



Using the data in the archive

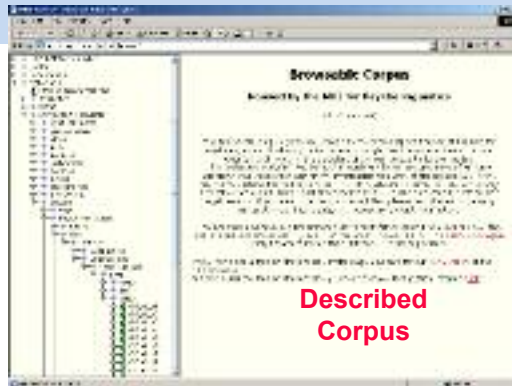
Jacqueline Ringersma

The Language Archive
Max Planck Institute for Psycholinguistics

A very rich archive



A very rich archive



Described Corpus

Multimedia Lexicon



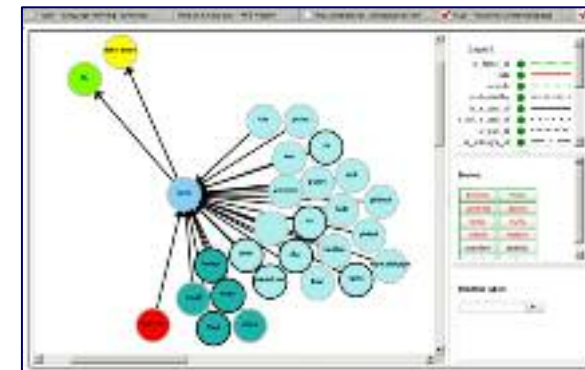
Video clips, sound files & images



Annotated Media



Typed Relations within the Lexicon





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

Content

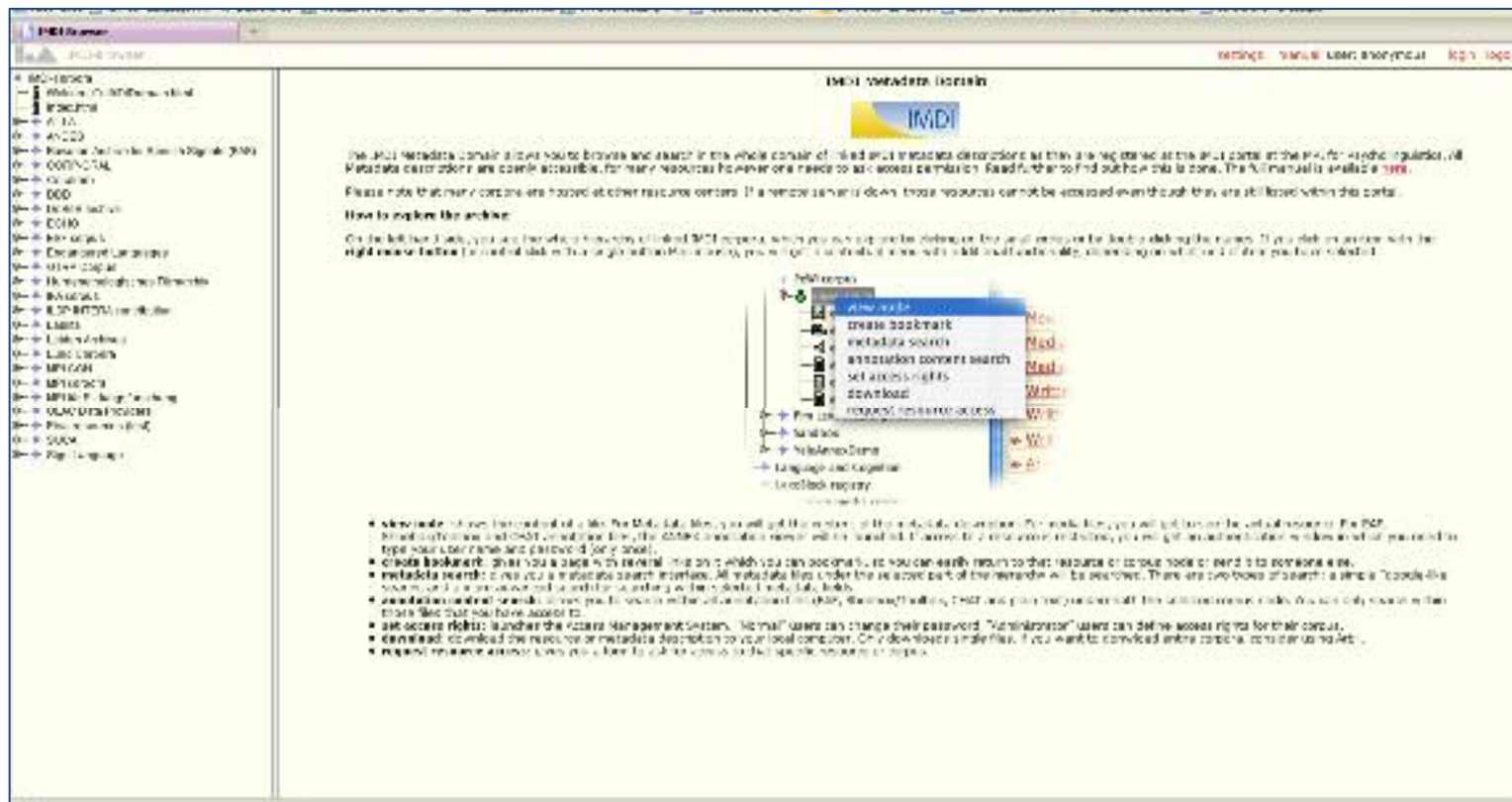


1. Online IMDI browser
2. Metadata search
3. Viewing resources and ANNEX viewer
4. TROVA content search

5. Virtual Language Observatory
 - GE overlay
 - Facetted browsing

IMDI browser

<http://corpus1.mpi.nl>



The screenshot shows the IMDI browser interface. On the left is a sidebar with a tree view of corpora, including categories like 'All corpora', 'ANCO', 'ANCO2', 'ANCO3', 'ANCO4', 'ANCO5', 'ANCO6', 'ANCO7', 'ANCO8', 'ANCO9', 'ANCO10', 'ANCO11', 'ANCO12', 'ANCO13', 'ANCO14', 'ANCO15', 'ANCO16', 'ANCO17', 'ANCO18', 'ANCO19', 'ANCO20', 'ANCO21', 'ANCO22', 'ANCO23', 'ANCO24', 'ANCO25', 'ANCO26', 'ANCO27', 'ANCO28', 'ANCO29', 'ANCO30', 'ANCO31', 'ANCO32', 'ANCO33', 'ANCO34', 'ANCO35', 'ANCO36', 'ANCO37', 'ANCO38', 'ANCO39', 'ANCO40', 'ANCO41', 'ANCO42', 'ANCO43', 'ANCO44', 'ANCO45', 'ANCO46', 'ANCO47', 'ANCO48', 'ANCO49', 'ANCO50', 'ANCO51', 'ANCO52', 'ANCO53', 'ANCO54', 'ANCO55', 'ANCO56', 'ANCO57', 'ANCO58', 'ANCO59', 'ANCO60', 'ANCO61', 'ANCO62', 'ANCO63', 'ANCO64', 'ANCO65', 'ANCO66', 'ANCO67', 'ANCO68', 'ANCO69', 'ANCO70', 'ANCO71', 'ANCO72', 'ANCO73', 'ANCO74', 'ANCO75', 'ANCO76', 'ANCO77', 'ANCO78', 'ANCO79', 