CLARIN Resource Families for Oral History

Jakob Lenardič,* Silvia Calamai, Stefania Scagliola,† Henk van den Heuvel‡**

*Institute of Contemporary History, Slovenia; **Università di Siena, Italy; †Independent researcher; ‡Radboud University, the Netherlands

Introducing the CLARIN Resource Families (CRF)

https://www.clarin.eu/resource-families

Background and Aims

- In a nutshell: Manually curated overviews of language resources and tools deposited in CLARIN repositories (Lenardič and Fišer 2022)
- Three-fold aim:
  1. User-friendly overviews focussing on reuse of CLARIN resources and tools in digital humanities and social sciences research
  2. Comprehensive and harmonized metadata documentation in line with FAIR
  3. Part of CLARIN curation efforts (Goosen, Haaf, and Windhouwer 2018)
- From the perspective of FAIR principles (Wilkinson et al. 2016), CRF mainly contributes to **findability** and **reusability**.
  1. **Findability**: collating typological characteristics (e.g., parliamentary vs. newspaper vs. oral history corpora); such typology is not encoded in a principled way in CMDI metadata (Windhouwer and Goosen 2022)
  2. **Reusability**: unified description of each of the tools and resources that is tailored to the unique technical features of each family, and their qualitative characteristics

In digital humanities and social sciences, the structure and quality of provided metadata are widely divergent; CRF aims to account for gaps and harmonize differences.

The Existing CRF Families

- 15 **corpus families**
  Parliamentary corpora, Computer-Mediated Communication Corpora, Sign Language Corpora, L2-Learner Corpora, etc. in addition to Oral History Corpora
- 6 **lexical resource** families
  Language Models, Lexica, Dictionaries, Conceptual Resources, Glossaries, and *Wordlists*
- 5 **tool** families
  Corpus Query, Normalisation, Named Entity Recognition, Part-of-Speech Tagging and Lemmatisation, and *Tools for Sentiment Analysis*

In total, over 1000 manually curated tools and resources across CLARIN repositories

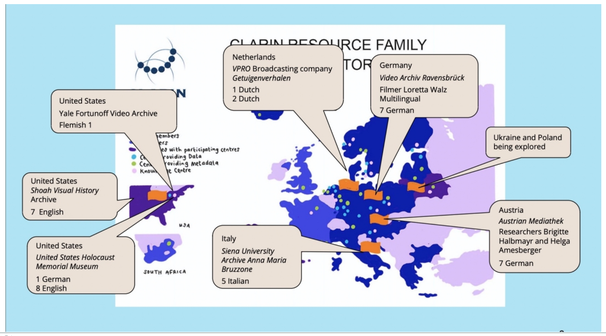
The Oral History Family

https://www.clarin.eu/resource-families/oral-history-corpora

Introduction

- Currently consisting of a single subfamily – the *Voices from Ravensbrück* corpora
- Organised and funded in the context of CRF Project Funding (Calamai et al. 2022)

Voices from Ravensbrück
38 oral history interviews with survivors from the Ravensbrück concentration camp for women



Corpus	Language(s)
Collection Bruzzone	Italian
Collection VPRO Selma van der Perre	Dutch
US Holocaust Memorial Museum	English, Hebrew, French, German, Dutch
USC Shoah Foundation Visual History Archive Online	English, German
Getuigenverhalen	Dutch
Die Frauen von Ravensbrück (Loretta Waltz collection)	German, French, English
Austrian Mediathek (Brigitte Halbmayr und Helga Amesberger)	German
Fortunoff Video Archive for Holocaust Testimonies	English, French, Hebrew, Slovak, German

Metadata

- *Collection Bruzzone* available for download (under a restricted licence) through CLARIN (The Language Archive, MPI; Nijmegen), see Bruzzone and Beccaria Rolfi (1976); the recordings in the other corpora can be streamed online directly or after registration.
- *Collection Bruzzone* was created specifically in the context of CLARIN:
  - Digitisation of 14 audio cassettes recorded in the 1970s
  - Orthographic transcription using the CLARIN Transcripion Portal (Draxler et al. 2020), which is tailored to interview data
- Extensive metadata about (i) context of creation (analogue recordings on cassettes) and (ii) context of digitisation are made available for each of the 38 interviewees separately

