al., 2023). Within the HIPE-2022 shared task on NER in historical documents (Ehrmann et al., 2022), Historical Multilingual BERT (Schweter et al., 2022) has been pretrained from scratch on 19th- and 20th-century newspapers and outperformed other participating systems in multiple languages.

A different approach to domain adaptation is the continued pretraining, where an already pretrained model is further trained on the same training objective, but on new data. For example, Gururangan et al. (2020) continued the Masked Language Modeling pretraining of RoBERTa in different domains (biomedicine, computer science, news, and Amazon reviews) and achieved significantly improved results in domain-specific tasks such as relation and topic classification. XLM-RoBERTa, a highly multilingual transformer model still used in state-of-the-art NER architectures (Straková and Straka, 2025), has been additionally pretrained on parliamentary proceedings, outperforming the original general-domain model in sentiment analysis in the legal domain (Mochtak et al., 2024).

### 2.2. NLP for Testimonial Data

While Named Entity Recognition is a well-established task, its applications in speech are still limited, and language models not traditionally trained on spoken language struggle with this task (Caubrière et al., 2020; Yu et al., 2025). This poses a problem in the Holocaust domain, as large amounts of its documentation exist only in oral testimony. These testimonies are often not manually transcribed, which has led to the emergence and reliance on domain-specific Automatic Speech Recognition technologies (Lehečka et al., 2023).

First steps in NER in testimonial data have been taken by Anuradha Nanomi Arachchige et al. (2023), who labeled English testimonies from the United States Holocaust Memorial Museum[3], Fortunoff Video Archive[4], and the Wiener Holocaust Library[5] with a highly domain-specific and granular entity type ontology. In addition to the common entity types Person, Location, and Organization, it distinguishes between different spatial entities (Location, Geopolitical Entity, Ghetto, Camp, Street, River) and temporal entities (Time, Date, Event), as well as Military organizations, Warships, Spousal Relationships, and Languages. These distinctions can lead to ambiguities, as the types of toponyms can be context-dependent, e.g., "Czestochowa" can refer either to a city (Location) or a Camp. Baseline experiments show that general-domain language models outperform Historical Multilingual BERT (Schweter et al., 2022): While the Holocaust undoubtedly belongs into the historical domain, many testimonies have been recorded at the end of the 20th century, which most historical data predates.

The same testimonies have recently served as training data for the domain adaptation of an English language model, HoloBERT, which outperforms the general-domain BERT on some, but not all, entity types in this granular ontology (Anuradha et al., 2025).

A more recent, multilingual NER dataset in this domain is EHRI-NER (Dermentzi and Scheithauer, 2024), which is based on EHRI Online Editions[6] in 9 languages. EHRI-NER uses a smaller, but still domain-specific entity ontology, extending the standardized CoNLL format (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) with Dates, Camps, and Ghettos. However, since these Online Editions were not originally annotated for Named Entity Recognition, but for Entity Linking, EHRI-NER contains non-standard annotations. For instance, "father" is not a named entity, but can be tagged as Person, since the word can be linked to a specific entity within the context of the testimony it appears in. EHRI-NER has been published, split into languages[7] or into multilingual training/validation/test data[8].

## 3. Training Data

This section describes the acquisition and augmentation of a training corpus for the domain adaptation

[3] https://www.ushmm.org
[4] https://fortunoff.library.yale.edu
[5] https://www.testifyingtothetruth.co.uk
[6] https://www.ehri-project.eu/ehri-online-editions/
[7] https://github.com/EHRI/EHRI-NER
[8] https://huggingface.co/datasets/ehri-ner/ehri-ner-all

of a new multilingual language model.

### 3.1. Source Corpus

The Visual History Archive[9], maintained by the USC Shoah Foundation, is the largest existing archive of Holocaust testimonies. It comprises more than 55,000 video interviews with survivors and covers more than 30 languages. Founded in 1994, these interviews are rather recent and thus not subject to orthographic reforms or other significant evolutions of written language. However, beside terminology related to World War 2, Nazi Persecution, and Jewish Identity in these oral testimonies, survivors often refer to places by historical names that have fallen out of regular use.

USC has provided the video interviews in six languages, the distribution of which is shown in Table 1. In total, these 33,902 testimonies amount to more than 60,000 hours of MP4 files at a size of 27 TB.

| Language | Testimonies | Fraction |
|---|---|---|
| English | 28,457 | 83.94% |
| Polish | 1,521 | 4.49% |
| Hungarian | 1,369 | 4.04% |
| Dutch | 1,077 | 3.18% |
| German | 917 | 2.70% |
| Czech | 561 | 1.65% |
| Total | 33,902 | 100.00% |

Table 1: The distribution of languages in the available testimonies from the Visual History Archive.

### 3.2. Data Preparation

#### 3.2.1. Automatic Speech Recognition

The audio tracks from all 27 TB of MP4s have been extracted with FFmpeg to single-channel MP3s at a sampling rate of 16,000 Hz and a variable bitrate of 190-250 kbps. While this encoding is rather lossy, its quality is sufficient for automatic speech recognition (ASR), and it reduces the size of the data tremendously to only 1.55 TB, which helps significantly with data transfer and processing.

The ASR processing was performed using a self-hosted, containerized version of the UWebASR service (Švec et al., 2025)[10], deployed using Singularity containers in the MetaCentrum HPC infrastructure. The system supports the languages relevant to the Holocaust testimonies and utilizes two primary architectures: Wav2Vec 2.0 (Baevski et al., 2020) and the more recent Zipformer architecture (Yao et al., 2023). To efficiently manage long audio inputs typical of Holocaust testimonies, the engine employs a sliding window approach during transcription.

The training methodologies and decoding strategies differed between the two architectures. The Wav2Vec 2.0 models for Czech, Slovak, German, and English were first pre-trained on large-scale, unlabeled speech datasets (e.g., 80,000 hours for Czech and German, 20,000 hours for Slovak). These base models were then fine-tuned in a two-phase process: first on general-domain speech and subsequently on oral-history-style recordings (Lehečka et al., 2023). For Dutch, we utilized a publicly available Wav2Vec 2.0 model fine-tuned on the Corpus Gesproken Nederlands (CGN) dataset[11]. For decoding, the Wav2Vec 2.0 models employ Connectionist Temporal Classification (CTC) (Graves et al., 2006) over graphemes, integrated with an external language model to enhance linguistic context.

In contrast, the Zipformer models were trained using supervised learning directly on labeled datasets (see Table 2 for data sizes) and employ greedy CTC decoding over subword units (SentencePiece). The Zipformer architecture utilizes a modified Transformer encoder operating at multiple lower frame rates, enabling faster decoding and improved performance.

The architectures also cover different sets of languages. Wav2Vec 2.0 models were used for Czech, Slovak, German, English, and Dutch. The Zipformer architecture was applied to the same set (with the exception of Dutch) and further extended to include Polish and Hungarian. Table 2 shows the Word Error Rates (WER) for the languages matching our corpus across these architectures, compared against the Whisper-large-v3 baseline. The service supports flexible downstream processing by providing outputs in multiple formats, including plain text, WebVTT, JSON, and Transcriber XML.

Table 2 summarizes the labeled training data composition for each language, reporting both the total amount of supervised speech and the proportion originating from the oral-history domain. For nearly all languages, oral-history recordings constitute only a small fraction of the available labeled data, highlighting the severe scarcity of in-domain supervision and the resulting difficulty of the ASR task. Consequently, the evaluated models must rely heavily on cross-domain generalization rather than extensive domain-matched training. The only exception is German, for which substantially larger in-domain resources are available: manual annotations exist for approximately 900 German-language interviews, totaling nearly 2,000 hours, prepared by researchers from Freie Universität Berlin.[12]

[9] https://vha.usc.edu

[10] https://uwebasr.zcu.cz

[11] https://huggingface.co/GroNLP/wav2vec2-large-xlsr-53-ft-cgn

[12] Transcripts are publicly available at https://

| Language | Sup. Data [h] | Oral Hist. [h (%)] | Whisper v3 | Wav2Vec 2.0 | Zipformer |
|---|---|---|---|---|---|
| Czech | 6,000 | 106 (1.8%) | 19.1 | 8.5 | 7.1 |
| Slovak | 3,800 | 98 (2.6%) | 22.0 | 11.6 | 10.3 |
| German | 6,100 | 1,800 (30.0%) | 25.9 | 16.6 | 12.4 |
| English | 12,500 | 255 (2.0%) | 18.0 | 12.9 | 11.5 |
| Polish | 1,400 | 53 (3.9%) | 22.8 | – | 15.7 |
| Hungarian | 3,800 | 24 (0.6%) | 30.9 | – | 16.4 |

Table 2: Supervised training data size in hours [h] and Word Error Rates (WER) [%] of ASR architectures evaluated on oral history archives. Columns report the total amount of supervised data per language, the amount originating from the oral history domain (hours and proportion) and measured performance on the test split of the oral history dataset. A lower WER value indicates a better model. Whisper-large-v3 (Radford et al., 2022) serves as the general-domain baseline. We omit Dutch in this table as we had no labeled oral history data to fine-tune or evaluate the models for this language.

The ASR output is further processed through a post-processing pipeline for automatic punctuation and casing restoration. For English, German, Czech, and Slovak, we employed the approach described in Švec et al. (2021), using monolingual BERT-based predictors trained on CommonCrawl web text dumps to restore sentence boundaries, punctuation (full stop, comma, question mark), and proper casing. For the remaining languages (Polish, Hungarian, and Dutch), we utilized the `xlm-roberta_punctuation_fullstop_truecase` model[13] (Guhr et al., 2021). This step ensures that the resulting 3.1 GB of plain text is well-formatted for the subsequent domain adaptation of the corpus.

#### 3.2.2. Data Augmentation and Sampling

Due to the significantly skewed language distribution shown in Table 1, the produced text has been further machine-translated into all six present languages plus Danish to overcome data scarcity and create language-balanced training data. These translations have been created with MADLAD400-3B-MT[14], which has been shown to outperform comparable state-of-the-art models such as NLLB (NLLB Team et al., 2022) on mid- and high-resource languages at decreased inference time (Kudugunta et al., 2023; Lanz and Pecina, 2025). This includes the seven targeted languages.

The resulting balanced dataset has been tokenized with the XLM-RoBERTa tokenizer (Conneau et al., 2020), since this is the model architecture to be domain-adapted. Similar to the pretraining data of this architecture, the tokens have then been language-wise concatenated to single long strings and split into continuous, equally sized batches of 512 tokens. The final shorter batch in each language, which counts less than $10^{-7}$% of the total number of tokens, has been truncated.

Finally, 10% of batches per language are randomly sampled for a test set, and their tokens are statically masked with 15% probability. The remaining batches will be dynamically masked during the training, as in the Masked Language Model objective during the original XLM-RoBERTa pretraining.