'ANCO80', 'ANCO81', 'ANCO82', 'ANCO83', 'ANCO84', 'ANCO85', 'ANCO86', 'ANCO87', 'ANCO88', 'ANCO89', 'ANCO90', 'ANCO91', 'ANCO92', 'ANCO93', 'ANCO94', 'ANCO95', 'ANCO96', 'ANCO97', 'ANCO98', 'ANCO99', 'ANCO100'. The main content area displays the 'IMDI metadata browser' title and the IMDI logo. Below the logo, there is a paragraph of text explaining the browser's purpose and a note about metadata access. A context menu is open over the 'All corpora' tree view, showing options like 'create bookmark', 'metadata search', 'annotation comment search', 'set access rights', 'download', and 'request metadata access'. At the bottom, there is a list of instructions for users.

IMDI metadata browser

The IMDI metadata browser allows you to browse and search in the whole corpus of linked and metadata descriptions as they are registered at the MPI for Linguistics. All Metadata descriptions are openly accessible, for many resources however one needs to ask access permission. Read further to find out how this is done. The full manual is available [here](#).

Please note that many corpora are hosted at other resource centers. If a resource is unavailable, those resources cannot be accessed even though they are still listed within this portal.

How to explore the archive:

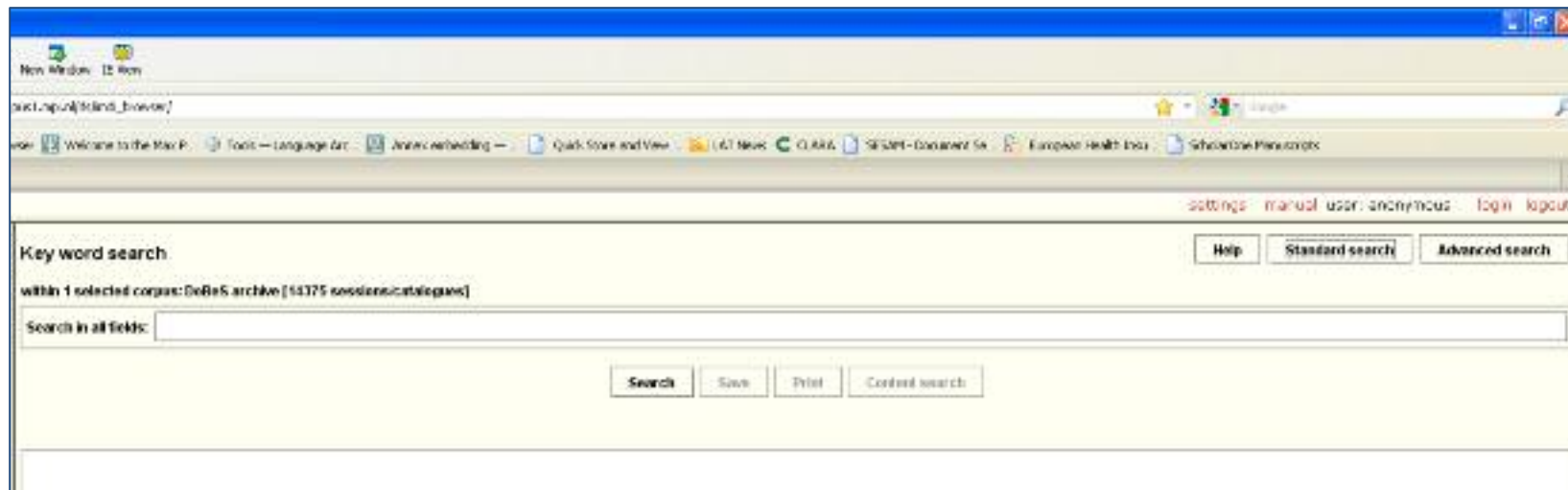
On the left-hand side, you can find the full hierarchy of linked MPI corpora, which you can explore by clicking on the small icons or by clicking on the names. If you click on a name with the right mouse button (or control click with the left button) a context menu will appear with the following options:

- view metadata: view the metadata of the selected corpus.
- create bookmark: create a bookmark for the selected corpus, so you can easily return to that resource or copy the URL to someone else.
- metadata search: view a metadata search interface. All metadata files under the selected part of the hierarchy will be searched. There are two boxes to search a simple "code-like" search and a more advanced search using regular expressions.
- annotation comment search: view a search interface for annotations and comments.
- set access rights: launch the Access Management system. "Normal" users can change their password. "Administrator" users can define access rights for their corpus.
- download: download the resource or metadata description to your local computer. Only download single files, if you want to download entire corpora, contact us via [mailto:corpus@mpi.nl](#).
- request metadata access: ask for access to the specific resource or corpus.

Metadata search



<http://corpus1.mpi.nl>



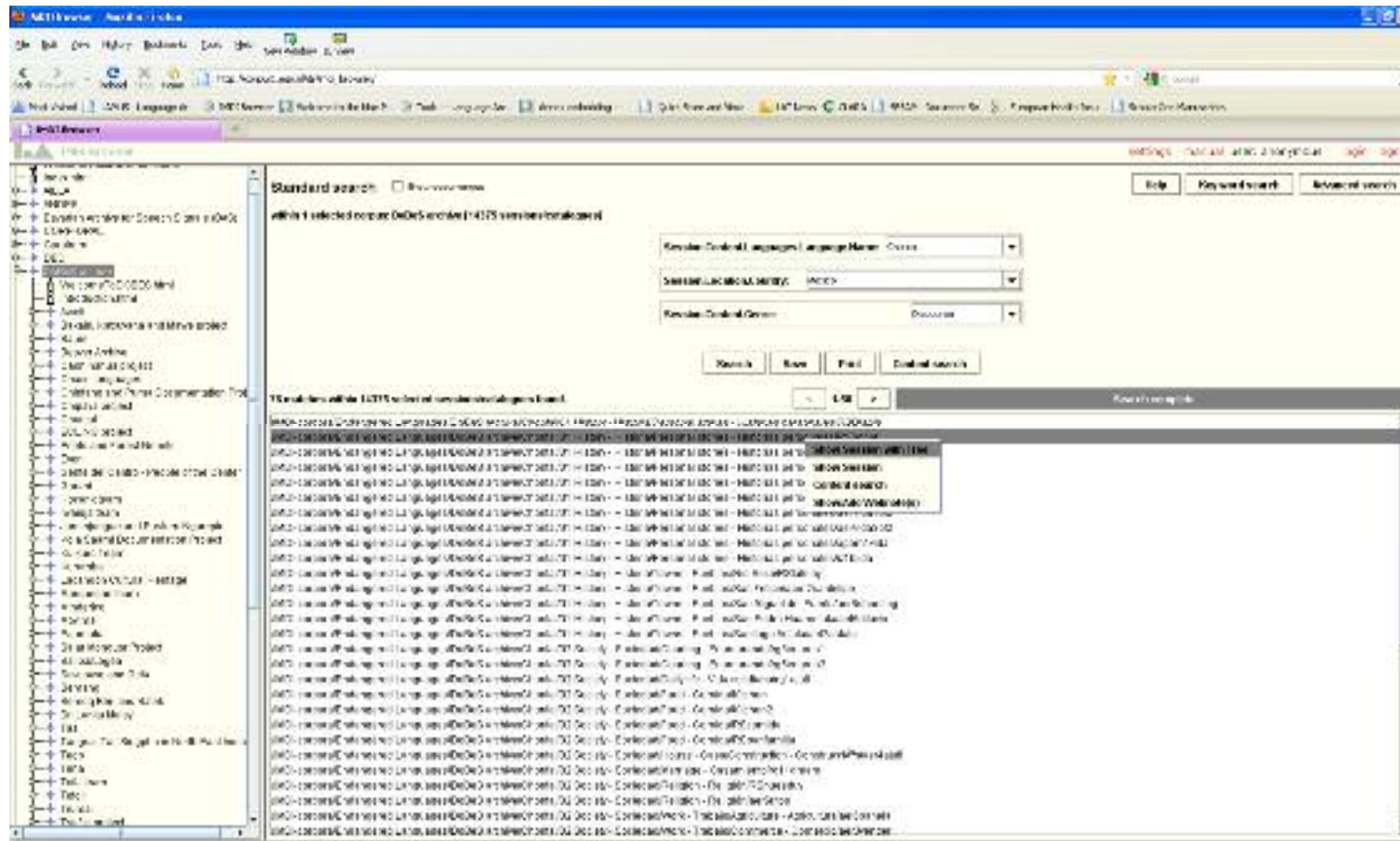
Keyword search

Standard search

Advanced search

Metadata search

<http://corpus1.mpi.nl>



The screenshot displays the MPI-Browser web application interface. The browser window title is "MPI-Browser - Search results". The address bar shows the URL "http://corpus1.mpi.nl". The page content includes a search interface with the following elements:

- Search Interface:** A "Standard search" section with a search input field containing "with 1 selected output 0 of 0 records (14375 items total) found". Below this are three dropdown menus: "Reverse Default Language: Language Name: Dutch", "Select Default Metadata: None", and "Reverse Default Query: Document". There are buttons for "Reset", "Keyword search", and "Advanced search".
- Search Results:** A table listing search results. The table has columns for "ID", "Language", "Document", "Date", "Time", "Speaker", "Gender", "Age", "Education", "Occupation", "Address", "City", "Country", "Region", "Postal code", "Phone", "Fax", "Email", "Web", "Other". The results are sorted by "ID" and show a list of records with their corresponding metadata.
- Navigation:** A "Page 1 of 1" indicator and a "Search complete" message.

