CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event.

Table of Contents

Voices from Ravensbrück. Towards the creation of an oral and multilingual resource family

Silvia Calamai
Università di Siena
Siena, Italy
silvia.calamai@unisi.it

Jeannine Beeken
University of Essex
Colchester, United Kingdom
jeannine.beeken@essex.ac.uk

Max Broekhuizen
Erasmus School of History
Rotterdam, The Netherlands
maksbroekhuizen@gmail.com

Christoph Draxler
Ludwig Maximilian University
Munich, Germany
draxler@phonetik.uni-muenchen.de

Arjan van Hessen
University of Twente
Enschede, The Netherlands
a.j.vanhessen@utwente.nl

Henk van den Heuvel
Radboud University
Nijmegen, The Netherlands
h.vandenheuvel@let.ru.nl

Stefania Scagliola
University of Luxembourg
Belval, Luxembourg
scagliolas@gmail.com

Abstract

This paper describes a pilot project aimed at introducing a new type of corpus in the CLARIN resource family tree, called 'narratives'. To this end, a multilingual corpus of existing interviews with survivors of concentration camp Ravensbrück will be curated following CLARIN compliant standards. During WWII this German camp imprisoned 130.000 women from 20 different nationalities. This diversity creates the opportunity to build a unique corpus of gender specific interviews, covering the same topic, narrated in a similar structure, but voiced in different languages. The corpus will also be enriched with various types of annotation (transcripts e.g.).

1 CLARIN Resource families and oral history

The CLARIN Resource Family is a user-friendly overview per data type of available language resources in the CLARIN infrastructure aimed at the needs of researchers from (digital) humanities and social sciences and human language technologies.[1] Within this overview, there is only one entry for 'spoken corpora', which contains 90 data sets mainly targeted at phonetic, linguistic and speech technology research. We argue for a new type of entry, namely the datatype 'narratives', covering oral history interviews and other types of spoken narrative discourse, in both audio and textual form. Interviews, aside from oral history, are a central object of research in a broad variety of fields such as anthropology, psychology, literary studies, sociology, health studies, education, linguistics and cognitive science. Yet there seem to be scarce opportunities for comparison and cross-fertilisation between these disciplines.

It is our contention that oral history interviews represent an under-utilized but promising datatype outside the realm of history, and that the CLARIN infrastructure is the ideal 'home' for this type of 'family' to illustrate its multidisciplinary potential. With this project we want to substantiate this position, by bringing together and comparing similar interview data – retrospective spoken narratives of women – from different countries and languages – about the same topic: going through war and trauma in concentration camp Ravensbrück. First the characteristics of oral history and opportunities for digital humanities will be discussed. This is followed by a sketch of the data that is currently available and data

[1] https://www.clarin.eu/resource-families

in other languages that will be explored. We conclude with a description of the envisioned workflow for collecting, enriching and publishing the corpus.

2 The underutilized potential of Oral History

The key characteristic of oral history is that it is co-created and mediated through the use of language, speech and memory. Its content is thus *multi-layered*. It can also be appreciated in *multimodal* ways (i.e., seeing the interview, hearing the recording, reading the transcript). An oral history interview offers information through a single person, but from the perspective of interactions in a social context (e.g., the family, the village, the military unit, the company). These descriptions reveal the creation of identity vis-à-vis social, economic, political and cultural contexts.

Aside from what is said, one can also reflect on the organization of the narrative: what is not said, what is repeated again and again, the language and terms that are used. Another dimension to study is the interaction that takes place during the interview: the mediating process between interviewer and interviewee who together are co-creating an historical source. Typical questions relate to differences in gender, background, status and the relation between standard and vernacular speech in the course of the interview.

With regard to modalities, the encounter can be studied: (1) by listening to the audio-signal, (2) by identifying speaker-specific physiological information with digital tools (segmental and suprasegmental acoustics, silence, emotions, speech rate and timing), and (3) by reading the textual representation of what is uttered, which is characterized by the transcription convention. Regardless of the approach, the interview has value as singular source, as well as in the form of a collection (similarities and differences).

With regard to ethical and legal issues and re-use of data, oral history data has a different background than spoken corpora that are specifically created to study language. This is evident with regard to recordings made before the digital era, when giving consent for use of the source meant that it could be consulted at the premises of an archive under supervision of someone who can be held accountable. But even for projects initiated in the digital era, that can be accessed through the web in secured ways, the problem is encountered that consulting the data is a different research practice then processing the data.