Since 1/7 of the dataset is the output of different domain-specific and general-domain ASR models, and the remaining 6/7 are machine translations of the ASR output, this corpus can be considered 100% synthetic, although it is 100% a representation of real testimonies of Holocaust survivors. Due to possible errors and biases introduced by ASR artifacts and MT hallucinations, the corpus should not be used to train generative models, but only encoders for natural language understanding tasks such as NER.

### 3.3. Corpus Statistics

The created corpus has a total size of 4.9 billion tokens. The sizes per language and split are shown in Table 3. Although the same 33,902 testimonies are present in all seven languages, minus the truncated final batches, the numbers of tokens are not perfectly balanced across the languages. Instead, they represent how many tokens are required in each language to describe the same data.

| Language | Training | Test | Total |
|---|---|---|---|
| Czech | 637 | 71 | 708 |
| Danish | 612 | 68 | 680 |
| Dutch | 626 | 70 | 696 |
| English | 612 | 68 | 680 |
| German | 634 | 70 | 704 |
| Hungarian | 642 | 71 | 713 |
| Polish | 645 | 72 | 717 |
| Total | 4,407 | 490 | 4,897 |

Table 3: The VHA corpus size in **million [M]** tokens.

Since the test splits have been randomly sam-

`transcripts.vha.fu-berlin.de`.

[13] `https://huggingface.co/1-800-BAD-CODE/xlm-roberta_punctuation_fullstop_truecase`

[14] `https://huggingface.co/google/madlad400-3b-mt`

pled from each language individually, some data leakage has to be assumed: All test samples are likely seen during the training, albeit in different languages and with different positional embeddings. This can lead to an underestimated intrinsic perplexity of the language model. However, it does not affect extrinsic evaluation metrics on other datasets and downstream tasks.

In addition, we repurpose the full EHRI-NER dataset (Dermentzi and Scheithauer, 2024) to a second MLM test dataset by removing its annotations and applying the same tokenization and masking steps to it. We consider two variants of this test set: **EHRI-6** contains the six languages that overlap with our corpus (minus Danish), and **EHRI-9** additionally contains French, Slovak, and Yiddish, which our model does not see during the continued pre-training. The language distribution of this dataset is shown in Table 4. While significantly smaller in size (0.02%) and imbalanced, it is unbiased.

| | | Language | Tokens [k] |
|---|---|---|---|
| EHRI-9 (874) | EHRI-6 (713.5) | Czech | 195 |
| | | Dutch | 2.5 |
| | | English | 81 |
| | | German | 356 |
| | | Hungarian | 45 |
| | | Polish | 34 |
| | | French | 3.5 |
| | | Slovak | 6 |
| | | Yiddish | 151 |

Table 4: The size of the EHRI dataset in thousand [k] subword tokens. The total sizes of EHRI-6 and EHRI-9 are given in parentheses.

## 4. Domain Adaption

This section describes the domain adaption process and the internal evaluation of the resulting model on its pretraining objective.

### 4.1. Continued Pretraining Setup

We adapt the large-sized XLM-RoBERTa model[15] to the domain of Holocaust testimony by continuing its pretraining with the Masked Language Modeling objective on the produced VHA corpus. To do so, we replicate most of the hyperparameters reported Liu et al. (2019) for the original large-sized RoBERTa model[16], which uses Adam optimization (Kingma and Ba, 2017) with $\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon =$ 1e-6, and 0.01 weight decay, a peak learning rate of $4e^{-4}$ that is warmed up for the first 6% of steps and then decays linearly, and an effective batch size of 8192 for 500k steps.

In contrast, we use the improved AdamW optimizer (Loshchilov and Hutter, 2019), decrease the peak learning rate to prevent overfitting, and train the model on 4 L40 GPUs with per-device batch size 8, using 64 gradient accumulation steps for an effective batch size of 2048 for 34k steps. These smaller parameters are in line with other domain-adapted RoBERTa and XLM-RoBERTa models trained on smaller corpora (Gururangan et al., 2020; Mochtak et al., 2024).

While the test set is already pre-masked, the training set is dynamically masked with the same probability of 15%. We experiment with the learning rates 2e-5 and 1e-5 and evaluate the model's cross-entropy after every 10% of total steps on the VHA test data and on the EHRI dataset. The training lasts approximately 25 hours for 1 pass through the whole dataset, after which the model is rolled back to the checkpoint with the smallest VHA cross-entropy.

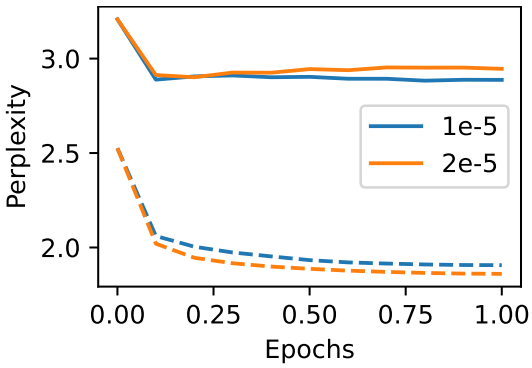


Figure 1: Perplexity of the model during 1 training epoch on the **VHA (dashed line)** and **EHRI-6 (solid line)** test datasets using peak learning rates of 1e-5 (blue) and 2e-5 (orange).

### 4.2. Resulting Model

Starting at 2.5257, the continued pretraining reduces the base-2 perplexity on the VHA data to 1.9064 (peak learning rate 1e-5) or 1.8603 (peak learning rate 2e-5). While Figure 1 indicates that neither model converges in 1 epoch, it also shows that the model with the greater learning rate starts overfitting to the automatically transcribed and translated VHA data 30% into the epoch, whereas the smaller learning rate keeps the perplexity on the EHRI-6 data stable at around 2.89. In both cases, the best checkpoint is the final checkpoint after 1 epoch.

Table 5 reports the perplexity of XLM-RoBERTa-large and both domain-adapted XLM-RoBERTa-

[15] https://huggingface.co/FacebookAI/xlm-roberta-large

[16] https://huggingface.co/FacebookAI/roberta-large

| Model | cs | de | en | fr* | hu | nl | pl | sk* | yi* |
|---|---|---|---|---|---|---|---|---|---|
| XLM-RoBERTa-large | 3.1553 | 3.4038 | 3.0588 | 2.0579 | 2.8928 | 2.9133 | 2.5284 | 2.6245 | **4.0217** |
| Malach 2e-5 | 2.8553 | 3.2277 | 2.9072 | *2.0966* | *2.9210* | *2.9404* | 2.4187 | *2.6592* | *5.3259* |
| Malach 1e-5 | **2.8023** | **3.1704** | **2.9022** | **2.0254** | **2.8285** | **2.8797** | **2.4003** | **2.5914** | *4.0910* |
| Support [k] | 195 | 356 | 81 | 3.5 | 45 | 2.5 | 34 | 6 | 151 |

Table 5: Base-2 perplexity of 3 models for 9 different languages in EHRI documents: The original XLM-RoBERTa-large checkpoint and ours, with two different peak learning rates. The last row shows the number of thousand [k] subwork tokens for each language. Languages marked with asterisks were not seen during the continued pretraining. Best scores are marked in **bold**, worse scores in *italics*.

malach variations in all 9 languages in the EHRI-9 corpus. Compared with the starting checkpoint, the peak learning rate of 2e-5 decreases the perplexity in only 4 of the seen languages, but increases it in Dutch, French, Hungarian, Slovak, and Yiddish. The increase from 4.0217 to 5.3259 in Yiddish, which is the only present language using a non-Latin alphabet, is particularly severe. In contrast, the peak learning rate of 1e-5 achieves the minimum perplexity in 8 languages, including 2 unseen ones. For Yiddish, the perplexity increases only to 4.0910. The greatest improvements can be observed in the Czech and German splits, which are also the greatest in size.

## 5. Named Entity Recognition

This section addresses the additional evaluation of the domain-adapted model on a domain-specific downstream task.

### 5.1. Experimental Setup

In addition to the internal evaluation of the domain-adapted models, we further evaluate them on the NER downstream task. The underlying dataset for this evaluation is the multilingual EHRI-NER (Dermentzi and Scheithauer, 2024), which has already served as an additional test dataset to measure the model perplexity on domain-specific data. EHRI-NER is annotated with the entity types Person, Organization, Location, Camp, Ghetto, and Date, in the languages Czech, Dutch, English, French, German, Hungarian, Polish, Slovak, and Yiddish.

`xlm-roberta-large-ehri-ner-all`[17], which has been fine-tuned on this dataset, serves as the baseline model. It is based on the same model as XLM-RoBERTa-malach, but without previous domain adaptation.

We replicate the training process of the baseline model on the published EHRI-NER training/validation/test splits: XLM-RoBERTa-malach is fine-tuned for 3 epochs using a learning rate of 3e-5, weight decay of 0.01, and a batch size of 16. The training is repeated 3 times using 3 different random seeds, so that not only the overall and tag-wise mean $F_1$ scores can be reported, but also their 95% confidence intervals.

[17] https://huggingface.co/ehri-ner/xlm-roberta-large-ehri-ner-all

### 5.2. Results

Mean $F_1$ scores and their 95% confidence intervals are shown in Table 6. Note that we were not able to exactly reproduce the results from Dermentzi and Scheithauer (2024) using their provided fine-tuned model and fixed test split, possibly due to differences in the processing and evaluation code.

The overall $F_1$ score does not increase significantly compared with the state of the art. Using a peak learning rate of 2e-5, it does not change at all; using a peak learning rate of 1e-5, it increases by 0.67% $F_1$ on average. More interesting are the differences per tag: While there is a slight decrease (less than 1%) for PER and LOC entities, the scores increase significantly (up to 5% on average) for the rarer, domain-specific CAMP and GHETTO entities.

While XLM-RoBERTa-malach pre-trained with a learning rate of 1e-5 tends to outperform the variant pre-trained with a learning rate of 2e-5 on the NER downstream task, its experimental results also come with increased variance, in particular with respect to ORG and GHETTO. Organizations appear to be generally difficult to predict in EHRI-NER, and ghetto examples are sparse. Furthermore, a typical error for all models is the confusion of camps, ghettos, and general locations.

In comparison, fine-tuning the English-centric domain-adapted HoloBERT (Anuradha et al., 2025) achieves an overall $F_1$ score of only 75%, with comparable scores only for the domain-specific entity types CAMP (72.00), GHETTO (82.00), and DATE (81.67). Its results are significantly worse for PER (77.00), ORG (56.67), and LOC (75.33).

