Using the data in the archive

Jacqueline Ringersma

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

Described Corpus

Multimedia Lexicon

Typed Relations within the Lexicon

Video clips, sound files & images

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
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

![Metadata search interface](image_url)
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: 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

Domain: MPJ CLG

EAF (12767)

Find

groter dan

#hits : 12

#annotations with a hit : 12

Progress

Cancel

Hit 1 - 10 of 12 hits

ze waren niet groter dan van mijn pols tot de top van mijn middelvinger. ze stonden op een

fen maar moesten klark die net binnen kwam wij hebben de hele morgen samen gemaakt. Mark zat al in groep zes en zeide zich haast wat groter dan zijn zus. 'kinderen voor ik hier ik er is hard gewerkt kaste moe
ter en daar waren wasloeken. en er hoorde ook toe bij de camping. gaan we daar doordring vandaar vroeg Mark. ja dat is best

ze gaf Mangalo de klokkelein. ga je die aangenaam niet ongezagen? waarom zou ik? het is jeur werkzaak; geduig huwelijk aanvaarljijk maar direct je op toeval en geluk of hardlocht en romantiek dan is de kans op een misluk of tegenig huwelijk groter dan vrijwel 80 procent. de huidde kan verduisteren meer dan avenhonde
der gij in bij zijn moeder en naar karmew. Mangalo had ook een gelo baak aan met een gele bloem in zijn knoo gat hij was een half hond groter dan de moeder van Pfffuun en hij was ook eigenlijk voor het vooral niet Pfffuun va

het aantal verpleeghuisen dat wegens het personeelskort en de tekorten een zomerstop inkoest is groter dan vorig jaar. dat heeft de vereniging van verpleeghuisen Aardes bij

nieuws geboden van één commissie zevens miljard gulden. het elektronicaconcern had al gewaarschuwd voor rode cijfers maar het verlies is wel een heel groter dan was verwacht. vooral bij de productie van chips voor computers en

ja hoe groot is 1 2 of 3 nog nooit gekeur? 1s 2s 3s is het best schijnt groter te zijn van Nederland. nog groter dan de lading, maar dat is goed ik eigenlijk niet voor volgens i

die nip jeker ook twee meter hoog want Raam is voeger aangevallen door een haan en naar zijn idee was die naam ook echt drie koppen groter dan hij en die heeft 'm ervan meer geplot ja en gangen die...}


eenheid dan veel mensen hadden gedacht en 1x gaat dan koeriers bij de standaard magistratuur de hedendaagse rechtspraak was groter dan ooit werd vermoed. minister Winnie Sorgelap van Justitie zag zo

chronisch tot zwaar depressieve patiënten en de kansen toenomen dat al evenal met hondert procent bij zinnen zijnde mensen heel groot is groter dan uh... ja 322 als we 1 toch over zijn fechten 322 ja.
TROVA: Content search

Domain: MPJ_CGN

Find (12767)

#hits: 12
#annotations with a hit: 12
#annotations investigated: 4931050

Options

Show Frequency View
Show Selected Hit in Transcription
Show Selected Hit in Corpus Tree
Show Info Balloons
Context Size
Change Font
Font Size
Save Hits

Hit 1 - 12 of 12 hits

1. naarm was zo kleiig en droog
2. fen maar misschepart klank die meer
3. ter en daar waren watlonk
4. ze gaf Marga
5. gedacht... huwelijksaantrekking
6. der giben, maar zijn moeder en
7. onles geboden van één kermis zo
8. Ja, hoe
9. s die mij jeker ook twee meter hi

Show Selected Hit in Transcription
Show Selected Hit in Corpus Tree
Show Info Balloons
Context Size
Change Font
Font Size
Save Hits
TROVA: Content search

(Over and within) Annotation
Tier selection (speech vs. words)
Regular expressions

Examples:

. = any character

[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’
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 INDI 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

CLARIN Virtual Language Observatory - Resources

Powered by Flamenco

Show tooltips previews of subcategories

ORIGIN
- plaza (7087)
- mcCrindell (3286)
- ted (2844)
- sigm (1590)
- self (1284)
- signlanguage (2758)
- math (2122)
- 1947 (1618)
- http (1521)
- nli (917)
- mmm

CONTINENT
- Europe (5471)
- Asia (1064)
- South-America (7346)
- North-America (5227)
- Oceania (2897)
- Australia (2017)
- Africa (9059)
- Middle-America (1268)
- Unknown (31)

COUNTRY
- Netherlands (20476)
- Denmark (18204)
- Germany (15054)
- France (14387)
- Japan (7995)
- Belgium (7946)
- Turkey (2993)
- United States (2072)
- Belgium (2089)
- France (2794)
- Mexico (2735)
- Canada (2085)
- United States (2072)

LANGUAGE
- English (35)
- Dutch (15195)
- German (14501)
- French (4226)
- Japanese (4182)
- Swedish (4146)
- Undetermined (3655)
- English (35)
- Dutch (15195)
- German (14501)
- French (4226)
- Japanese (4182)
- Swedish (4146)
- Undetermined (3655)

ORGANISATION
- Max Planck Institute for Psycholinguistics (1849)
- Language Archive (2333)
- University of Manchester
- School of Languages
- University of Bremen (1390)
- University of Leipzig (1335)
- Max Planck Institute for Evolutionary Anthropology
- Department of Linguistics (1390)
- University of California (1442)
- Language Archive - University of Dijon (1442)
- Babel University, Roehm (784)

GENRE
- Discourse (54370)
- spontaneous speech (5865)
- interview (2212)
- simulacrum (1660)
- dialogue (12329)
- narrative (1197)
- simulacrum (1139)

SUBJECT
- language description (12428)
- phonology (2708)
- semantics (3483)
- morphology (2014)
- speech analysis a speech analysis sheet via telephone (1956)
- more...
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 Sout-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 DoBeS team together with the elders.

<table>
<thead>
<tr>
<th>Stories</th>
<th>Learn about …</th>
<th>Materials</th>
<th>The Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal history</td>
<td>Drum</td>
<td>Movies</td>
<td>Searching</td>
</tr>
<tr>
<td>Traditional stories</td>
<td>Food and cooking</td>
<td>Clickables</td>
<td>Navigating</td>
</tr>
<tr>
<td>Animals</td>
<td>Handgames</td>
<td>Dictionary</td>
<td>Download</td>
</tr>
<tr>
<td>Place</td>
<td>Horses</td>
<td>Phrasebook</td>
<td>Tours</td>
</tr>
<tr>
<td></td>
<td>Moccassins</td>
<td>Calendar</td>
<td>Google Earth</td>
</tr>
<tr>
<td></td>
<td>Moosehide</td>
<td>Alphabet</td>
<td>Studies</td>
</tr>
<tr>
<td></td>
<td>Preparing meat</td>
<td>Posters</td>
<td></td>
</tr>
</tbody>
</table>
Summary

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