Working with Digital Collections of Newspapers
Leuven, 19-21 September 2016
CLARIN Data, Services and Tools

Menzo Windhouwer
Dieter van Uytvanck
CLARIN ERIC
menzo.windhouwer@meertens.knaw.nl

Working with Digital Collections of Newspapers
2016-09-19
Leuven
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
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 • Hungary • Italy • Latvia • 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 34 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
newspaper

The text was recorded at Madison University in the 1960s. The text was recorded indoors.; The overall goal of the project is the documentation and preservation of the Hoocal language. The project therefore includes the following sub-projects: (1) (audio- and video-)recording, analysing, processing and archiving a rep...

Newspaper articles

The Norwegian Newspaper Corpus

Dynamic, web-based newspaper corpus; 700 000 000 ws and growing; multilabeled
Our services for researchers

Federated Content Search
Ein gewaltiger Bau ist auch das Palais der Brisbane Newspaper Company, welches allein über 100.000 Pfund Sterling kostete und sieben Geschosse aufweist.


Die Verbreitung der Interviews wurde durch das Manhattan Newspaper Syndicate besorgt.

Die Engländer finden außer den örtlichen Zeitungen ihre andere Kost, Lloyds Weekly und Reynolds’ Nusipepa = Brief, Schriftstück, Druckwerk (von newspaper, Zeitung); he savee look along newspaper = er kann lesen (er wissen schauen entlang Zeitung).
Our services for researchers

Web services and applications
Our services for researchers

Easy access to protected resources
Our services for researchers

Virtual Collections
**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)</td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
<tr>
<td>Absolute versus absolute transpositional pointing signs</td>
<td>Resource</td>
</tr>
<tr>
<td>Footnote 4 - video</td>
<td></td>
</tr>
<tr>
<td>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
Seamless integration within CLARIN

- Centres and Services are not isolated islands but part of a well-integrated setup

24/7 monitoring

language observatory

content search

centre registry
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
Digital Collections of Newspapers – some first steps

- Would be nice to integrate them (all) 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
- Connect search engines to the Federated Content Search?
Digital Collections of Newspapers – 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: menzo.windhouwer@meertens.knaw.nl