## 6. Discussion

Overall, the model domain-adapted with a peak learning rate of 1e-5 processes domain-specific text better than the domain-agnostic model and the 2e-5 variant, in terms of internal metrics (Masked Language Modeling and perplexity) and external

| | EHRI | Malach 1e-5 | Malach 2e-5 |
|---|---|---|---|
| PER | **87.00** | 86.67 ± 1.43 | 85.67 ± 1.43 |
| ORG | 63.00 | 64.33 ± 6.25 | **64.33 ± 1.43** |
| LOC | **82.00** | 81.67 ± 1.43 | 81.67 ± 1.43 |
| CAMP | 70.00 | **75.00 ± 2.48** | 73.33 ± 2.87 |
| GHETTO | 80.00 | **85.00 ± 6.57** | 84.67 ± 1.43 |
| DATE | 84.00 | **85.00 ± 4.30** | 84.67 ± 3.79 |
| Overall | 81.00 | **81.67 ± 1.43** | 81.00 ± 0.00 |

Table 6: Mean micro $F_1$ scores (%) and their 95% confidence intervals on the EHRI-NER dataset. The compared models are `xlm-roberta-large-ehri-ner-all` (Dermentzi and Scheithauer, 2024) and our XLM-RoBERTa-malach, pre-trained with peak learning rates 1e-5 and 2e-5. Despite the variance of individual tags, the last model achieves the same overall score across 3 experiments.

metrics (Named Entity Recognition and $F_1$ scores). The perplexity decreases even for two languages unseen during the continued pretraining, which suggests that the model has learned genuine domain-specific representations, rather than simply memorising language patterns from the training data. However, the perplexity slightly increases for Yiddish, which is the only present language not using the Latin alphabet. Given the relevance of Yiddish in this domain, the increased perplexity is unfortunate, and additional data in Yiddish is required to tackle this issue. Such data can be generated via ASR (Marmor et al., 2025); however, machine translation from English to Yiddish is often of low quality (Kudugunta et al., 2023), and its suitability for data augmentation has to be further investigated.

In the NER task on multilingual written testimony, the performance noticeably increases on domain-specific entities, while the overall performance improves only slightly. This is because more general entities, such as people, occur much more frequently than camps and ghettos, which are of particular interest when extracting entities from testimonies. Slight degradations can be observed for PER and LOC entities. The latter one can be explained by the occasional ambiguity of LOC, CAMP, and GHETTO.

Although adapted exclusively to speech, produced by automatic speech recognition and machine translation, XLM-RoBERTA-malach is an interesting candidate for the processing of oral and written testimonies in multilingual settings. Its vastly improved performance in multilingual NER over HoloBERT (Anuradha et al., 2025) outlines the importance of multilingual pretraining in this domain. However, the model has not been evaluated on downstream tasks in speech data due to the limited availability of annotated corpora. In particular, the model has been trained on Danish, but no annotated domain-specific Danish text is yet available.

## 7. Conclusion and Future Work

We presented XLM-RoBERTa-malach, a variation of the large-sized multilingual XLM-RoBERTa model, domain-adapted to oral testimonies of Holocaust survivors. These testimonies have been produced via Automatic Speech Recognition of video testimonies in 6 languages from the Visual History Archive, the largest available archive of testimonies, and the resulting corpus has been further augmented via Machine Translation to add a seventh language, tackle data scarcity in six of the seven languages, and balance out the language proportions.

Although based on real testimonies, the resulting corpus can be considered synthetic, and both steps in the corpus creation can create errors and biases. Despite these issues, the continued pretraining on the Masked Language Modeling objective decreased the model perplexity not only on the speech-based training data, but also on written testimonies in 8 languages using the Latin alphabet: Czech, Dutch, English, French, German, Hungarian, Polish, and Slovak. Notably, French and Slovak were not seen during the domain adaptation, whereas the model has been additionally adapted to Danish. On the other hand, the model perplexity slightly increased on Yiddish, which is based on the Hebrew alphabet.

In the same languages, XLM-RoBERTa-malach outperforms its non-adapted variant XLM-RoBERTa-large on the Named Entity Recognition task in written testimonies. While it handles the frequent, general-domain entity type Person slightly worse, it exhibits significant improvements in the recognition of domain-specific entities, namely ghettos, camps, and numerical dates. Overall, the domain-adapted model appears to be an interesting baseline for NLP tasks in this domain. A very small learning rate, even smaller than the learning rate used during the NER fine-tuning (1e-5 < 3e-5), has proven to be beneficial in the domain adaptation of this model.

In the future, XLM-RoBERTa-malach should be evaluated in additional domain-specific downstream tasks, including NER in speech, such as the new MalachNER dataset (Brückner et al., 2026). The model itself can be improved in several ways:

- More language can be included in the domain adaptation corpus. For example, the used ASR system additionally supports Croatian and Serbian. Languages with non-Latin alphabets often spoken by survivors, e.g., Yiddish, Hebrew, Russian, and Ukrainian, can be added for better cross-lingual generalization. SOTA ASR models for Yiddish and Hebrew are available (Marmor et al., 2025), and further languages can be added via machine translation.

- In addition to speech transcripts, manual transcripts and written testimonies can be added to the training data to cover a larger variance of language and named entities.

- The possible data leakage in the model training, which affects the internal evaluation and convergence criteria, can be tackled by separating the testimonies in the training and test splits more clearly. I.e., no translations of training samples should appear in the test data, and vice versa.

Furthermore, the domain adaptation corpus can be used as a parallel corpus to train cross-lingual sentence embeddings (Reimers and Gurevych, 2019; Feng et al., 2022) for sentence similarity tasks, such as document retrieval and sequence classification.

## 8. Ethics Statement

Holocaust testimony is a sensitive domain and should always be handled with consideration. The automatic speech recognition and machine translation used to produce the training corpus may introduce errors that should not be present in published data related to the Holocaust. The domain-adapted model is an encoder-only model to be used for downstream tasks such as Named Entity Recognition, which mitigates the risk of reproducing biases often seen in causal language modeling, i.e., in autoregressive or diffusion models for text generation. However, this does not prevent the model entirely from misuse: We emphasize that results produced with XLM-RoBERTa-malach should still be carefully validated, e.g., before automatically processing archival material.

## 9. Limitations

The corpus used for the domain adaptation cannot be published, as the processed video data is licensed only for use within the project this research has been conducted in. The published domain-adapted model has only been evaluated on one downstream task, as, to our knowledge, no more annotated data in this domain is currently openly available.

## 10. Acknowledgements

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work described herein has also been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

This research was partially supported by Charles University, project GA UK No. 380126 and SVV project number 260 821.

## 11. Bibliographical References

Isuri Anuradha, Le An Ha, and Ruslan Mitkov. 2025. HoloBERT: Pre-trained transformer model for historical narratives. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 105–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián, and Pavel Pecina. 2026. From oral history to structured data: The MalachNER dataset. In *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, Palma de Mallorca, Spain. ELRA.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

pages 4514–4520, Marseille, France. European Language Resources Association.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180. CEUR-WS.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Interspeech 2023*, pages 201–205.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yanir Marmor, Yair Lifshitz, Yoad Snapir, and Kinneret Misgav. 2025. Building an accurate open-source hebrew asr system through crowdsourcing. In *Proc. Interspeech 2025*, pages 723–727.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*

*(LREC-COLING 2024)*, pages 16024–16036, Torino, Italia. ELRA and ICCL.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129. CEUR-WS.org.

Jana Straková and Milan Straka. 2025. NameTag 3: A tool and a service for multilingual/multitagset NER. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–39, Vienna, Austria. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

Jiawei Yu, Xiang Geng, Yuang Li, Mengxin Ren, Wei Tang, Jiahuan Li, Zhibin Lan, Min Zhang, Hao Yang, Shujian Huang, and Jinsong Su. 2025. "i've heard of you!": Generate spoken named entity recognition data for unseen entities. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jan Švec, Jan Lehečka, and Pavel Ircing. 2025. Current State of the UWebASR - Web-Based ASR Service for Czech, Slovak, German, and English. In *CLARIN Annual Conference Proceedings 2025*, page 95.

Jan Švec, Jan Lehečka, Luboš Šmídl, and Pavel Ircing. 2021. Transformer-Based Automatic Punctuation Prediction and Word Casing Reconstruction of the ASR Output. In *Text, Speech, and Dialogue. TSD 2021*, pages 86–94.

# Evaluating Automatic Speech Recognition for Holocaust Testimonies: A Large-Scale Analysis of Whisper Performance on the Fortunoff Video Archive

**William J.B. Mattingly, Christy Bailey-Tomecek**

Yale University
New Haven, CT, United States
william.mattingly@yale.edu, christy.tomecek@yale.edu

**Abstract**

Holocaust testimonies are key primary sources documenting survivors' experiences, yet many remain inaccessible due to the labor-intensive nature of manual transcription. This paper presents a comprehensive evaluation of OpenAI's Whisper automatic speech recognition (ASR) system on 1,847 testimonies from the Fortunoff Video Archive for Holocaust Testimonies at Yale University. We assess transcription quality across multiple languages including English, French, German, Hebrew, Yiddish, Ladino, Slovak, and American Sign Language (with English voice-over), using human-reviewed captions as ground truth. Our analysis reveals a mean Word Error Rate (WER) of 15.28%, with 90.9% of testimonies achieving "Fair" or better quality (WER ≤25%). We identify systematic error patterns including challenges with disfluencies, interrupted speech, and language-specific orthographic conventions, particularly in Ladino, where Whisper's normalization to modern Spanish orthography creates systematic divergences from traditional Judeo-Spanish spelling. For Hebrew and Yiddish, we evaluate specialized models from ivrit-ai and find promising results for heritage language preservation. Our findings demonstrate that current ASR technology can substantially accelerate Holocaust testimony transcription while highlighting the need for domain-specific fine-tuning and post-processing for optimal results.

**Keywords:** Holocaust testimonies, automatic speech recognition, Whisper, oral history, Ladino, Yiddish, Hebrew, digital humanities

## 1. Introduction

The Fortunoff Video Archive for Holocaust Testimonies at Yale University holds over 4,500 testimonies comprising more than 12,000 hours of recorded interviews in 20 languages. Testimonies serve as both a research source and as a memorial to survivors and victims of the Holocaust. In both contexts, the archive has a responsibility to make testimonies accessible, and transcripts are an important piece of this.

Creating transcripts has many challenges. Historically, transcripts required cultural heritage institutions to manually transcribe audiovisual recordings with their own staff, or to contract out to vendors that utilize audio speech recognition (ASR). Manual transcription takes dozens of hours, and vendor costs, whether it involves human review, reach into the thousands of dollars (Bailey-Tomecek, 2025; Rodriguez & Brown, 2023). As a result, many institutions such as the Fortunoff Archive have historically been unable to transcribe their materials at scale. This landscape has changed with the advent of newer open-source ASR applications that use AI speech-to-text like OpenAI's Whisper.

Cultural heritage institutions have been exploring the use of Whisper because of its general accuracy and lower cost of business. (Rodriguez & Brown, 2023). However, Whisper's accuracy is dependent on features of the recording itself. (Graham & Roll, 2024; Lynema & Dunn, 2025). Holocaust testimonies have unique characteristics that can potentially impact Whisper's results.

