LAT

Language Archiving Technology

an eHumanities infrastructure

http://www.lat-mpi.eu

Overview

- What is LAT?
  - History, context
  - Current state
- Case study: LAT and the corpus spoken Dutch
- Future evolutions

MPI for Psycholinguistics

- 4 main research groups:
  - Language Acquisition
  - Language Comprehension
  - Language & Cognition (“field linguists”)
  - Language Production
- Large technical group: support & software development

MPI digital language archive

- Started creating a digital archive about 10 years ago for a number of reasons:
  - Preventing “data cemeteries”: central, well organized data archive with rich metadata descriptions
  - Taking care of long-term preservation of data: continuous migration to latest storage technology, making use of “archivable” file formats, data distribution
  - Data access: Making data discoverable and accessible in various ways over the internet

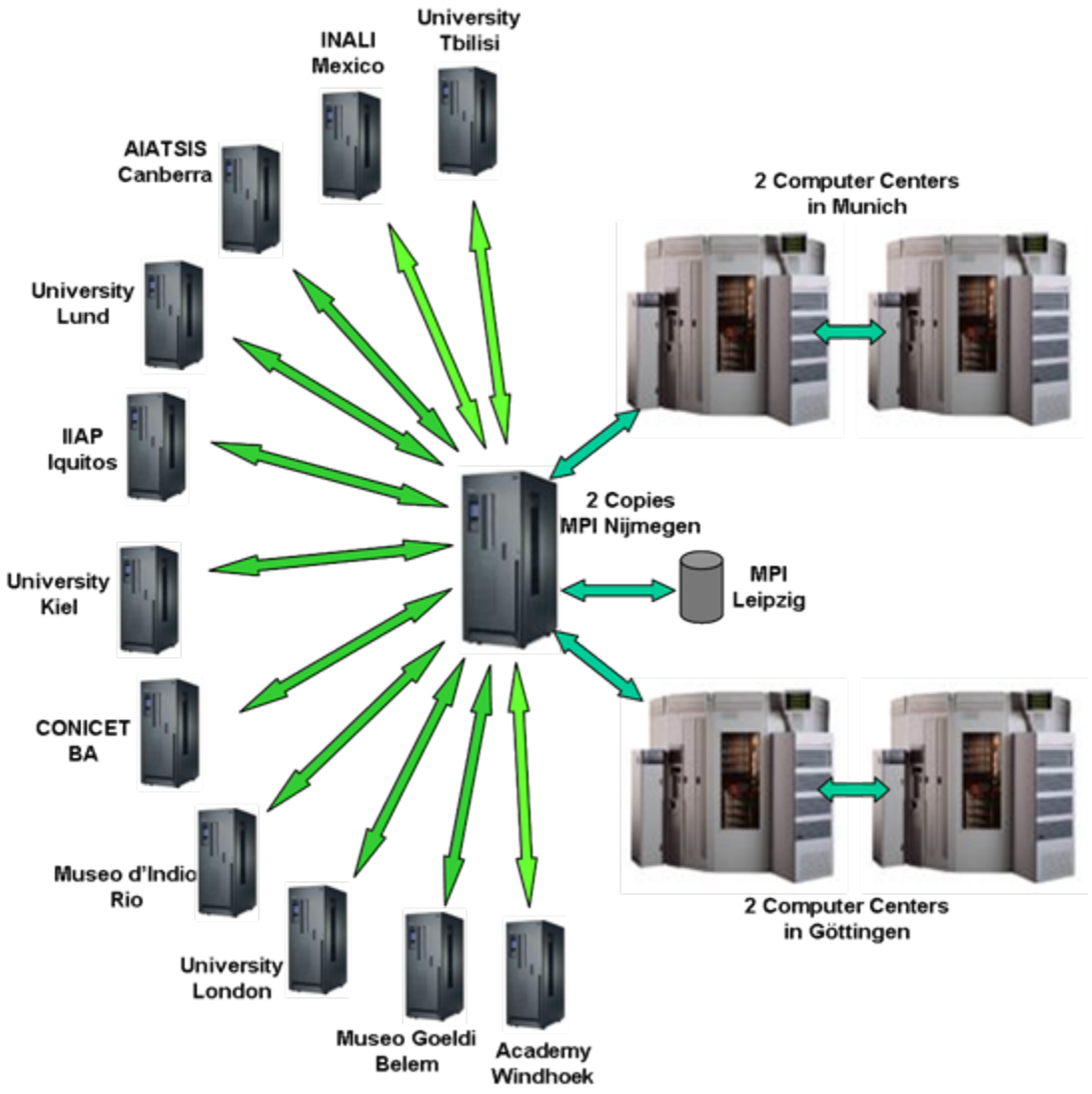
MPI digital language archive

- Contains a large number of languages:
  - Languages from all over the world studied by MPI field linguists
  - First and second language acquisition studies
  - Endangered languages documented for the VolkswagenStiftung DoBeS programme
  - Spoken Dutch corpus
  - Sign languages
  - etc.

MPI digital language archive

- Kinds of data:
  - video recordings
  - audio recordings
  - images
  - transcriptions/annotations
  - lexica
  - various other types of written documents (descriptions, grammars, ...)
- All described using IMDI metadata scheme

Data distribution



- at MPI about 45 Terabyte in total (20 TB in IMDI-described corpora)
- > 250.000 resources
- > 15.000 hours recording

- every 5 years new storage technology
- 4 copies at large computer centers

- synchronized regional archives
  - another copy
  - data back to regions
  - trust

Tasks of a digital archive

- long-term preservation
- giving access in the ***Live*** Archives sense
  - archive as a center of ORGANIZED information
  - archive as a center for continuous extension
  - archive as a center of collaboration and interaction
  - archive as a center for commentary and relation drawing (enrichment)
  - archive as a center supporting cross-corpus/language work

