Using the data in the archive

Jacquelijn Ringersma

The Language Archive
Max Planck Institute for Psycholinguistics
A very rich archive
A very rich archive

Video clips, sound files & images

Described Corpus

Multimedia Lexicon

Typed Relations within the Lexicon

Annotated Media
Two warnings

Only metadata are open: for resources you need access rights from the owner of the data

Only well described resources can be found
1. Online IMDI browser
2. Metadata search
3. Viewing resources and ANNEX viewer
4. TROVA content search

5. Virtual Language Observatory
   GE overlay
   Facetted browsing
IMDI browser

http://corpus1.mpi.nl
Metadata search

http://corpus1.mpi.nl

Keyword search
Standard search
Advanced search
Metadata search

http://corpus1.mpi.nl
Find some resources in the DoBeS archive which are:

1. Spoken discourse with at least one consultant (2634)
2. Spoken discourse with at least two consultants (1575)
3. Spoken discourse with at least two consultants in Asia (36)
4. Or Spoken discourse with at least two consultants in a Face to Face conversation (391)
Metadata search
Viewing resources
Viewing resources

Viewing video (mpg, mpeg) and audio (wav)
Viewing resources

Viewing images (jpg/tiff)
Viewing resources

Viewing text files (pdf/txt)

ANNEX:
Viewing ELAN, Toolbox and Chat annotations
Content search

TROVA 🍀
TROVA: Content search

Three options:

Simple keyword search
Single layer search (in one annotation tier)
  but: Annotation/Over annotations/Within annotations
  and: case (in)sensitivity
  and: substring/exact match and regular expressions
Multiple layer search:
  complex searches over multiple layers
TROVA: Content search

Simple search
TROVA: Content search
TROVA: Content search

(Over and within) Annotation

Tier selection (speech vs. words)
Regular expressions

Examples:

[abc] = a, b or c
[^abc] = any character, but not a,b, or c
b[a-zA-z]ng matches ‘bang’ but not baang
X* = x zero or more times
X+ = X one or more times
X|Y = X or Y
^ = beginning of an annotation, $ is end of an annotation
Regular expression:

`[^n]g$` finds all ending ‘g’, but not ‘ng’
Virtual Language Observatory

Welcome to the Virtual Language Observatory
From here you can explore the world of language resources and technology from different perspectives:

- Travel around the Virtual Language World (requires Google Earth)
- Navigate through the CLARIN Language Resource inventory (with IMDI archive, OLAC, ELRA and CLARIN data) or Language Tool inventory (combining CLARIN tools and DFKI NLP Software Registry data) using a faceted search
- Browse through the CLARIN catalogue of harvested metadata (requires a Java plugin)
- Have a look at the CLARIN Language Resource inventory or the Language Tool inventory

Are your resources still missing here? Contact us about how to incorporate them

www.clarin.eu/vlo
Virtual Language World

Google Earth overlay:

• Geographic navigation: approach for novice users
• Google Earth is a popular, freely available tool
• KML format is widely used and easily convertible
Virtual Language World

Place marks for
- linguistic archives
- language sites
- entry point for sets of resource bundles
place marks can be enriched with introductory texts, photos and direct links to the MPI archive
Facetted browser

![Facetted browser example](image)

This image demonstrates a facetted browser interface, typically used in the CLARIN Virtual Language Observatory - Resources. It allows users to search and filter resources based on various categories such as origin, continent, country, language, organisation, genre, and subject. Each category has a list of subcategories with corresponding numbers, indicating the size of the data collection under each category. For example, under the origin category, there are entries for Asia and South America with respective counts of 2765 and 1265. This type of interface is useful for users to navigate and find specific resources efficiently.
Facetted browser

www.clarin.eu/vlo

Find some resources in the DoBeS archive which are:

1. Spoken discourse with at least one consultant (2634)
2. Spoken discourse with at least two consultants (1575)
3. Spoken discourse with at least two consultants in Asia (36)
4. Or Spoken discourse with at least two consultants in a Face to Face conversation (391)
Facetted browser

www.clarin.eu/vlo

Find some resources in the catalogue which are:

1. Personal anecdotes recorded in South-America
2. Telephone conversation recordings in Nepal

Open the metadata files in the IMDI browser and check the full content of the metadata files
Speech community portals

Dane-zaa Community Portal

This page is created for the members of the Dane-zaa community to facilitate the use of the archive collected by the DobBeS team together with the elders.

Dane-zaa Community Portal

Stories  Learn about ...  Materials  The Archive

Personal history  Drum  Movies  Searching
Traditional stories  Food and cooking  Clickables  Navigating
Animals  Handgames  Dictionary  Download
Place  Horses  Phrasebook  Tours
Moccassins  Calendar  Google Earth
Moosehide  Alphabet  Studies
Preparing meat  Posters

The Language Archive
Summary

- browser
- metadata search
- Portals
- facetted browsing
- TROVA
- ANNEX viewer
- GE overlay