This paper evaluates Whisper's efficacy with Holocaust testimonies from the Fortunoff Archive. We compared raw transcripts from Whisper to human-edited ground truth ones for 1,847 testimonies in eight languages. Along with evaluating Whisper's general capabilities, we examined the specific needs and challenges of Hebrew, Yiddish, and Ladino testimonies in the corpus, including using specialized models for Hebrew and Yiddish from the ivrit-ai project. Based on these results, we recommend a workflow for creating and editing ASR transcripts for Holocaust testimonies.

## 2. Related Work

### 2.1 Automatic Speech Recognition for Oral History

Complete transcription has long been a goal for many cultural heritage institutions, to provide better accessibility, as well as enhance searchability (Rao et al., 2025). Additionally, transcripts can provide opportunities for digital humanities scholars for research and analysis, such as the Fortunoff Archive's *Let Them Speak: In Search of the Drowned*[1] project and Critical Editions series.[2] Institutions have tried many vendor-based audio speech recognition (ASR) products, including 3play, Trint, Rev, Amazon Transcribe, Google speech-to-text, Kaldi,

[1] http://lts.fortunoff.library.yale.edu/

[2] https://editions.fortunoff.library.yale.edu/



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 84–92

Microsoft Stream, and Sonix (Bailey-Tomecek, 2025; Dunn et al., 2024; Lundgard, 2024; Lynema & Dunn, 2025; Myntti & Steed, 2019; Rao et al., 2025, Rodriguez & Brown, 2023). The efficacy of these services varied and were dependent on the quality of the source recordings as much as the service itself. It was also impacted by language used. (Bailey-Tomecek, 2025; Lundgard, 2024, Myntti & Steed, 2019; Rodriguez & Brown, 2023).

The emergence of large-scale pretrained models has transformed the landscape. Whisper was trained on 680,000 hours of multilingual audio and demonstrates strong zero-shot performance across languages and domains (Radford et al., 2022). Institutions such as Emory University have reported as much as a 35% decrease in labor time for correcting Whisper output versus fully human-edited transcripts (Rao et al., 2025). For specialized applications, fine-tuned variants have shown substantial improvements. The ivrit-ai project achieved state-of-the-art results for Hebrew and Yiddish ASR through crowdsourced training data collection (Marmor et al., 2025).

However, Whisper's accuracy has been demonstrated to be dependent on the quality of the recording itself; the volume, gender, speech rate and accent of the speaker; and the presence of background noise or music (Graham & Roll, 2024; Lynema & Dunn, 2025). Depending on the type of recording, institutions like Indiana University reported word error rates between 1-8% for educational recordings, which utilize a professionally recorded English language speaker, and 20-40% for field recordings, which were done in less-than-ideal conditions with speakers with non-standard accents. (Lynema & Dunn, 2025). Similarly, Emory University reported word error rates between 5.24% for educational recordings to 24.01% for variety television shows that contain musical performances among other features. (Rao et al., 2025).

Historically, a drawback of Whisper for oral history collections such as the Fortunoff Archive is that it is not necessarily trained on corpora that have the speech qualities of oral histories (e.g. accents, disfluencies, halting speech), nor have domain-specific information (e.g. placenames in Europe that may not be commonly referred to in an English-language corpus). Researchers at University of West Bohemia Pilsen working with a similar collection of testimonies held by the USC Shoah Foundation's Visual History Archive sought to bridge that gap using a fine-tuned, monolingual Wav2Vec that trained from multiple corpora and utilized text from the Common Crawl project to build a robust vocabulary that would contain less common entities for Czech, English, and German language materials. The results were more accurate than transcripts utilizing XLS-R and Whisper models. (Lehečka et al., 2023). However, Whisper continues to improve in accuracy and is attractive to institutions due to its wide availability and ease of use.

After reviewing available literature and feedback, the majority consensus is that Whisper is accurate enough for at least topical research purposes, if not for full accessibility, and institutions consider it a realistic solution for large scale transcription of collection materials (Dunn et al., 2024; Harbert, 2025; Lundgard, 2024; Lynema & Dunn, 2025; Rao et al., 2025, Rodriguez & Brown, 2023).

## 2.2 Challenges in Testimony Transcription

Holocaust testimonies from the Fortunoff Archive provide significant challenges for modern ASR platforms. These include:

- **Condition of the original carriers and associated impact on digitization.** Most of the testimonies were recorded on magnetic tape formats. Magnetic tape has a limited lifespan, on average lasting 15 years. The tape degradation will impact audio quality in digitization
- **Surrounding environment during the recordings.** Many testimonies were recorded in non-studio environments, both indoors and outdoors, with background noise, and may have lacked professional audio recording equipment
- **Elderly speakers.** Most testimonies were recorded when survivors were in their 60s-90s, with age-related voice characteristics
- **Emotional speech.** Testimonies frequently include crying, pauses, and voice breaking during traumatic recollections
- **Language switching:** Survivors often mix languages, e.g. they may use the language(s) spoken at the time of the recalled event rather than the primary language of the interview
- **Non-native accents for English and Hebrew testimonies:** Many survivors speak English and Hebrew as second or third languages, with accents and pronunciations that are not standard for a native speaker
- **Disfluencies:** As testimonies are unscripted, survivors will pause, stutter, repeat words multiple times in a row, and use fillers (um, eh, uh)
- **Less well-known locations and context-specific jargon:** Names of camps, geographic locations, and institutions are less likely to be recognized by ASR. Additionally, survivors may use ghetto and camp specific jargon.

# 3. Dataset

## 3.1 The Fortunoff Video Archive

The Fortunoff Video Archive for Holocaust Testimonies holds over 4,500 testimonies recorded between 1979-2024. The testimonies were recorded in 20 languages throughout North America, South America, Europe, and Israel. Testimonies range from 30 minutes to over 40 hours in length, with an average length of one and a half hours. Interviews were unscripted and unguided by the interviewers except for occasional, clarifying questions based on the survivor's discussion.

For this study, we analyzed a subset of 1,847 testimonies for which both Whisper-generated transcripts and human-reviewed captions were available. The distribution by primary language is shown in Table 1.

| Language | Code | Count | Percentage |
|---|---|---|---|
| English | eng | 1,699 | 91.9% |
| Unknown | -- | 71 | 3.8% |
| German | ger | 35 | 1.9% |
| French | fre | 25 | 1.4% |
| Hebrew | heb | 5 | 0.3% |
| Ladino | lad | 5 | 0.3% |
| Yiddish | yid | 4 | 0.2% |
| American Sign Language | sgn | 2 | 0.1% |
| Slovak | slo | 1 | 0.05% |
| Total | -- | 1,847 | 100% |

Table 1: Testimony distribution by language

## 3.2 Ground Truth Captions

Most ground truth captions were created by several vendor products and later corrected by archive staff. Vendors either used a hybrid of ASR with human editors employed by the vendor or purely ASR. These include 3play for English (hybrid), Trint for English, German, French, and Slovak (ASR only), Verbit for Hebrew (hybrid), and Sonix for all languages except Yiddish and Ladino. (ASR only). Yiddish and Ladino captions were created manually by scholars associated with archive projects. Yiddish transcripts were created in the Elan editor, which includes time synchronization while Ladino transcripts were transcribed without time synchronization and later aligned using Sonix's forced alignment tool with the Spanish model.

The archive's style guide includes the following:

- Diarization, either with the speaker names, or general labels like "interviewer" or "subject"
- Verbatim transcription, including disfluencies (e.g. "uh", "um"). Grammar and phrasing are not normalized or corrected as the archive prioritizes fidelity to the speaker. Mispronunciations are not documented in the text
- Notation of interrupted speech or false starts with em-dashes (e.g., "I was—")
- Bracketed annotations for non-verbal sounds using all-caps (e.g. [LAUGHS], [CRYING])
- Bracketed annotations for unclear or inaudible speech using all-caps (e.g. [INAUDIBLE])
- Bracketed annotations for words and phrases that the transcriptionist thinks are being used, but not completely certain of, using question marks directly after the opening bracket and directly before the closing bracket (e.g. "[? He said ?]")

Ground truth captions may have errors that contribute to false error rates. These include:

- Time misalignment in the vendor's editing platform. Most of the editing platforms used by the archive's vendors do not show any sort of timestamping beyond line break/paragraph levels. If more than a couple of words need to be edited, it can accidentally misalign individual syllables or even pin an entire sentence to one timestamp
- Time misalignment by manual transcriptionists. While Elan provides time-synchronization, some amount of misalignment still can happen.
- Choices by human editors. Some editors corrected grammar when asked not to, or did not transcribe sections of audio, e.g. conversation with crew members. These were corrected by archive staff whenever found, but may have been missed in other transcripts

Ladino transcripts present additional challenges as ground truth captions due to the language's Romanized orthography not being fully standardized. (Kohen & Kohen-Gordon, 2000) Choices by one transcriptionist may not be the same as another.

## 3.3 ASR System Configuration

We employed Whisper using the faster-whisper implementation with the large-v3-turbo model variant for optimal speed-accuracy tradeoffs. For Hebrew and Yiddish testimonies, we utilized specialized models from the ivrit-ai project:

- Hebrew: ivrit-ai/whisper-large-v3-turbo-ct2

- Yiddish: ivrit-ai/yi-whisper-large-v3-turbo-ct2

These models were fine-tuned on crowdsourced Hebrew and Yiddish audio data, achieving substantially lower word error rates than the base Whisper model on these languages.

As Ladino lacks a dedicated model, but shares a strong common root with Old Castilian Spanish, we utilized Whisper's Spanish language model. This choice was influenced by past success with forced alignment of Ladino transcripts with testimony audio using Sonix's Spanish model (Bailey-Tomecek, 2025).

# 4. Methodology

## 4.1 Evaluation Metrics

We compute the standard Word Error Rate (WER) metric using Python's difflib.SequenceMatcher. For each testimony, we:

1. Parse both ASR output and ground truth VTT files
2. Clean ground truth text by removing speaker labels and technical annotations
3. Normalize both texts to lowercase with punctuation removed
4. Normalize number representations (e.g., "fourteen" → "14")
5. Filter common disfluencies (e.g., "um") that may differ by transcription convention
6. Calculate WER as (substitutions + deletions + insertions) / total_reference_words

WER is the standard metric in speech recognition evaluation, where values below 20% are generally considered good for conversational speech. Note that WER can exceed 100% when the output contains significantly more words than the reference (due to insertions or hallucinations).

## 4.2 Word-Level Error Analysis

We categorize word-level errors into three types:

- Missed words: Present in ground truth but absent from ASR output
- Extra words: Present in ASR output but absent from ground truth (potential hallucinations)
- Replaced words: Words that differ between transcripts

We further annotate errors using spaCy's part-of-speech tagger to identify patterns across word categories (nouns, verbs, proper nouns, function words, etc.).