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

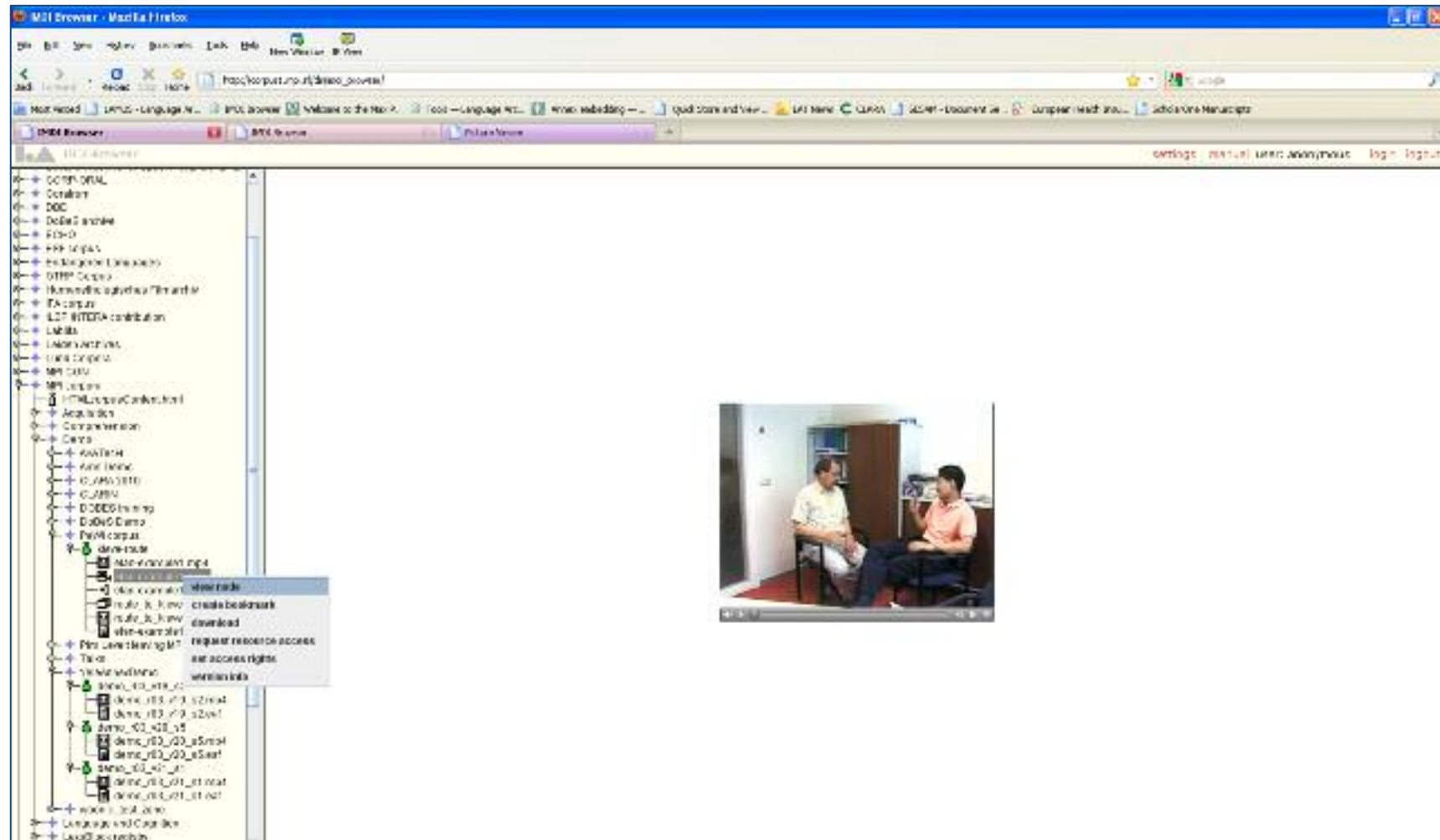
The screenshot displays the IMDI (International Metadata Database Interface) web application. The browser address bar shows the URL `http://corpus.lingpi.de/imdi-browser/`. The interface includes a navigation sidebar on the left with a tree view of metadata categories such as 'All Languages', 'Endangered Languages', and 'Linguistics'. The main content area is titled 'Advanced search' and shows a search for 'DeDeS archive' within 14374 selected corpora. Below the search criteria, there are buttons for 'Add constraint', 'Delete', 'Search', 'Save', 'Print', and 'Content search'. The search results section shows 391 matches with a 'Search console' button and a '1-90' pagination control. The results list various metadata entries with their IDs and titles, such as 'IMDI-corpora/Endangered Languages/DeDeS archive/From Language and Cultural Linguistic Data/El dhalaniars'.