The first is a one-to-one encounter: you listen to a signal (the audio of an interview) or read one transcript at a time, while 'processing' the data, for example to detect signals or patterns, entails an automated process: you need access to the file itself to run it through the software. In a way this is taking Alessandro Portelli's adage to 'bring back orality to oral history' one step further. He pledged for shifting the focus of research from the transcript to the auditory features of the source: the tone and rhythm of a voice (Portelli 1981).

We strive to go back to the signal itself and discover patterns in speech and non-verbal features. Technically this does not pose problems, but with regard to privacy, copyright and access control, we will have to enter a 'trading zone' with the 'guardians' of oral history data and design special measures to guarantee that the interviews will be treated in a respectful way (Calamai *et al.* 2019). Once such measures will be taken up, they will secure access to a 'bonanza' of data that is awaiting to be understood in 'new ways'.

3 The historical context of Ravensbrück and its suitability for our objective

Such 'bonanza' of data in terms of richness of voices, speakers, narrative styles and topical coverage, requires a clear structure for orientation. The biggest challenge is cutting down the abundance of variables that become visible when joining data from different contexts of creation, to a set of parameters that make a comparison meaningful within a particular paradigm. That too many variables within one theme, is not productive for such a comparison, is what we learned during a CLARIN workshop in Munich in 2018 (*Oral History: Users and their scholarly practices in a Multidisciplinary world*, Munich 2018). The idea was to bring together interviews on the topic of migration in different languages as a basis for experimenting with annotation, analysis, and emotion recognition tools. There was however too much diversity: different discourses on migration and different tools to apply on this data. The focus of the workshop was therefore put on breaking down 'silos' of knowledge, identifying the obstacles for the uptake of these tools in different disciplines that work with interview data.

With the donation of the archive of the Italian scholar Anna Maria Bruzzone, who interviewed five survivors of Ravensbrück for her book, to the University of Siena in 2016, a new opportunity for cross-

disciplinary multilingual research presented itself. Interviews about experiences in one camp have a more specific set of variables. Moreover the spoken memories of these women have been collected extensively in many of the 20 countries to which the women returned at the end of the war. This makes comparisons across languages and diverse historical contexts more viable.

The camp, built in 1939 and located in northern Germany, 90 km north of Berlin, was initially intended for social outcasts or so-called 'inferior beings' that had to be re-educated (Romani people, political dissenters, criminals and prostitutes). As the German occupation expanded to new territories, new types of female prisoners with either a Jewish background or who had been involved in rescue or illegal operations, were deported to the camp. This evolution explains the widely divergent background of the prisoners. Of the 130,000 women from 20 different nationalities that passed through it, 48,500 came from Poland, 28,000 from the Soviet Union, 24,000 from Germany and Austria, 8,000 from France, and thousands from other countries. More than 20,000 women among this population were Jewish, and 80 percent were political prisoners. Many of these prisoners were employed as forced laborers by Siemens & Halske. From 1942 to 1945, medical experiments were undertaken as well.

What all narratives about being imprisoned in Ravensbrück have in common is the gender perspective. What makes the diversity of the narratives interesting is the socio-cultural context in which the retrospective account is expressed. The story of a former Polish prisoner who immigrated to the USA and tells her story for the Shoah Visual Archive, can be quite different from someone with exactly the same background at the time, who returned to Poland after the liberation of the camp and contributes to a Polish interview project.

4 Data that is currently available

Web research and consultation of a number of authors has yielded enough data and commitment to be able to search for profiles that match those of the five narrators in Bruzzone's archive. We intend to first complement the Italian interviews with English, Dutch and German ones. In the long run we intend to expand the project to Eastern Europe, especially to Poland and Russia, where many survivors came from.

Bruzzone's Ravensbrück interviews consist of 14 audio cassettes, with a total duration of about 18 hours and 20 minutes. The analogue audio cassettes were digitized according to IASA standards (.wav format, 96000 Hz, 24 bit). The archive contains 4 long interviews. We know that for her publication, Anna Maria Bruzzone transcribed the recordings step by step, writing everything down that she heard, but unfortunately, the handwritten transcriptions were lost. [2] In 2016, the book was translated to German.

For Dutch, we have access to interviews held with Dutch Ravensbrück internees between 2007 and 2010, for a PhD study on the memory of Ravensbrück by Susan Hogervorst (2010). In case this material may pose legal problems, due to the absence of consent from narrators who have deceased,[3] we can fall back on the many project-generated interview collections that are publicly available in archives. These interviews have been created in the wake of the 50th anniversary of the second World War from the 1990s onwards, and their proliferation has been strongly influenced by the availability of digital technology and a push towards presence on the web. In the Netherlands this has resulted in the online resource *Getuigenverhalen*[4] which contains 3 interviews on Ravensbrück. The Visual History Archive contains 5 interviews in Dutch about Ravensbrück. The United States Holocaust Memorial Museum contains 1 interview in Dutch about Ravensbrück.

