



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

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

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 77.7%. On 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

<i>Integrating TEI Publication, Guided Exploration, and Vector Databases for Semantic Search in the Voci dall'Inferno Project</i> Angelo Mario Del Grosso, Elvira Mercatanti, Carla Congiu and Marina Riccucci	1
<i>Towards Semantic Searching in Diverse Multimodal Collections</i> Václav Kučera, Martin Bulín, Jan Švec and Pavel Ircing	12
<i>Automatic Transcription of Holocaust Testimonies in Yiddish: Orthographic Comparison and Cross-Domain Validation</i> Isaac L. Bleaman	20
<i>From Consensus to Split Decisions: ABC-Stratified Sentiment in Holocaust Oral Histories</i> Daban Q. Jaff	29
<i>EHRI Annotator: A Web-Based Tool for Named Entity Recognition and Linking in Holocaust-Related Texts</i> Maria Dermentzi	37
<i>The Shape of Testimony: A Scalable Framework for Oral History Archive Comparison</i> Renana Keydar, Amit Pinchevski and Itamar Trainin	47
<i>From Oral History to Structured Data: The MalachNER Dataset</i> Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián and Pavel Pecina	59
<i>Emotions In Oral History Interviews: A Multimodal Approach to Holocaust Testimonies</i> Nele Mantaj, Vaibhav Agarwal and Ines Matres	66
<i>Modeling the Language of Holocaust Survivors' Testimony with Domain-Adapted Transformers</i> Christopher Brückner, Jan Lehečka, Jan Švec and Pavel Pecina	74
<i>Evaluating Automatic Speech Recognition for Holocaust Testimonies: A Large-Scale Analysis of Whisper Performance on the Fortunoff Video Archive</i> William J.B. Mattingly and Christy Bailey-Tomecek	84
<i>Cross-Modal Modeling of Emotional and Thematic Trajectories in Holocaust Survivor Oral Histories</i> Henry Gagnier	93

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	<i>Modeling the Language of Holocaust Survivors' Testimony with Domain-Adapted Transformers</i> Christopher Brückner, Jan Lehečka, Jan Švec and Pavel Pecina
16:50–17:10	<i>Evaluating Automatic Speech Recognition for Holocaust Testimonies: A Large-Scale Analysis of Whisper Performance on the Fortunoff Video Archive</i> William J.B. Mattingly and Christy Bailey-Tomecek
17:10–17:30	<i>Cross-Modal Modeling of Emotional and Thematic Trajectories in Holocaust Survivor Oral Histories</i> Henry Gagnier
17:30–18:00	Pannel Discussion
	Final Remark