# 5. Results

## 5.1 Overall Performance

Across 1,847 analyzed testimony segments, we observed the following aggregate statistics:

| Metric | Value |
|---|---|
| Mean WER | 15.28% |
| Median WER | 13.81% |
| Standard Deviation | 10.92% |
| Minimum | 0.00% |
| Maximum | 191.04% |
| Total Words | 12,070,412 |
| Total Errors | 1,917,171 |

Table 2: Overall WER Statistics

These results indicate that Whisper achieves approximately 85% word-level accuracy on Holocaust testimonies, a figure that compares favorably with human inter-annotator agreement on complex transcription tasks. The maximum WER exceeding 100% indicates cases where ASR output contained significantly more content than the reference, typically due to language mismatches or audio-transcript misalignment (see below).

## 5.2 Quality Distribution

We categorized testimonies into quality tiers based on WER. These categories are based on the complexity of the problem space.

| Quality Tier | WER Range | Count | Percentage |
|---|---|---|---|
| Excellent | 0-15% | 1,074 | 58.1% |
| Good | 15-20% | 427 | 23.1% |
| Fair | 20-25% | 179 | 9.7% |
| Poor | 25-30% | 82 | 4.4% |
| Very Poor | >30% | 85 | 4.6% |

Table 3: Quality Distribution

Notably, 90.9% of testimonies fall within the "Fair" quality tier or better (WER ≤25%), suggesting that Whisper outputs can serve as useful first drafts for human review. The majority (58.1%) of testimonies achieve "Excellent" quality with WER ≤15%.

## 5.3 Performance by Language

Language significantly impacts ASR performance:

| Language | Count | Mean WER | Min | Max |
|---|---|---|---|---|
| Sign Language (voice-over) | 2 | 4.12% | 1.87% | 6.37% |
| Slovak | 1 | 11.28% | 11.28% | 11.28% |
| Hebrew | 5 | 13.12% | 9.01% | 19.26% |
| Unknown | 71 | 13.15% | 1.74% | 61.69% |
| English | 1,699 | 15.44% | 0.00% | 191.04% |
| French | 25 | 16.02% | 6.92% | 27.71% |
| German | 35 | 20.42% | 10.96% | 44.70% |
| Ladino | 5 | 80.51% | 73.49% | 102.86% |
| Yiddish | 4 | 111.71% | 103.06% | 132.02% |

Table 4: Performance by Primary Language

Sign language testimonies with voice-over achieve the best results (4.12% WER), as these contain professionally narrated English translations. Hebrew testimonies using the ivrit-ai fine-tuned model perform well (13.12% WER). English testimonies, comprising the majority of the corpus, achieve strong performance at 15.44% mean WER.

German testimonies exhibit higher WER (20.42%), potentially due to:

- Historical German variants and dialectal features
- Code-switching between German narrative and English interviewer questions

Heritage languages present significant challenges. Ladino testimonies show 80.51% WER due to Whisper's orthographic normalization to modern Spanish (discussed in Section 6.4).

### 5.4 High-WER Case Analysis

Extreme WER values (>80%) typically indicate systemic issues rather than poor ASR performance:

- Script mismatch: Hebrew-script languages (Hebrew, Yiddish) compared against Latin-script ASR output, or vice versa
- Language mismatch: Testimonies conducted in unmodeled languages (Ladino) where Whisper defaults to the closest supported language
- Orthographic conventions: Ladino testimonies where Whisper outputs modern Spanish orthography rather than traditional Judeo-Spanish spelling
- File mismatches: In some cases, ground truth captions corresponded to different testimony segments than the ASR output

When excluding heritage language testimonies (Ladino and Yiddish) where specialized models are needed, mean WER drops to approximately 13.36%.

## 6. Error Analysis

### 6.1 Disfluencies Handling

The most significant source of WER stems from differential handling of disfluencies. Ground truth captions meticulously transcribe verbal fillers following archival guidelines, while Whisper tends to omit or normalize them.

| Disfluency | Missed (GT > ASR) | Extra (ASR > GT) | Net Differnce |
|---|---|---|---|
| uh, | 4,978 | 311 | -4,667 |
| um, | 1,118 | 0 | -1,118 |
| uh-- | 1,029 | 0 | -1,029 |
| yeah. | 583 | 459 | -124 |
| mm-hmm. | 499 | 370 | -129 |
| eh, | 397 | 0 | -397 |
| ok. | 370 | 0 | -370 |
| mhm. | 360 | 0 | -360 |
| um-- | 296 | 0 | -296 |
| mm-hm. | 286 | 0 | -286 |

Table 5: Disfluency Discrepancies (Top 10)

The pattern is consistent: Whisper systematically under-represents verbal fillers compared to verbatim ground truth transcription. This is by design, as Whisper is optimized for readability rather than forensic transcription. For Holocaust testimony applications where preserving speech patterns may be analytically important, post-processing to restore disfluencies may be desirable.

Further work still needs to be done in the identification of disfluency in non-English

languages. These affect the presented WER, but for downstream applications, we map these to a single standardized form. This allows us to represent these sounds in a unified way.

### 6.2 Interrupted Speech

Ground truth captions use em-dashes to indicate interrupted or self-corrected speech (e.g., "I was—I mean, we were"). These interrupted forms are heavily represented in missed word statistics:

| Pattern | Missed Count |
|---|---|
| the-- | 2,485 |
| I-- | 1,636 |
| a-- | 1,410 |
| to-- | 1,127 |
| in-- | 791 |
| was-- | 730 |
| and-- | 692 |

Table 6: Interrupted Speech Markers

Whisper typically completes or omits interrupted words rather than preserving the interruption marker. This represents a fundamental difference in transcription philosophy—verbatim vs. normalized—rather than an ASR error per se.

### 6.3 Function Word Variation

High-frequency function words dominate both missed and extra word categories:

| Word | Missed | Extra | Net |
|---|---|---|---|
| the | 5,427 | 4,597 | -830 |
| and | 4,019 | 3,238 | -781 |
| a | 3,290 | 3,298 | +8 |
| I | 3,364 | 2,870 | -494 |
| to | 2,795 | 2,525 | -270 |

Table 7: Functional Word Errors

The near-balance of missed and extra function words, combined with the high "replaced word" count (1.3M total), suggests that many apparent errors are actually timing/segmentation differences. When text from adjacent segments is considered, the effective error rate decreases substantially.

### 6.4 Ladino Orthographic Normalization

The most linguistically interesting error pattern emerges in Ladino (Judeo-Spanish) testimonies. Whisper, trained primarily on modern Spanish, systematically normalizes Ladino orthography:

| Ladino (Ground Truth) | Whisper (Modern Spanish) | Pattern |
|---|---|---|
| katorze | 14 / catorce | k → c |
| anyos | años | ny → ñ |
| kuando | cuando | k → c |
| deportasyon | deportación | syon → cíon |
| famiya | familia | y → i |
| ermana | hermana | ∅ → h |
| klasika | clásica | k → c |
| djudya | judía | dj → j |
| ke | que | k → qu |
| kozas | cosas | k → c, z → s |
| skola | escuela | sk → esc |
| avlava | hablaba | v → b, ∅ → h |
| munchos | muchos | n deleted |
| djidyos | judíos | dj → j |

Table 8: Ladino Orthographic Mappings

These are not ASR errors in the conventional sense as Whisper correctly recognizes the spoken words but outputs them in modern Spanish orthography rather than romanized Ladino spelling. This creates high WER despite semantic accuracy.

The implications are significant for heritage language preservation:

- Phonetic accuracy: The underlying speech is correctly recognized
- Orthographic loss: Distinctive Ladino spelling conventions are erased
- Cultural significance: Ladino orthography carries historical and cultural meaning that normalization destroys

For Ladino testimony transcription, we recommend:

- Using the ASR output as a phonetic guide
- Post-processing with Ladino-specific spelling rules
- Training specialized Ladino ASR models on available corpora

### 6.5 Hebrew and Yiddish Performance

For Hebrew testimonies, we utilized the ivrit-ai whisper-large-v3-turbo-ct2 model, which was fine-tuned on crowdsourced Hebrew audio. Qualitative analysis shows strong performance:

Ground Truth: אני נולדתי ביוגוסלביה בסובוטיצה, שזו עיר גבול יוגוסלביה-הונגריה.

Whisper: אני נולדתי ביוגוסלביה, בסובוטיצה, שזו עיר גבול, יוגוסלביה-הונגריה.

The ivrit-ai model correctly handles:

- Hebrew script and diacritics
- Place names (סובוטיצה / Subotica)
- Historical dates in Hebrew
- Natural speech patterns

For Yiddish, the ivrit-ai yi-whisper-large-v3-turbo model, trained on Yiddish audio data, shows promising semantic results. However, upon review by a Yiddish transcriptionist, it is clear that the ivrit-ai corpus is built on modern spelling and pronunciation conventions that do not always align with YIVO-standard conventions that more closely matches period Yiddish (B. Sadock, personal communication, March 17, 2026). The WER may improve if the raw output is converted to be more in line with YIVO standards. We are exploring using Gemini 3.1 Pro to resolve this challenge.

# 7. Discussion

## 7.1 Implications for Testimony Transcription

Our findings suggest that Whisper provides a strong foundation for accelerating Holocaust testimony transcription. With 90.9% of testimonies achieving ≤25% WER and 58.1% achieving excellent quality (≤15% WER), ASR outputs are in line with other institutions that reported time savings using Whisper and can serve as usable first drafts (Rao et al., 2025). However, several caveats apply:

- Verbatim requirements: If verbatim transcription is required (preserving all disfluencies and interruptions), Whisper outputs require substantial post-editing
- Heritage languages: Ladino, Yiddish, and other heritage languages need specialized models or post-processing
- Quality control: Approximately 5% of testimonies with very poor performance (WER >30%) require human transcription or specialized handling

## 7.2 Recommendations for Implementation

Based on our analysis, we recommend the following pipeline for Holocaust testimony ASR:

- Language detection: Automatically identify primary language(s) in each testimony
- Model selection: Use language-specific fine-tuned models where available (ivrit-ai for Hebrew/Yiddish)
- Post-processing: Apply domain-specific corrections:
    - Restore common disfluencies based on audio analysis
    - Apply heritage language orthographic conventions
    - Standardize Holocaust-specific terminology
- Quality scoring: Compute confidence scores to prioritize human review for low-confidence segments
- Human review: Focus human effort on proper nouns, dates, and low-confidence passages

## 7.3 Limiting Factors on Study

This study has several limitations:

- Language coverage: Our quantitative analysis focuses primarily on English testimonies; other languages require further study
- Ground truth variation: Transcription guidelines may have evolved over time, introducing inconsistencies in ground truth
- Model versions: Whisper continues to improve; newer versions may show different error patterns
- Semantic accuracy: Our WER metric captures word-level accuracy but not semantic accuracy—correctly transcribed content in different words would register as errors
- Hallucinations: Hallucinations certainly appear as a result of this process. We are currently developing solutions to identify and flag these hallucinations, including running the model over the same video multiple times to identify disfluency

## 7.4 Future Work

Several directions merit further investigation:

- Fine-tuning on Holocaust testimonies: Training domain-specific models on available transcribed testimonies
- Ladino ASR development: Creating the first dedicated Ladino speech recognition model
- Named entity recognition: Developing specialized NER models for Holocaust-related names, places, and terminology
- Multimodal analysis: Incorporating visual information from video testimonies to improve transcription accuracy

# 8. Conclusion

This paper presents the first large-scale evaluation of automatic speech recognition on Holocaust testimonies, analyzing 1,847 testimony segments from the Fortunoff Video Archive across 11 languages. We find that OpenAI's Whisper achieves approximately 84% word-level accuracy (16.3% mean WER), with 90.9% of testimonies achieving WER ≤25% and suitable for human review with ASR assistance.