MPI “LAT” Archiving Framework & Tools

Preparation

Primary resources

ELAN / LEXUS / SYNPATHY Annotation, Lexicon & Syntax

Integration

IMDI: Metadata descriptions, data organization

LAMUS: Data Uploading and Management, Access Management

Archive

>20 Terabytes
>15.000 hours A/V

MPI “LAT” Archiving Framework & Tools

Preparation

Primary resources

ELAN / LEXUS / SYNPATHY Annotation, Lexicon & Syntax

Integration

IMDI: Metadata descriptions, data organization

LAMUS: Data Uploading and Management, Access Management

Archive

>20 Terabytes
>15.000 hours A/V

Utilization

IMDI / GIS: Metadata Browsing & Searching, Google Earth

ANNEX / LEXUS / TROVA: Complex Access via the Web

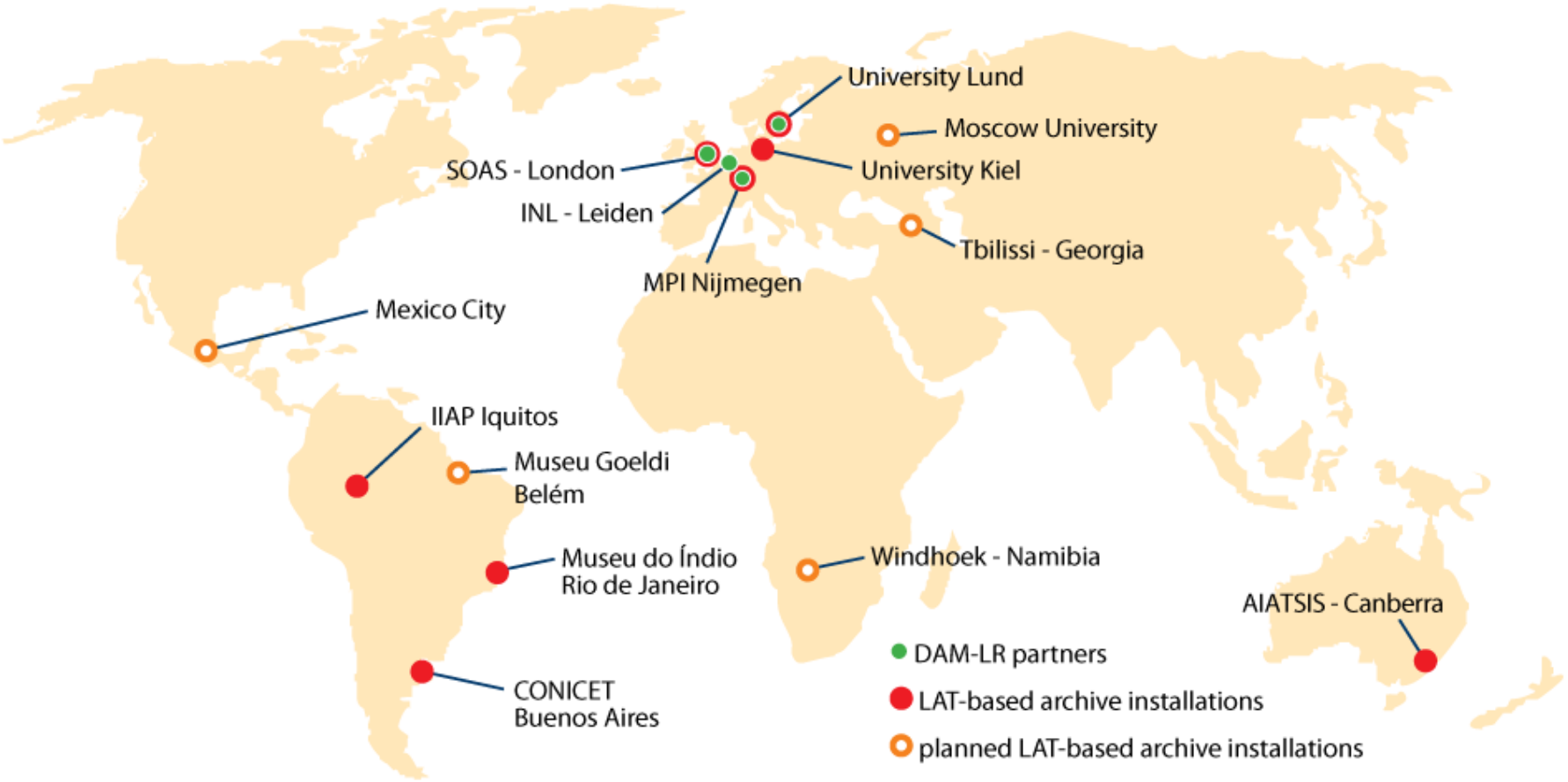
ADDIT / VICOS: Enrichments/Views

So what is LAT about?

- Openness:
  - Open standards
  - Open source (& VMware image)
  - Long-term preservation
- Compatibility:
  - Between all LAT applications
  - With outside world (import/export)
    - Praat
    - Shoebox / toolbox / ...
    - OAI-PMH / REST-based access
- Research-centered:
  - What are researchers looking for?
  - Cross-media: links between the archive and
    - paper/ electronic publications
    - websites

