The Language Archive at the Max Planck Institute for Psycholinguistics

Alexander König
(with thanks to J. Ringersma)
1. The Language Archive
   • Why Archiving?
   • Stats and Facts
   • Archiving Software

1. Archive Exploitation
   • The IMDI Browser / Metadata search
   • Viewing resources and ANNEX
   • TROVA content search
   • Virtual Language Observatory
Why archiving?
Misconceptions about archiving

1. Your stuff is buried here and gone forever
Misconceptions about archiving

2. Other linguists will take advantage of your hard work and take away your good ideas
Why Archiving?

But the actual truth is that:

“The coolest thing to do with your data will be thought of by someone else”
Why Archiving?

Is there a danger that we lose digital data?

YES,
UNESCO: 80% of our recordings is endangered
How much of your data and files on the notebook is organized, backed-up?
How long can media and formats be accessed?
Why Archiving?

Correct conceptions about archiving

1. It requires discipline
2. It creates a bit of techno noise

The rewards:

Long term preservation and access

Different ways of presenting your data are possible
The repository of the MPI contains different types of linguistic material: the DOBES endangered languages archive, the ESF second Learner corpus, the Dutch Spoken National Corpus, corpora of the different research groups at the MPI, e.g. language documentations of the language and cognition research group, data of the acquisition group, and various gesture and sign language corpora.

The archive covers more than 637,615 objects, mostly organized in sessions that are described with the XML-based metadata description format IMDI. Mostly, these sessions contain digitized audio/video signals and layers of annotations. In general the access to these resources is limited and can be made available upon request.
Current State of the Archive

- Metadata: 144,729 files, 1.7 GB
- Audio Data: 100,416 files, 6.2 TB
- Video Data: 93,126 files, 17.0 TB
- Annotations: 77,959 files, 13.9 GB
- Lexica: 86 files, 22.7 MB
- Other: 103,722 files, 8.7 TB
- Total: 520,000 files, ~33 TB

- Included formats:
  - Text (UNICODE): txt, XML, HTML, Chat, Shoebox, pdf
  - Media: Wav, Mpeg1, Mpeg2, Mpeg4 (for streaming), jpg, tiff
- A persistent Unique Resource Identifier for each archive object
- We keep multiple copies.
Stats and Facts

- Multimedia Lexicon
  - Described Corpus
  - Photos
  - Video Clips
  - Annotated Media
  - Typed Relations within the Lexicon

The Language Archive

Stats and Facts
Archiving Software

Clear principles: Data organization and access infrastructure

**IMDI**  Metadata, Arbil
Browser and search

Coherent, consistent and persistent: Data management

**LAMUS**  Checks the content of the files, and file type check
Assigns a persistent identifier to the uploaded file
Allows the creation of corpus structures
Web based, easy to use

Safe access: Data access rights and protection

**AMS**  All metadata in the archive is open
All resource access can be controlled by AMS (web based)
Users remain the owners and stay in control of the access
Setting of licenses and code of conducts
1. Online IMDI browser
2. Metadata search
3. Viewing resources and ANNEX viewer
4. TROVA content search
5. Virtual Language Observatory
   • Google Earth Overlay
   • Facetted Browsing
1. Community Portals
IMDI browser

http://corpus1.mpi.nl
Metadata search

http://corpus1.mpi.nl

Keyword search
Standard search
Advanced search
Metadata search

http://corpus1.mpi.nl
Metadata search

Advanced search

within 1 selected corpus: DeBeS archive (14374 sessions/catalogues)

Add constraint  Delete

Search  Save  Print  Content search

391 matches within 14374 selected sessions/catalogues found.

The Language Archive
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

Content Navigation and Search

How To Facilitate Online Content Exploration

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
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’
TROVA: Content search
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
Speech community portals
Summary

browser

metadata search

TROVA

ANNEX viewer

GE overlay

Portals

facetted browsing