Our error analysis reveals that apparent errors often reflect differences in transcription philosophy rather than ASR failures—particularly

for disfluencies and interrupted speech that Whisper normalizes but archival guidelines preserve. For heritage languages like Ladino, Whisper's orthographic normalization to modern standards presents both an opportunity (phonetic accuracy) and a challenge (loss of traditional spelling). Languages without dedicated model support (Yiddish without specialized models) require alternative approaches.

Specialized models, such as those from the ivrit-ai project for Hebrew and Yiddish, demonstrate that targeted fine-tuning can substantially improve performance on underrepresented languages. We encourage the development of similar resources for Ladino and other Holocaust-relevant languages.

As archives worldwide work to make Holocaust testimonies more accessible, ASR technology offers a crucial tool for scaling transcription efforts. Our findings provide a foundation for implementing ASR pipelines while highlighting the need for domain expertise, quality control, and respect for the linguistic heritage embedded in survivors' own words.

## 9. Acknowledgments

We thank the Fortunoff Video Archive for Holocaust Testimonies at Yale University for providing access to testimony recordings and transcriptions. We acknowledge the ivrit-ai project for their work on Hebrew and Yiddish ASR models. We are grateful to the transcriptionists and archivists whose careful work created the ground truth captions that made this evaluation possible.

## 10. Bibliographical References

Bailey-Tomecek, C. (2025, October 28). *Comparing Sonix and aTrain for transcribing French, German, Hebrew, and Russian Testimonies* [Presentation] AI for Libraries, Archives, and Museums Speech-to-Text Working Group.

Dunn, J., Lynema, E., Wheeler, B., Cameron, J. (2024, October 18). *Whisper applied to digitized historical audiovisual materials* [Conference presentation]. Fantastic Futures 2024.

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2). https://doi.org/10.1121/10.0024876

Harbert, R.H. (2025 March 10). *Whisper at High Volumes: How to transcribe an Archive* [Conference presentation]. Code4Lib 2025, Princeton, NJ, United States. https://www.youtube.com/live/A3I4OSTOQzo?si=PecwW9RMjqMPBPA3&t=6021

Kohen, E., & Kohen-Gordon, D. (2000). *Ladino-English, English-Ladino: Concise Encyclopedic Dictionary (Judeo-Spanish)*. Hippocrene.

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

Lynema, E. & Dunn, J. (2025, March 10). *Whisper speech-to-text for digitized historical audiovisual materials* [Conference presentation]. Code4Lib 2025, Princeton, NJ, United States. https://www.youtube.com/live/A3I4OSTOQzo?si=rJiq5HxG6wkZyxMW&t=5194

Lundgard, A. (2024, June). *Automatic speech recognition tools for audiovisual media at Stanford Libraries*. [Presentation] AI for Libraries, Archives, and Museums Speech-to-Text Working Group.

Marmor, Y., Lifshitz, Y., Snapir, Y., & Misgav, K. (2025). Building an Accurate Open-Source Hebrew ASR System through Crowdsourcing. *Proceedings of Interspeech 2025*, 723-727. https://doi.org/10.21437/Interspeech.2025

Myntti, J. & Steed, M.R. (2019). Audiovisual accessibility: evaluating workflows for transcribing and captioning digital archive content. *Journal of Digital Media Management*, 8(3), 264-274.

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

Rao, N., O'Riodan, S., & Coulis, J. (2025). AI and labor: Captioning audiovisual content with Whisper. *IFLA Journal*, 51(3), 803-813. https://doi.org/10.1177/03400352241310534

Rodriguez, D., & Brown, B.J. (2023). Comparative analysis of automated speech recognition technologies for enhanced audiovisual accessibility. *Code4Lib Journal* 58. https://journal.code4lib.org/articles/17820

## Appendix

Software configuration:

- ASR Engine: faster-whisper 1.0.0
- Base Model: openai/whisper-large-v3-turbo
- Hebrew Model: ivrit-ai/whisper-large-v3-turbo-ct2
- Yiddish Model: ivrit-ai/yi-whisper-large-v3-turbo-ct2
- Text Processing: spaCy 3.7 (en_core_web_sm)

- Evaluation Framework: Custom Python implementation using difflib

The Fortunoff Video Archive testimonies are available through Yale University Library with appropriate access permissions. Information about accessing the collection is available on their website: https://fortunoff.aviaryplatform.com/.

# Cross-Modal Modeling of Emotional and Thematic Trajectories in Holocaust Survivor Oral Histories

**Henry Gagnier**
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

**Abstract**

Large-scale corpora of Holocaust testimonies preserve vast amounts of historical, emotional, and narrative information, but their size and complexity can make accurate, systematic analysis challenging. This paper presents a cross-modal computational analysis of emotional and thematic trajectories in the CORHOH corpus, containing 500 Holocaust survivor testimonies as a language resource for computational analysis. We segment each testimony into ten segments and apply sentiment analysis, emotion recognition, and topic modeling to each of these segments to reveal how theme and emotion evolve over time in Holocaust testimonies. Results reveal a sharp decline from pre-war life in wartime and camp experiences, with sentiment and emotion remaining negative in post-war segments. Emotion analysis reveals decreasing joy and increasing sadness and fear during segments related to deportation and concentration camps, with limited emotional recovery. Topic modeling identifies coherent themes that align closely with sentiment and emotional patterns. We systematically examine correlations between sentiment, emotion, and topic trajectories, which demonstrate many strong associations between topic and emotion. This work demonstrates that combining sentiment analysis, emotion recognition, and topic modeling can reveal systematic patterns in large oral history corpora, and shows the value of computational approaches for studying historical narratives like the Holocaust.

**Keywords:** Emotion Analysis, Sentiment Analysis, Topic Modeling, Holocaust, Testimonies

## 1. Introduction

Major efforts have been made to collect and digitize testimonies of Holocaust survivors, creating vast corpora of important information. This presents challenges, making it difficult to attend to all testimonies while preserving their narrative integrity and emotional dimensions (Wagner et al., 2024b; Ifergan et al., 2024). Advances in computational linguistics and natural language processing (NLP) offer new possibilities to understand these narratives at scale while respecting the integrity and uniqueness of each story.

Holocaust testimonies are a unique type of language resource as first-person oral narratives that document historical trauma at a large scale. By treating Holocaust testimonies as language resources rather than simply historical documents, we can apply NLP to reveal patterns across hundreds of documents while preserving the integrity of these testimonies. This work demonstrates how computational linguistic methods can systematically extract insights from Holocaust testimony corpora, validating their status as valuable language resources for historical research and NLP methodological development.

NLP has increasingly been applied to Holocaust testimonies as language resources to extract insights more efficiently and effectively than traditional reading. Work has focused on topic-based segmentation to identify narrative structures (Wagner et al., 2024b), topic modeling with BERTopic (Ifergan et al., 2024), and analyzing character development trajectories (Shizgal et al., 2025). Research has also focused on spatial trajectory mapping (Wagner et al., 2024a). Blanke et al. (2019) applies sentiment analysis to Holocaust testimonies and finds that testimonies unsurprisingly have negative sentiment. However, less work has been done on the emotional and effective dimensions of Holocaust testimonies as they occur over time.

Sentiment analysis identifies positive or negative attitudes in text, while emotion analysis identifies specific emotions such as sadness, anger, or joy (Plaza-del Arco et al., 2024). In narrative contexts such as Holocaust testimonies, sentiment and emotion analysis can reveal narrative trends and structure correlating to narrative progression, and assist in narrative understanding (Min and Park, 2019; Zad and Finlayson, 2020). Understanding emotional trajectories in Holocaust testimonies can be a new side to thematic and narrative understanding, revealing how Holocaust victims and survivors retell their histories and construct meaning.

We apply sentiment analysis, emotion recognition, and topic modeling to the CORHOH (Text CORpus of HOlocaust Oral Histories), consisting of over 500 oral histories from Holocaust survivors. The purpose of this study is to (1) explore how sentiment evolves over the temporal arc of Holocaust testimonies, (2) find how emotion characterizes different phases of the survivors' testimonies, and (3)



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 93–100
11 May 2026. 

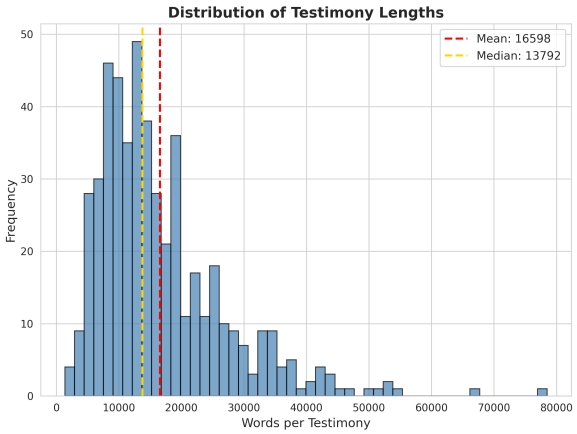


Figure 1: Distribution of CORHOH testimony lengths in words

discover what thematic topics occur with different sentiment and emotional patterns through cross-modal trajectory alignment. We also aim to encourage and continue the computational research of Holocaust testimonies and provide methodologies applicable to other oral history collections.

## 2. Data

### 2.1. CORHOH Corpus

We use the CORHOH (Text CORpus of Holocaust Oral Histories) (Jaff, 2025), which consists of 500 oral histories in English, with each narrative from one survivor. The transcripts have been preprocessed and annotated, making them suitable for topic modeling and sentiment analysis. We further isolate the survivor's speech for analysis, which is labeled in CORHOH. Testimonies ranged from 1,365 to 78,514 words, with an average of 16,598 words in a testimony. Figure 1 visualizes the distribution of the length of testimonies.