Viewing resources



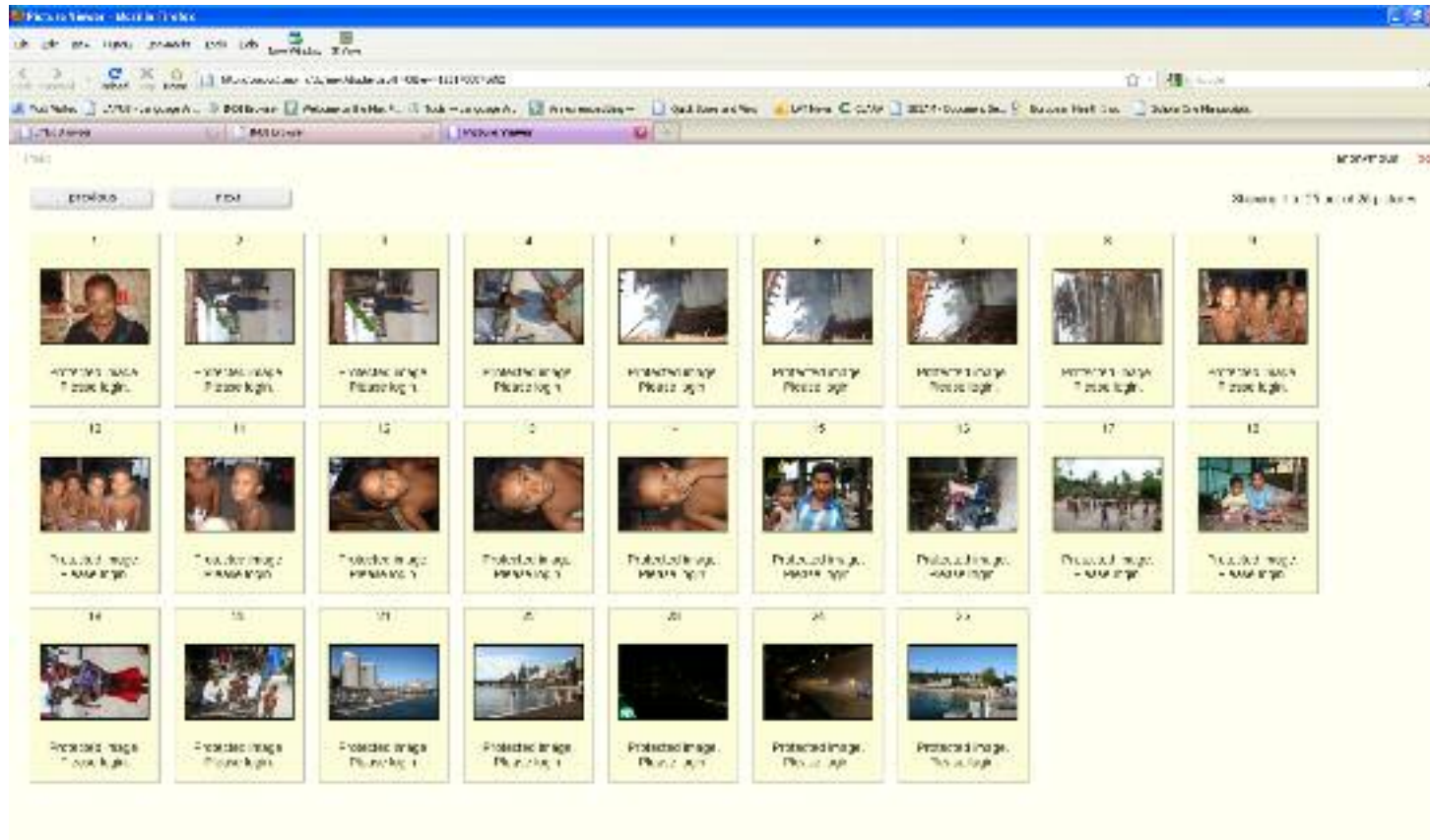
Viewing resources

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



Viewing resources

Viewing images (jpg/tiff)

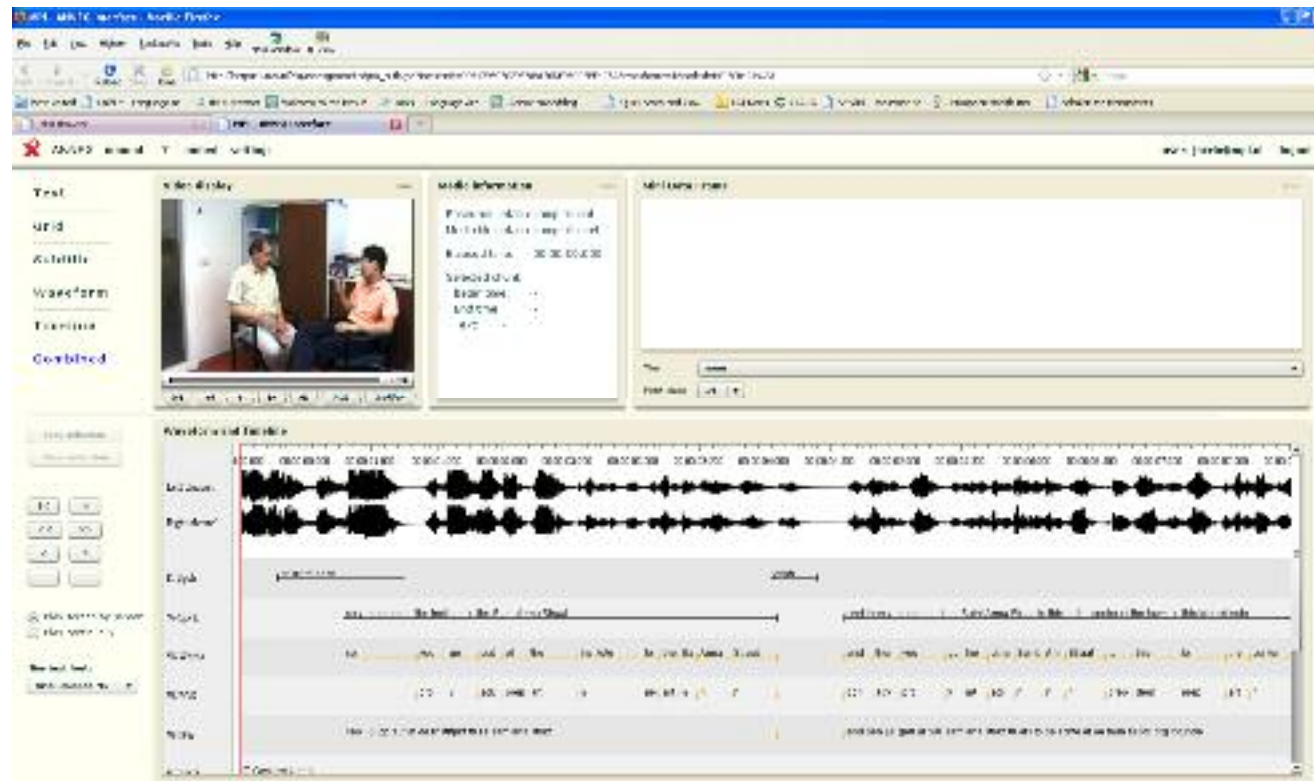


Viewing resources

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:

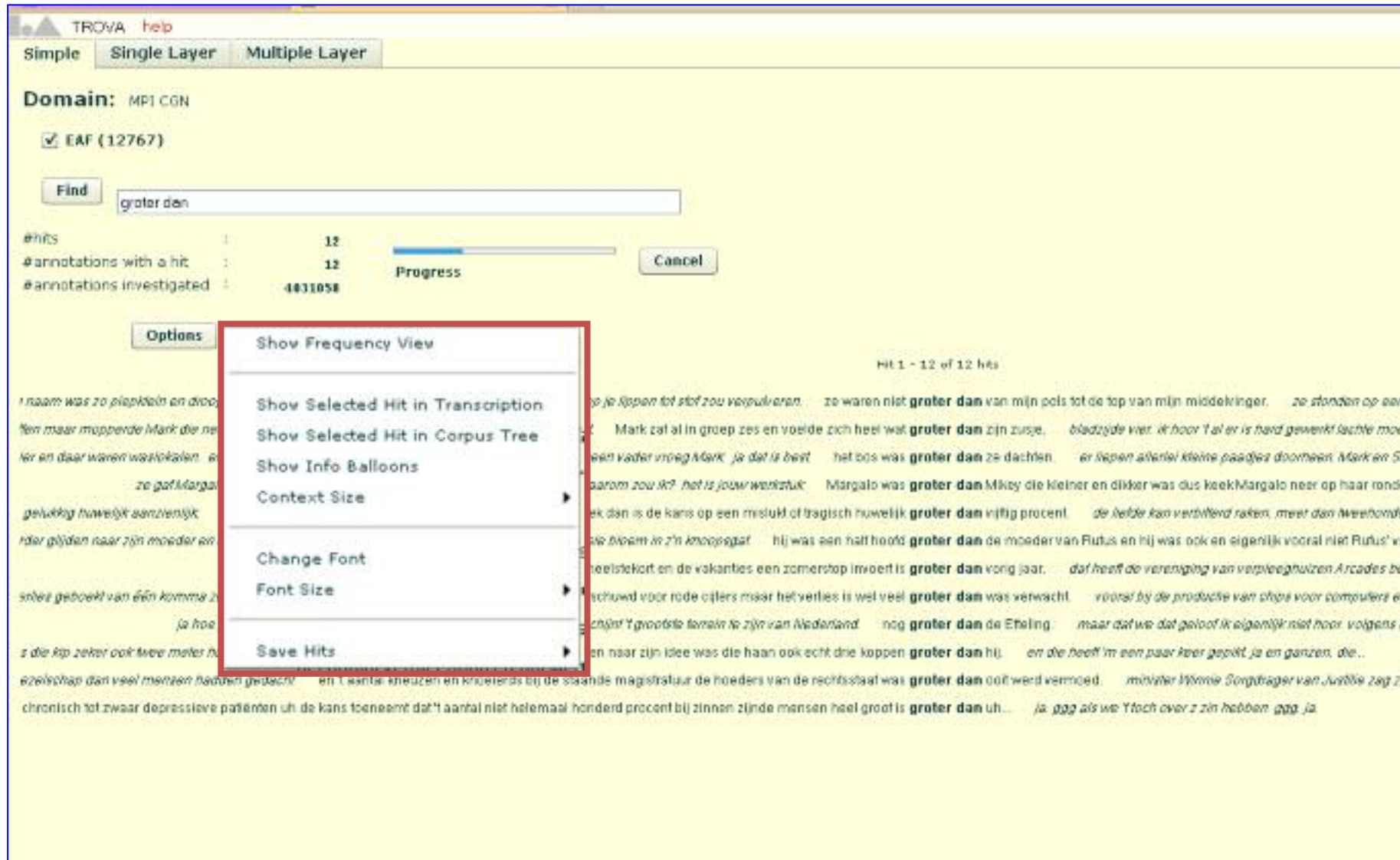
complex searches over multiple layers

TROVA: Content search

The screenshot shows the TROVA search interface. At the top, there are tabs for 'Simple', 'Single Layer', and 'Multiple Layer', with 'Simple' selected. Below the tabs, the domain is set to 'MPI CGN'. A search filter 'EAF (12767)' is checked. The search term 'groter dan' is entered in the search box. The results show 12 hits, with 12 annotations and 4831058 annotations investigated. A progress bar is visible. The search results are displayed as a list of text snippets, with the search term highlighted in bold. The first snippet is: 'naam was zo piepklein en droog dat je hem nauwelijks uit durfde te spreken uit angst dat hij op je lippen tot stof zou vergulveren. ze waren niet **groter dan** van mijn pols tot de top van mijn middelvinger. ze stonden op een...