With regard to interviews in German, we have access to the *Videoarchiv "Die Frauen von Ravensbrück"* (200 interviews), *Österreichische Lagergemeinschaft Ravensbrück und Freundinnen* (34 interviews). In English, we have the *Visual History Archive* (20 interviews) and the *Imperial War*

[2] Hopefully, with digital repositories this will never happen again.

[3] Legal issues may occur at different levels in case of oral archives, due to the transposition of the General Data Protection Regulation (GDPR) into different national laws. In Italian law, just as an example, GDPR applied also to dead people. At a general level, one may consider the right of the interviewers, that of the interviewees, and that of third parties mentioned during an interview. In some national laws, interviews are also protected by the laws on authors' rights.

[4] http://getuigenverhalen.nl/home

Museum GB (8 interviews). We are confident that in some of the countries involved in the CLARIN network, additional similar material might be found, in further languages.

5 Towards a CLARIN Resource Family for Oral History

The planned project consists of two phases: 1. basic curation of available data and exploration of data for expansion, 2. curation and enrichment of all available data. In the first phase, five Italian interviews from the Bruzzone collection will be described and transcribed via the Transcription Chain in the Oral History portal, a project supported by Clarin[5]. We will prepare the hosting of this corpus in a CLARIN Centre and generate the metadata according to an appropriate CMDI profile for oral history. This means that data created within a historical framework will be described in a format and structure that adheres to CLARIN methodology. To search in other oral history collections for interviews on Ravensbrück in English, German and Dutch, that match with the Bruzzone archive, we will use a profile of the five Italian narrators and take into consideration their background, the length of the interviews and the interview approach (chronological semi-structured interviews with probing for details). We will use the archival material identified in section 4 as a backlog.

In the expansion phase we will curate and enrich the collected data by adding transcriptions, time stamps at word level, and phonetic and suprasegmental information. We will also add annotations on e.g. the use of specific language, emotions expressed in non-verbal modes of communication (i.e., laugh, pauses and silences, breathing). Aside from exploring the content, we will also document the feasibility of the re-use of oral history data, within the framework of FAIR open data, taking into account the particularity of the interview as source of knowledge, and identifying technological and legal obstacles for the re-use and dissemination of such material. Overcoming these obstacles will hopefully increase the visibility of this type of data and foster the interest for cross-disciplinary and multilingual approaches to research with interviews.

References

Beccaria Rolfi, L., Bruzzone, A.M. 2020. *Le donne di Ravensbrück. Testimonianze di deportate politiche italiane*, Torino, Einaudi.

Beccaria Rolfi, L., Bruzzone, A.M. 2016. *Als Italienerin in Ravensbrück. Politische Gefangene berichten über ihre Deportation und ihre Haft im Frauen-Konzentrationslager*, Herausgegeben von Johanna Kootz, Metropol Verlag Berlin.

Calamai, S., Kolletzek C., Kelli, A. 2019. *Towards a protocol for the curation and dissemination of vulnerable people archives*. In: Skadina, I. & Eskevich, M.: Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018. Linköping University Electronic Press, Linköpings universitet: 28-38 https://ep.liu.se/ecp/159/003/ecp18159003.pdf.

Draxler, C., Van den Heuvel, H., Van Hessen, A., Calamai, S., Corti, L., and Scagliola, S. 2020. A CLARIN *Transcription Portal for Interview Data. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020)*. pp. 3346-3352 http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.411.pdf

Hogervorst, S. *Onwrikbare Herinnering. Herinneringsculturen van Ravensbrück in Europa*. 2010. Uitgeverij Verloren, Hilversum.

Portelli, C. 1981. On the peculiarities of oral history. *History Workshop Journal*, 12: 96-107. https://doi.org/10.1093/hwj/12.1.96

Scagliola, S., Corti, L., Calamai, S., Karrouche, N., Beeken, J., Van Hessen, A., Draxler, Chr., Van den Heuvel, H., Broekhuizen, M., and Truong, K.. 2020. *Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow.* In: Simov, K., & Eskevich, M.: Selected Papers from the CLARIN Annual Conference 2019 Leipzig, 30 September - 2 October 2019. Linköping Electronic Conference Proceedings 172:15, 126-136. https://doi.org/10.3384/ecp2020172.

[5] https://www.phonetik.uni-muenchen.de/apps/oh-portal/