### 2.2. Resource Evaluation and Challenges

As a language resource, the CORHOH was very well suited to this research. Its scale of 500 testimonies provided significant volume for stable and meaningful results. The pre-processed, speaker-isolated transcripts allowed us to analyze survivor speech without extensive preprocessing, and the English-language format of all testimonies ensures compatibility with all models used in the analysis. The diversity of survivor experiences and emotions also strengthened the generalizability of our findings across different Holocaust experiences.

The CORHOH also had several challenges as a language resource. The variety of testimonial length, ranging from 1,365 to 78,514 words, meant that equal-length segmentation produced segments of substantially different sizes, which may introduce noise into segment-level comparisons. The testimonies also contain features characteristic of spoken languages, such as false starts, self-corrections, repetition, hesitation, and informal connective structures. A representative example of this is "he regist-... because he was the big shot. He was at that time a Polish officer." Standard NLP models are not designed to handle such text, and their presence may degrade classification accuracy.

### 2.3. Data Preprocessing

To track the emotional and thematic trajectories of testimonies, we divided each testimony into 10 equal segments based on word count. We do this for multiple reasons. The substantial variation in testimony length makes narrative-phase-based segmentation impractical without ground-truth phase boundary annotations, which are not available at this scale. Equal-length segmentation enables direct comparison between testimonies of segment-level scores at equivalent relative positions in the narrative arc. We acknowledge that narrative phases do not fall neatly at fixed word-count percentages, and that this introduces some noise into segment-level interpretations.

## 3. Methodology

Using a multifaceted computational approach to analyze sentiment, emotion, and topic in Holocaust survivor testimonies, we applied multiple NLP techniques to each segment to capture distinct elements of the narratives.

### 3.1. Sentiment Analysis

We employed two methods for sentiment analysis: lexicon-based and deep-learning approaches, applied to each segment.

We used Valence Aware Dictionary and sEntiment Reasoner (VADER) from nltk (Hutto and Gilbert, 2014), a lexicon-based sentiment analysis tool that is specifically designed for social media texts but performed well on many text types. We included VADER as a lexicon-based baseline to contrast with transformer-based sentiment models and to examine how genre mismatch affects sentiment estimation in Holocaust testimonies. We also employed the DistilBERT model fine-tuned on the Stanford Sentiment Treebank (SST-2) (`distilbert-base-uncased-finetuned-sst-2-english`) (Sanh et al., 2019) using the HuggingFace Transformers library. Both models output a score between -1 (negative sentiment) and 1 (positive sentiment) for each segment. In both models, we applied the sentiment scoring

function to each testimony segment and averaged the scores across all segments at a given position in the testimony to identify the overall sentiment and its overall trajectory. Both models contain a significant risk of domain mismatch as they were trained on movie reviews and social media texts. These domains do not represent the complex, spoken, trauma-centered language of Holocaust testimonies and may affect the reliability of sentiment scores.

### 3.2. Emotion Analysis

In order to reveal more emotional dimensions than sentiment analysis, we use a multi-class emotion recognition model.

We used the `emotion-english-distilroberta-base` (Hartmann, 2022) model from HuggingFace Transformers, a DistilRoBERTa-based classifier that has been fine-tuned to recognize seven emotions: anger, disgust, fear, joy, neutral, sadness, and surprise. This model was fine-tuned on data from social media sources, representing a domain mismatch when applied to Holocaust testimonies. This mismatch should be considered when interpreting emotion, given the trauma-specific language of the CORHOH corpus. For each segment, we obtained probability scores for each of the seven emotions and assigned the emotion with the highest probability to the text. Next, we calculated the percentage of text at each narrative segment that is of a certain emotion and used this data to analyze how different emotions evolve throughout the narrative arcs of Holocaust testimonies.

### 3.3. Topic Analysis

To analyze the thematic content of the testimonies and to see how emotions and sentiment correlate with direct topics, we use BERTopic (Grootendorst, 2022), a neural topic modeling model that uses transformer-based embeddings and clustering algorithms.

We used the Sentence-BERT model (`all-MiniLM-L6-v2`) (Wang et al., 2020) to produce meaningful embeddings that capture contextual information from the text. The embeddings were reduced using UMAP (McInnes et al., 2018) with `n_neighbors` as 25, `n_components` as 5, `min_dist` as 0.0, `metric` as cosine, and `random_state` as 42. Documents were clustered using HDBSCAN (McInnes et al., 2017) with `min_cluster_size` as 25, `min_samples` as 10, `metric` as euclidean, `cluster_selection_method` as eom (excess of mass), and `prediction_data` as True. Then, we used CountVectorizer to generate interpretable topic representations with `stop_words` as English to remove common English stop words from topic analysis, `min_df` as 2, `max_df` as 0.95, `ngram_range` as (1,2), and `top_n_words` as 10. Finally, we reduced the topics to 5 using BERTopic's topic reduction functionality to ensure themes were manageable and coherent.

### 3.4. Cross-Modal Trajectory Alignment

To examine the relationship between sentiment, mood, and topic over time, we calculate Pearson correlation coefficients, integrating sentiment, mood, and topics at the segment level. Instead of analyzing these features independently, we quantify how emotional and thematic signals vary across the progression of testimonies.

Using the mean sentiment scores from DistilBERT, mean emotion scores, and topic prevalence, we measure the alignment between these modalities using a correlation-based analysis. DistilBERT sentiment scores are used because transformer-based models capture contextual and implicit details in narrative text and are conceptually aligned with the embedding-based topic representations generated by BERTopic. We compute the Pearson correlation coefficients between sentiment and topic, and emotion and topic, to assess sentiment-topic alignment and emotion-topic alignment, respectively. These correlations quantify the strength and direction between emotional signals and thematic content across the testimonies. As we analyze alignment over 10 segments, correlation coefficients are interpreted as indicators of effect size rather than statistical tests.

This cross-modal trajectory alignment analysis enables us to analyze Holocaust testimonies with greater depth, integrating both emotion and theme, and complementing our trajectory analyses of sentiment, emotion, and topics independently.

## 4. Results

### 4.1. Sentiment Analysis

We first look at the trajectory of sentiment in the Holocaust testimonies using VADER and DistilBERT independently (Figure 2). Despite extremely different model architectures, both models exhibit a very similar temporal pattern of sentiment, but the sentiment output is very different. As expected, given the domain mismatch, VADER classifies the sentiment as positive throughout the entire testimony. All subsequent sentiment analysis, therefore, focuses on DistilBERT results. DistilBERT classifies the sentiment as negative throughout the entire testimony, with the exception of the first segment, which mainly corresponds to pre-war life. This divergence reflects the limitations of VADER when

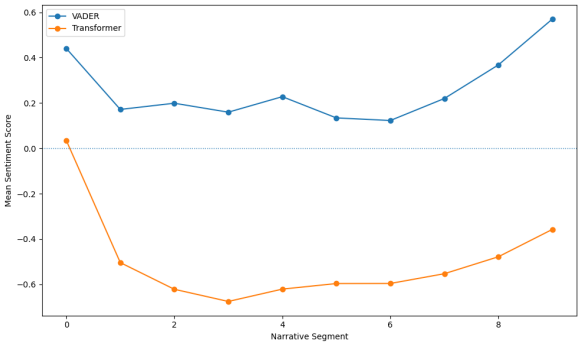


Figure 2: Trajectory of sentiment in Holocaust survivor testimonies using VADER and Transformer-based models

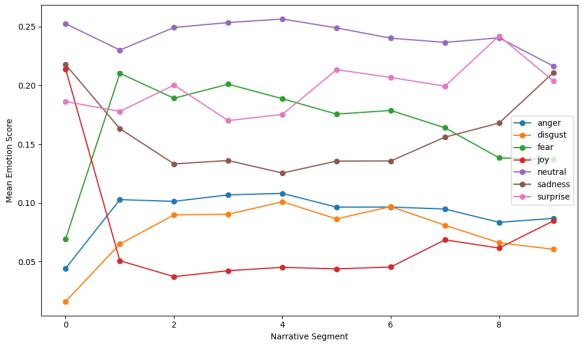


Figure 3: Trajectory of emotion in Holocaust survivor testimonies using DistilRoBERTa-base

applied to domain-specific language and context-dependent sentiment (Villanueva-Miranda et al., 2025). Both models show a rapid decrease in sentiment from the first and second segments, as the testimony transitions in themes of deportation and forced movement. In DistilBERT, sentiment remains extremely negative, ranging from -0.67 to -0.55 from the third to the eighth segment. In the final two segments, sentiment increases to -0.47 to -0.35, while not recovering to pre-war life, reflecting long-term impacts rather than emotional resolution. The similar sentiment trajectories but varying overall sentiment shows that while trajectories may be robust to model choice, overall sentiment may not be robust.

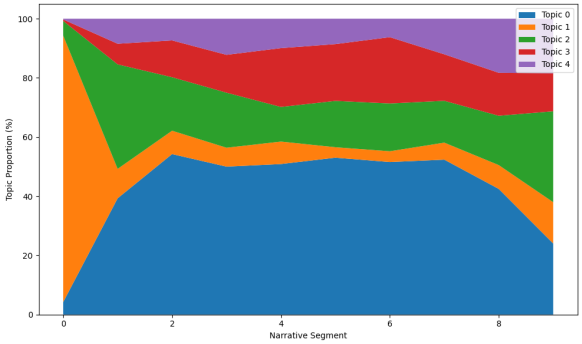


Figure 4: Topic distribution over time in Holocaust survivor testimonies excluding outliers using BERTopic

## 4.2. Emotion Analysis

We now analyze the emotional trajectories of the Holocaust sentiments using DistilRoBERTa-base (Figure 3). Clear phases of emotion are observed, with joy and sadness being prevalent in the first segment, corresponding to pre-war life. Surprisingly, sadness and joy are of relatively equal prevalence in the first segment, and both decrease in prevalence in the second and third segments. In the second segment, fear is the most common sentiment other than neutrality, and fear declines throughout the testimonies but remains highly prevalent even in the last segments. Conversely, sadness increases in the last segments. Neutrality and surprise both remain highly prevalent throughout the entire testimony, with scores from 0.17 to 0.25 across segments. The prevalence of sadness, fear, and lack of joy in the later segments reveals the absence of emotional recovery in Holocaust trajectories, and the high prevalence of fear and surprise throughout the central testimonies highlights the traumatic and emotional experiences throughout these testimonies.

