CLARIN-PLUS workshop: 
"Exploring Spoken Word Data in Oral History Archives"

Overview of CLARIN infrastructure & language technologies

CLARIN Data, Services and Tools,

Dieter van Uytvanck,
CLARIN ERIC
CLARIN Data, Services and Tools

Dieter Van Uytvanck
Technical Director CLARIN ERIC
dieter@clarin.eu
2016-04-18

Exploring Spoken Word Data in Oral History Archives
Oxford
CLARIN?

- **Common Language Resources and Technology Infrastructure**
- Research Infrastructure for the **humanities and social sciences**
- Provides easy and sustainable access for scholars
  - to **digital language data** (in written, spoken, video or multimodal form)
  - to **advanced tools** to discover, explore, exploit, annotate, analyse or combine them
Language resources: more than linguistics

Source:
- doi:10.3133/pp1707
CLARIN centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres** ([http://clarin.eu/centres](http://clarin.eu/centres))
Organisation CLARIN

- European (ESFRI) Research Infrastructure
- ERIC since 2012
- Landmark since 2016
- **Members:**
  - Austria • Bulgaria • Czech Republic • Denmark • Dutch Language Union • Estonia • Finland • Germany • Greece • Italy • Lithuania • Netherlands • Norway • Poland • Portugal • Slovenia • Sweden • United Kingdom (observer)
Benefits for countries

- Access to the CLARIN Infrastructure, i.e. to all CLARIN language resources and technology services
- Access to expertise via the CLARIN Knowledge Sharing Infrastructure
- Embedding in the humanities research community, with access to the same data
- Better visibility of their language, their research results, their resources and their cultural heritage
- Opportunities
  - for cross-lingual and -cultural research
  - to participate in research projects in which CLARIN ERIC participates as a beneficiary
The 33 CLARIN centres
Our services for researchers

- Concrete and usable services
- All available via [http://clarin.eu/services](http://clarin.eu/services)
  (+ a whole set of technical services behind the scenes)
Our services for researchers

Depositing & Archiving
Our services for researchers

Virtual Language Observatory
SEARCH

interviews

SEARCH RESULTS

27869 results

Collection

All values in this facet

Search:

Sort by Name

Only show values that occur at least 2 times

Academia Sinica Balanced Corpus of Modern Chinese (1)
African Language Materials Archive (1)
Archive of the Indigenous Languages of Latin America (7)
Bavarian Archive for Speech Signals (BAS) (414)
Berliner Wendekorpus (1)
Center of Estonian Language Resources (2)
CLARIN Centre (969)
COLlections de CORpus Oraux Numeriques (CoCoON ex-CRDO) (259)
European Language Resources Association (5)
Hamburger Zentrum für Sprachkorpora (HZSK) (1)
Hamburger Zentrum für Sprachkorpora (HZSK) (563)
LINDAT / CLARIN Data & Tools (1)
LRT + Open Submissions Data & Tools (15)
Meertens Collection: Diversity in Dutch DP Design (DIDDD) (1)
Meertens Collection: Dynamische Fonologische en Morfologische Atlas van de Nederlandse Dialecten (GTRP) (1)
Meertens Collection: Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND) (1)
Meertens collection: Liederenbank (53)
Meertens collections: PILNAR (15)
Multimodal Learning and teaching Corpora Exchange (1)
Nederlands Instituut voor Beeld en Geluid Academia collectie (12311)
Oxford Text Archive (4)

NARROW DOWN

Use the categories below to limit the search results to those matching the selected value(s).

+ LANGUAGE

- COLLECTION

Nederlands Instituut voor Beeld en Geluid Academia collectie (12311)
TLA: DoBeS archive (3344)
UBU Clarin Set (1558)
TalkBank (1553)
TLA: Acquisition (1073)
TLA: Donated Corpora (972)
CLARIN Centres (969)
TLA: Language and Cognition (938)
TLA: MPI CGN (816)
TLA: MPI für Bildungsforschung (740)

+ RESOURCE TYPE

+ COUNTRY

+ MODALITY

+ GENRE

+ SUBJECT
Our services for researchers

Federated Content Search
Umzüge und Erschienungen sind aus Tbilisi berichtet worden.

Die Delegation besuchte Moskau, Leningrad, Tbilisi, Chabarowski, Irkutsk und Nachodka.
Our services for researchers

Web services and applications
Our services for researchers

Easy access to protected resources
Our services for researchers

Virtual Collections
Absolute spatial deixis and proto-toponyms in Kata Kolok

Name: Absolute spatial deixis and proto-toponyms in Kata Kolok
Type: extensional
Creation Date: 2014-09-26

Purpose: research
Reproducibility: intended
Persistent identifier: hdl:11372/VC-1001
Keywords: sign language, Kata Kolok

Creators
Person: Connie de Vos
Organisation: Max Planck Institute for Psycholinguistics
Website: http://www.mpi.nl/people/vos-connie-de
Role: Researcher

Resources

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal Article (fulltext) This paper presents an overview of spatial deictic structures in Kata Kolok, a sign language which is indigenous to a Balinese village community.</td>
<td>Resource</td>
</tr>
<tr>
<td>Footnote 3 - video Absolute versus absolute transpositional pointing signs</td>
<td>Resource</td>
</tr>
<tr>
<td>Footnote 4 - video COME-HERE-FROM-A and GO-FROM-HERE-TO-B</td>
<td>Resource</td>
</tr>
</tbody>
</table>
Our services for researchers

Consultancy
CLARIN technology pillars

- **Federated Identity** - letting users login to protected data and services with their own institutional username and password
- **Persistent Identifiers** - enabling sustainable citations of electronic resources
- **Sustainable repositories** - digital archives where language resources can be stored, accessed and shared
- **Flexible metadata and concept definitions** - to ensure semantic interoperability when describing language resources
- **Well-described and open protocols**, e.g.:
  - **Content search** - offering a search engine for a wide range of language resources
  - **Web service chaining** - giving users the possibility to freely combine language processing services
Centres and Services are not isolated islands but part of a well-integrated setup.
Seamless integration within CLARIN

- e.g. ELAN with WebLicht (tagging) and WebMaus (phonetic alignment)
- e.g. the Language Resource Switchboard
Seamless integration in the infrastructure landscape
While CLARIN is based on a federation of long-term archives and services, it is not a static infrastructure.

This workshop is an excellent opportunity for us to:

- listen to your needs
- learn from your experiences
- create additional bridges between the research infrastructure and potential users and data/service providers
Oral History collection database – some first feedback

- Would be nice to integrate it into the Virtual Language Observatory
  - Challenge: gathering comparable metadata from the different sources (and maintaining it!)
    - at least dublin core via OAI-PMH endpoint
    - reuse/create a specific CMDI metadata profile?
  - What to do with non-digitized resources?
    - at least provide contact information for interested parties
  - What to do with access restrictions?
    - provide at least detailed information about how to obtain access
    - suggest single-sign-on
  - Nice start at http://www.verteldverleden.org/?page_id=151
- Connect transcript search engines to the Federated Content Search?
Oral History collection database – some first feedback

- Most important: your ideas, suggestions and feedback!
Thank you for your attention!

For more details, please visit: www.clarin.eu

or feel free to contact me at: dieter@clarin.eu