Grid of Language Archives

Distribution of MPI “LAT” archive installations



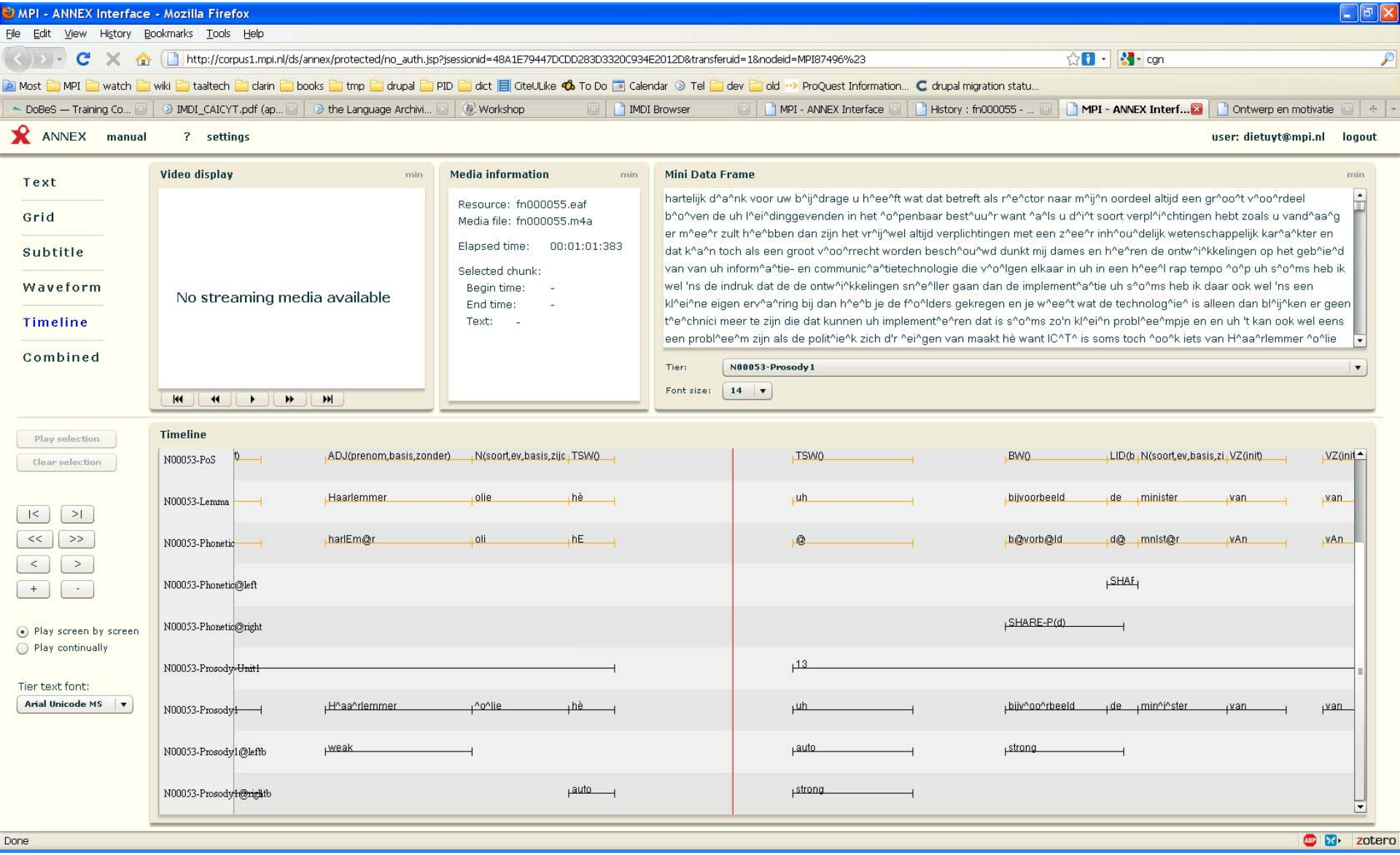
LAT and phonological corpora

- Example: Corpus Spoken Dutch (CGN)
  - Manually verified broad phonetic transcription: SAMPA-based
  - Prosodic annotation layer (Grover et al. 1998) :
    - strong break: ja dat weet ik **||** maar wanneer
    - weak break: dit is werkelijk on**|**ge**|**looflijk
    - vowel part of a prominente syllable: **^**i**^**k ben thuis
    - unusual lengthening of a sound: **%**ja**%**