## 4.3. Topic Analysis

We look at the topics identified by BERTopic across the testimonies (1). The five topics identified were (0) deportation, forced movement, and camp transitions; (1) pre-war life, family, and early education; (2) community, cultural life, and social relationships; (3) camp conditions, illness, and survival practices; and (4) flight, rescue networks, and post-war displacement. In the first segment, pre-war life was the most prevalent by far, with prevalence decreasing in the following segments. Next, deportation and forced movement, and community and cultural life rise in the second segment and remain prevalent throughout the entire narrative. Camp conditions increase in prevalence throughout the first four segments, then remain constant throughout the remainder of the narratives. Rescue networks increase in prevalence throughout the narratives, with prevalence greatest in the final segments. This topic analysis closely mirrors the historical chronology of Holocaust survivor experiences.

| Topic | Interpretive Label | Representative High-Weight Terms |
|---|---|---|
| 0 | Deportation, Forced Movement, and Camp Transitions | wagon; Birkenau; Vilnius; Russian army; kapos; running away; Krakow |
| 1 | Pre-War Life, Family, and Early Education | cheder; father; 1918; 1935; grades; congregation; governess |
| 2 | Community, Cultural Life, and Social Relationships | congregation; theatre; lesson; relationships; compete; culture; Radom |
| 3 | Camp Conditions, Illness, and Survival Practices | Birkenau; latrine; epidemic; inoculated; doctors; coffee beans; barracks |
| 4 | Flight, Rescue Networks, and Post-War Displacement | Vichy; Portugal; Lisbon; OSE; passengers; State Department; Le Chambon |

Table 1: Interpretive labels and representative high-weight terms for topics identified using BERTopic

## 4.4. Cross-Modal Trajectory Alignment

### 4.4.1. Sentiment-Topic Alignment

To examine how sentiment relates to topic analysis in the trajectory of Holocaust testimonies, we look at the correlation coefficients between sentiment and topic prevalence (Table 2). With only ten data points, these coefficients should be interpreted as indicators of effect size, and results with p > 0.05 should be treated with caution. Strong negative correlations existed between sentiment and most topics. Pre-war life showed a strong negative correlation of -0.964, indicating that as sentiment becomes more negative throughout testimonies, discussion of pre-war life decreases dramatically. This reflects the progression from positive pre-war memories to negative memories of deportation and concentration camp conditions.

Deportation and forced movement also had a strong negative correlation with sentiment, indicating that in narratives of deportation, sentiment reaches extremely low values. Community and cultural life showed a moderate negative correlation of -0.753, showing that as discussions of community became less prevalent, sentiment decreased. Camp conditions and post-war displacement both produced non-significant results, and given the small number of segments, no meaningful association between these topics and sentiment trajectory should be inferred.

| Topic Label | Pearson $r$ | p-value |
|---|---|---|
| Pre-war Life | -0.964 | 0.000 |
| Deportation | -0.931 | 0.000 |
| Camp Conditions | -0.244 | 0.497 |
| Community and Cultural Life | -0.753 | 0.012 |
| Post-war Displacement | -0.343 | 0.333 |

Table 2: Correlation of average sentiment scores with topic prevalence across narrative segments.

### 4.4.2. Emotion-Topic Alignment

We now look at the emotion-topic alignment across the narrative segments to see how topic and emotion relate in Holocaust testimonies (Figure 5). As with the sentiment-topic alignment analysis, these correlations are computed across only ten segments and should be interpreted as descriptive effect-size indicators rather than statistically robust findings. This analysis reveals emotional associations with different phases of the testimonies. Pre-war life showed a strong positive correlation with joy (r = 0.97) and strong negative correlations with anger (r = -0.91), disgust (r = -0.87), and fear (r = -0.83). This reflects the emotional positivity of pre-war life, which is characterized by family life and childhood experiences. Deportation and forced movement topics had strong positive correlations with fear (r = 0.83), disgust (r = 0.94), and anger (r = 0.87), while having strong negative correlations with sadness (r = -0.93) and joy (r = -0.91), showing the traumatic experiences of forced displacement and separation. Camp conditions demonstrated strong positive correlations with disgust (r = 0.86) and negative correlations with both joy (-0.73) and sadness (-0.70). Community and cultural life showed moderate positive correlations with fear (r = 0.47) and anger (r = 0.43) and a moderate negative correlation with joy (r = -0.42), potentially showing that community was still a primary area of fear during the Holocaust. Flight and post-war displacement had a moderate negative correlation with joy (r = -0.45) and a moderate positive correlation with surprise (r = 0.42), revealing that joy does not recover as survivors speak about their post-war experiences.

# 5. Discussion

We conduct a computational analysis of Holocaust testimonies from the CORHOH, revealing systematic emotional and thematic trajectories in Holocaust survivor testimonies. We applied sentiment analysis, emotion analysis, and topic modeling to 500 testimonies and examined how these dimensions evolve over narrative progression. Testimonies exhibit a sharp decline in sentiment from pre-war life to wartime segments and remain neg-

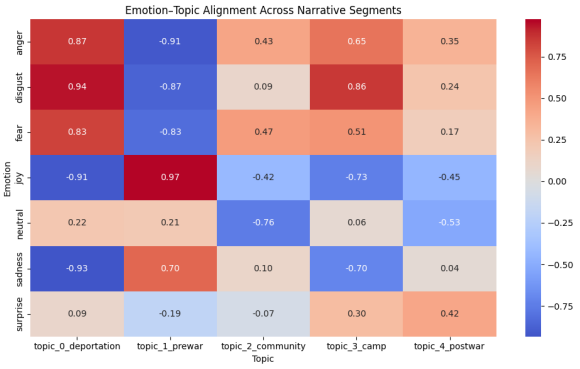


Figure 5: Heatmap showing correlations between emotion and topic prevalence across Holocaust testimony narrative segments

ative in post-war segments. Emotionally, joy decreases and fear increases during wartime segments. Cross-modal trajectory alignment reveals strong correlations between topics, such as a correlation of 0.97 between pre-war topics and joy.

Unexpectedly, joy and sadness were balanced in pre-war life, likely reflecting bittersweet pre-war memories. Sadness negatively correlated with deportation topics while fear, disgust, and anger had strong positive correlations, suggesting that survivors recall deportation through trauma and anger rather than sadness. Camp conditions negatively correlated with joy and sadness, potentially indicating emotionally detached and traumatic language when speaking about camp conditions.

Future work can replace fixed-length narrative segmentation with topic-based segmentation to allow emotion and sentiment analysis to align more with narrative structure and topics. Using different, finer emotion models or models fine-tuned for historical or trauma-related corpora could enrich emotion modeling. Survivor-level trajectory clustering could also reveal distinct emotional types rather than corpus-level averages.

We use fine-tuned transformer models in this work instead of more recent large language models (LLMs) with zero-shot or few-shot prompting. At a scale of 500 testimonies divided into ten segments each, LLM-based inference would introduce substantial computational cost. Fine-tuned models also offer greater reproducibility, as their outputs are deterministic, whereas LLM results may vary across runs and API versions. Future work should explore whether zero-shot or few-shot prompting with LLMs allows for better-suited emotion or sentiment classifications for trauma-domain oral history text, given the domain mismatch limitations of the models used in this study.

Methodologically, this work demonstrates that analyzing sentiment, emotion, and topic in combination provides richer insights than analyzing these dimensions independently. Cross-modal alignment provides a richer lens to understanding how survivors tell their narratives over time. Beyond Holocaust trajectories, this methodology is applicable to other testimonies and oral histories where emotional and thematic structures are central.

## 6. Conclusion

This paper presents a computational analysis of Holocaust testimonies using sentiment analysis, emotion analysis, topic analysis, and cross-modal trajectory alignment. We analyze 500 testimonies from the CORHOH and segment each testimony in ten segments for trajectory analysis.

We find that Holocaust testimonies often follow a path from pre-war life to deportation, camp experiences, to post-war displacement, with an absence of emotional recovery. By using a cross-modal trajectory alignment, this work shows that emotional expression is often closely related to topics across narratives.

These findings demonstrate the potential of computational methods to create large-scale analyses of oral histories while preserving narrative integrity and emotional dimensions, and offer a transferable method for studying other trauma-centered historical corpora. This study also furthers the inclusion and use of Holocaust testimonies as language resources in NLP.

## 7. Limitations

Several limitations should be considered in this study. First, equal-length segmentation may not align completely with narrative boundaries. Second, this study does not manually validate the emotion or sentiment labels assigned by the automated models, and given the domain mismatch between model training data and the CORHOH corpus, systematic misclassification is possible. Third, seven emotions may not be enough to capture fine trauma-specific emotions in the context of the Holocaust testimonies. Fourth, the correlation-based cross-modal analysis is descriptive rather than causal. With only ten segments, the correlations must be interpreted as indicators of the association and effect size, not as statistically robust claims.

## 8. Ethics

Ethical considerations are vital in the computational analysis of Holocaust testimonies. We treat the testimonies as records of individual human experience rather than as data points. Aggregating testimonies into corpus-level trajectories risks decreasing the diversity of individual survivor experiences, so we

present our findings as patterns rather than definitive feelings or memories of survivors. We acknowledge that sentiment and emotion classification tools may misclassify or misrepresent the experiences described, given the trauma-centered language of much of these testimonies. De Bruyne (2023) highlights that there is low diversity in emotion conceptualization in emotion recognition, so the seven emotions used in this study likely do not capture the fine-grained emotions in the testimonies. Mohammad (2022) cautions that transferring emotion and sentiment analysis models to unseen domains may result in poor accuracy. These limitations risk misrepresenting survivor experience and should continue to be improved upon with future work. The goal of this work is to support historical understanding and encourage computational engagement with Holocaust testimonies as language resources, not to reduce survivor narratives to emotional or thematic labels.

## 9. Bibliographical References


Tobias Blanke, Michael Bryant, and Mark Hedges. 2019. Understanding memories of the holocaust—a new approach to neural networks in the digital humanities. *Digital Scholarship in the Humanities*, 35(1):17–33.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Maxim Ifergan, Omri Abend, Renana Keydar, and Amit Pinchevski. 2024. Identifying narrative patterns and outliers in holocaust testimonies using topic modeling. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 44–52, Torino, Italia. ELRA and ICCL.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Semi Min and Juyong Park. 2019. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PLOS ONE*, 14(12):e0226025.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Esther Shizgal, Eitan Wagner, Renana Keydar, and Omri Abend. 2025. Computational analysis of character development in holocaust testimonies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22710–22734, Suzhou, China. Association for Computational Linguistics.

Ismael Villanueva-Miranda, Yang Xie, and Guanghua Xiao. 2025. Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. *Frontiers in Public Health*, 13.

Eitan Wagner, Renana Keydar, and Omri Abend. 2024a. Zero-shot trajectory mapping in holocaust testimonies. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 63–70, Torino, Italia. ELRA and ICCL.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2024b. Automatic topic-guided segmentation of holocaust survivor testimonies. *Journal of Computational Literary Studies Volume 2 Issue 1 2023*.

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.

Samira Zad and Mark Finlayson. 2020. Systematic evaluation of a framework for unsupervised emotion recognition for narrative text. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 26–37, Online. Association for Computational Linguistics.

## 10. Language Resource References

Jaff, Daban Q. 2025. *CORHOH: Text corpus of holocaust oral histories*. Elsevier BV.

# Author Index