Voices from Ravensbrück

Corpus	Language	Description	Availability
Collection Bruzzone **Size:** 4 interviews related to Ravensbrück **Annotation:** Orthographic **Licence:** Restricted access (CLARIN RES+PLAN+BY+NC+INF+PRIV+NORED+ND+DEP+*)	Italian	The corpus contains four interviews involving five ex-deportee in the female-only Nazi concentration camp of Ravensbrück (Lidia Rolfi, Bianca Paganini, Livia Borsi, Lina and Nella Baroncini). By clicking on the download button you are brought to a page with a description of the four interviews. By clicking on the name of the interviewee you are brought to a page with a brief description and extensive metadata of the single interview. On the right you can see the corresponding audio- and text files marked in orange. This means they are restricted and you need to contact the owner, prof. Silvia Calamai (silvia.calamai@unisi.it) to get access to this data. The files marked in green are short clips that you can download directly.	Download

New CMDI Metadata Profile for Oral History Corpora

In the *Voices of Ravensbrück* project, a new CMDI metadata profile was created that is tailored to to the needs of scholars who need to apply source criticism to digitized data.

OralHistoryInterviewCRF profile, see CLARIN's component registry (https://catalog.clarin.eu/ds/ComponentRegistry). The following bespoke components:

1. Interview General
   Number of Speakers, Creation Date, Publication Date, etc.
2. Modality
   Spoken
3. Creation Context
   Description, Contact, Role, Name, etc.
4. Digitization Context
   Description, Contact, Role, Name, etc.
5. Multilinguality
   Monolingual or *Multilingual*
6. Interview Content
   Language, Spatial Coverage, Temporal Coverage, Full Transcript, Interview Keywords, etc.
7. Interview Method
   Interview Sort, Recruitment Method, Pre-Interview Information, Data Collection Method, Topic List
8. Interviewee
   Birth Place, Residence Place, Role, Name, Ethnic Group, Birth Year, Education, Profession, Anonymised – yes/no, Birth Country, Actor Languages, etc.
9. Interviewer
   Relation to Interviewee, Relation to Project, Birth Place, Role, Name, Full Name, Ethnic Group, Age, Sex, Anonymised – yes/no, etc.
10. Country
11. Interview Media
    Media Type, Media Format, Media Quality, Recording Conditions, Speech Technical Metadata, etc.
12. Interview Annotation
    Annotation Protocol, Character Encoding, Annotation Types, Format, etc.

CRF Project Funding

https://www.clarin.eu/content/clarin-resource-families-project-funding

- EoIs for small projects (3–6 PMs) contributing to CRF
- Envisaged activities
  New CRF overviews, comprehensive metadata curation, enhancements of existing tool sets, harmonisation of heterogenous metadata schemas, etc., but clear CLARIN relevance required

References

Bruzzone, Anna Maria and Lidia Beccaria Rolfi (1976). *Collection "Anna Maria Bruzzone's Ravensbrück Interviews"*. The Language Archive. URL: https://hdl.handle.net/1839/e24ae5a4-be49-4c31-a7b8-b0c9ed84029e.

Calamai, Silvia, Stefania Scagliola, Fabio Ardolino, Christoph Draxler, Arjan van Hessen, and Henk van den Heuvel (2022). "Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant". In: *Proceedings of the CLARIN Annual Conference 2022*, pp. 1–9. DOI: 10.3384/ecp1891.

Draxler, Christoph, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti (2020). "A CLARIN Transcription Portal for Interview Data". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3353–3359. URL: https://aclanthology.org/2020.lrec-1.411.

Goosen, Twan, Susanne Haaf, and Menzo Windhouwer (2018). "CLARIN metadata best practices, curation and support". In: *Clarin Annual Conference 2018*. URL: https://www.clarin.eu/sites/default/files/CLARIN2018_BazaarPoster_CMDI-TF-MD-TF.pdf.

Lenardič, Jakob and Darja Fišer (2022). "The CLARIN Resource and Tool Families". In: *CLARIN: The Infrastructure for Language Resources*. Vol. 1, pp. 343–372. DOI: 10.1515/9783110767377-013.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1, pp. 1–9. DOI: 10.1038/sdata.2016.18.

Windhouwer, Menzo and Twan Goosen (2022). "Component Metadata Infrastructure". In: *CLARIN: The Infrastructure for Language Resources*, pp. 191–222. DOI: 10.1515/9783110767377-008.