Simple search

TROVA: Content search



The screenshot displays the TROVA search interface. At the top, there are tabs for 'Simple', 'Single Layer', and 'Multiple Layer'. Below these, the 'Domain' is set to 'MPI CGN'. A search box contains the query 'groter dan', and a 'Find' button is visible. The search results show 12 hits, with 12 annotations and 4831058 annotations investigated. A progress bar is shown next to the 'Progress' label. An 'Options' menu is open, listing several actions: 'Show Frequency View', 'Show Selected Hit in Transcription', 'Show Selected Hit in Corpus Tree', 'Show Info Balloons', 'Context Size', 'Change Font', 'Font Size', and 'Save Hits'. The search results are displayed in a list format, with the first hit highlighted. The text of the first hit is: 'op je lippen tot stof zou vergulveren. ze waren niet **groter dan** van mijn pols tot de top van mijn middelvinger. ze stonden op een...'. The interface also shows a 'Cancel' button and a 'Hit 1 - 12 of 12 hits' indicator.

TROVA: Content search



(Over and within) Annotation

Tier selection (speech vs. words)

The screenshot shows the TROVA search interface. At the top, there are tabs for 'Simple', 'Single Layer', and 'Multiple Layer'. The 'Domain' is set to 'MPI CGN'. A search history entry shows 'grater das: N-gram over annotations case insensitive exact match in Tier Type: Word'. The search mode is 'N-gram over annotations', 'case insensitive', and 'exact match'. The search term 'grater das' is entered in the search box. The search results show 11 matches, with 11 annotations with a hit and 1816586 annotations investigated. A progress bar is visible. The search results are displayed in a table with columns for 'grater das' and 'Tier Type: Words'.

grater das	Tier Type: Words
grater das	grater das
grater das	grater das
grater das	grater das

TROVA: Content search



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

TROVA: Content search

Regular expression:

[^n]g\$ finds all ending 'g', but not 'ng'

The screenshot shows the TROVA search interface. At the top, the search mode is set to "N-gram over annotations", "case insensitive", and "regular expression". The search query is "[^n]g\$". The results show 2063 hits, with 2063 annotations containing a hit and 89377 annotations investigated. A progress bar is visible. Below the search controls, there is a list of search results, each consisting of a snippet of text, a highlighted word, and a corresponding word from a dictionary or glossary.

Mode: N-gram over annotations | case insensitive | regular expression

Find: [^n]g\$

#hits : 2063
#annotations with a hit : 2063
#annotations investigated : 89377

Options

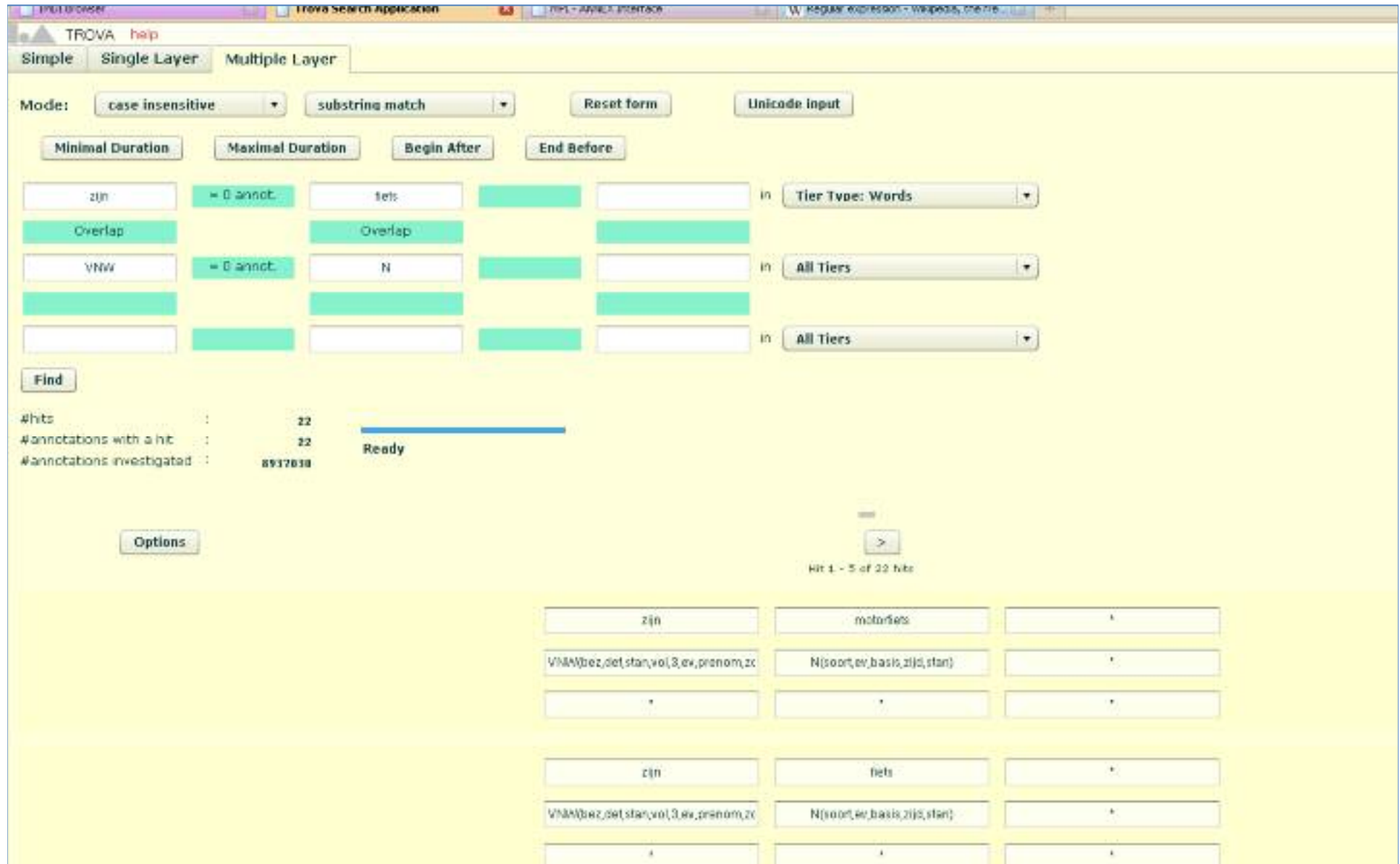
Progress

Cancel

Hit 1 - 15 of 2063 hits

Hans Bavinck 'i	oog	van de naald
Robert Puddington kwam	terug	van de dokter
toen hij haar	veertig	jaar eerder voor
had ontmoet was	zeg	je ik ga
herhaalde hij Susan	zeeg	neer op een
ze koken elkaar	ernstig	aan de dokter
er vorige week	nog	over weef je
over weef je	nog	't is gemakkelijker
kameel door 'i	oog	van een naald
naaste niet lie	genoeg	oh Bob zeg
genoeg oh Bob	zeg	zulke dingen loch
bijbel bijna elke	dag	je hebt een
je dan doen	vraag	zijn vrouw huverend

TROVA: Content search



The screenshot shows the TROVA search application interface. At the top, there are tabs for 'Simple', 'Single Layer', and 'Multiple Layer'. Below these are search options: 'Mode' (set to 'case insensitive'), 'substring match', 'Reset form', and 'Unicode input'. There are also buttons for 'Minimal Duration', 'Maximal Duration', 'Begin After', and 'End Before'. The search criteria are entered in a table with columns for 'Minimal Duration', 'Maximal Duration', 'Begin After', 'End Before', and 'Tier Type'. The first row shows 'zijn' and 'iets' with 'Tier Type: Words'. The second row shows 'VNW' and 'N' with 'Tier Type: All Tiers'. The third row is empty with 'Tier Type: All Tiers'. A 'Find' button is located below the search criteria. The results section shows '#hits: 22', '#annotations with a hit: 22', and '#annotations investigated: 8917838'. A progress bar is labeled 'Ready'. Below the results, there are 'Options' and '>' buttons. The bottom part of the screen shows a table of search results with columns for 'zijn', 'motorfiets', and 'iets', and their corresponding annotations.

Minimal Duration	Maximal Duration	Begin After	End Before	Tier Type
zijn	= 0 annot.	iets		Words
Overlap		Overlap		
VNW	= 0 annot.	N		All Tiers
				All Tiers

#hits : 22
#annotations with a hit : 22
#annotations investigated : 8917838

Ready

Options >

Hit 1 - 5 of 22 hits

zijn	motorfiets	*
VNW(bez,det,stan,vol,3,ex,pronom,zc	N(soort,er,basis,zijd,stan)	*
*	*	*
zijn	iets	*
VNW(bez,sef,stan,vol,3,ex,pronom,zc	N(soort,er,basis,zijd,stan)	*
*	*	*

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

Virtual Language World



place marks can be enriched with introductory texts, photos and direct links to the MPI archive

Facetted browser

Facetted Browsing - Resources

CLARIN Virtual Language Observatory - Resources Powered by Flamingo
Demonstrator with IMDI, DLAC, ELISA and CLARIN data (contact: viv@clarin.eu)

Show tooltip previews of subcategories

ORIGIN	
elac (70871)	enclanguage (2768)
mp Corpora (82884)	ebd (2122)
endangeredlanguages (10478)	Leipzig (1616)
esd (12767)	bip (1321)
las (7418)	sll (217)
lud (8198)	more...
sal (2054)	

CONTINENT	
Europe (54517)	Australia (2617)
Asia (11647)	Africa (2020)
North America (7346)	Middle America (1268)
North America (4227)	Unknown (31)
Oceania (2887)	

COUNTRY	
Netherlands (20676)	Bolivia (2858)
Germany (19404)	Australia (2826)
Sweden (2701)	France (2794)
Japan (1993)	Mexico (2733)
Malaysia (2946)	Canada (2003)
Turkey (2462)	more...
United States (2872)	

LANGUAGE	
English (26749)	Turkish (2768)
Dutch (19193)	Spanish (2609)
German (14951)	Undetermined (1458)
French (488)	Tzeltal, Teneca (1398)
Japanese (4183)	Arabic, Standard (1205)
Swedish (4146)	more...
Undetermined (2658)	

ORGANISATION	
Max Planck Institute for Psycholinguistics (38894)	German Research Foundation (DFG) (1890)
Max Planck Institute for Evolutionary Anthropology (1515)	University of Manchester, School of Languages, Linguistics and Culture (1240)
University of Cologne (1448)	University of Leipzig (1333)
IAS, IFA, Department of Italianistica - Università di Firenze (1442)	Max Planck Institute for Evolutionary Anthropology, Department of Linguistics (1890)
Boh University Bochum (784)	more...

GENRE	
Discourse (24070)	Music description (1122)
spontaneous speech (5845)	Singing (865)
interview (2213)	Conversation (702)
Stim., act-out (1849)	Elicitation (694)
dialogue (1228)	Unspecified, narrative (522)
narrative (1197)	more...
Stimul. (1129)	

SUBJECT	
language description (12428)	phonology (3706)
typology (7582)	semantics (2493)
sociolinguistics (7410)	phonetics (2662)
syntax (7335)	morphology (2614)
grammar tool (5400)	tools systems for a speechdat project sheet via toolchains (1956)
monologue about free topic (3929)	more...
lexicon (3905)	

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 DoBeS team together with the elders.

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
	Moccasins	Calendar	Google Earth
	Moosehide	Alphabet	Studies
	Preparing meat	Posters	

Summary