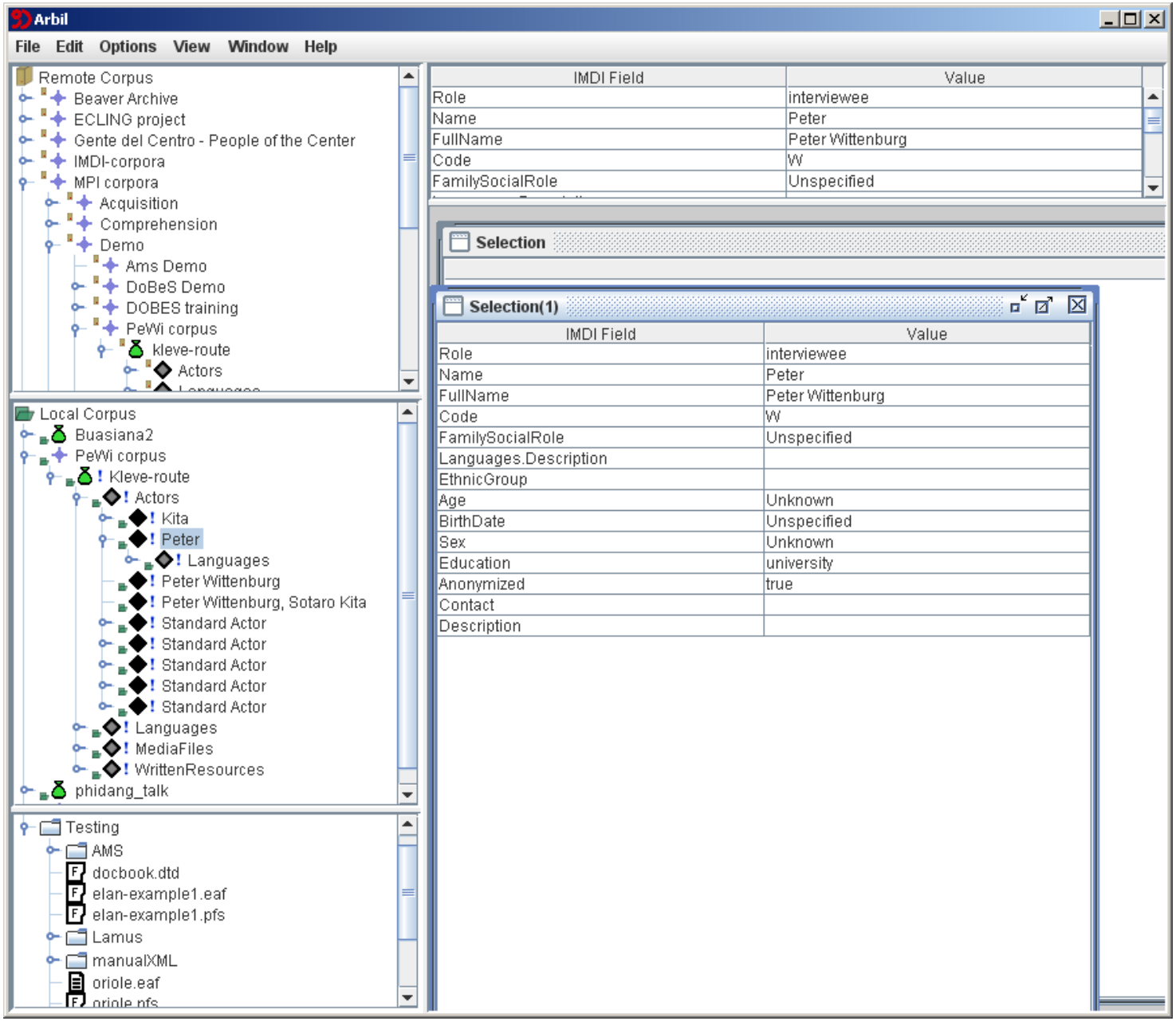
LAT and phonological corpora (2)

- Example: find words ending in shwa, followed by a word with the stress on a syllable with /aa/

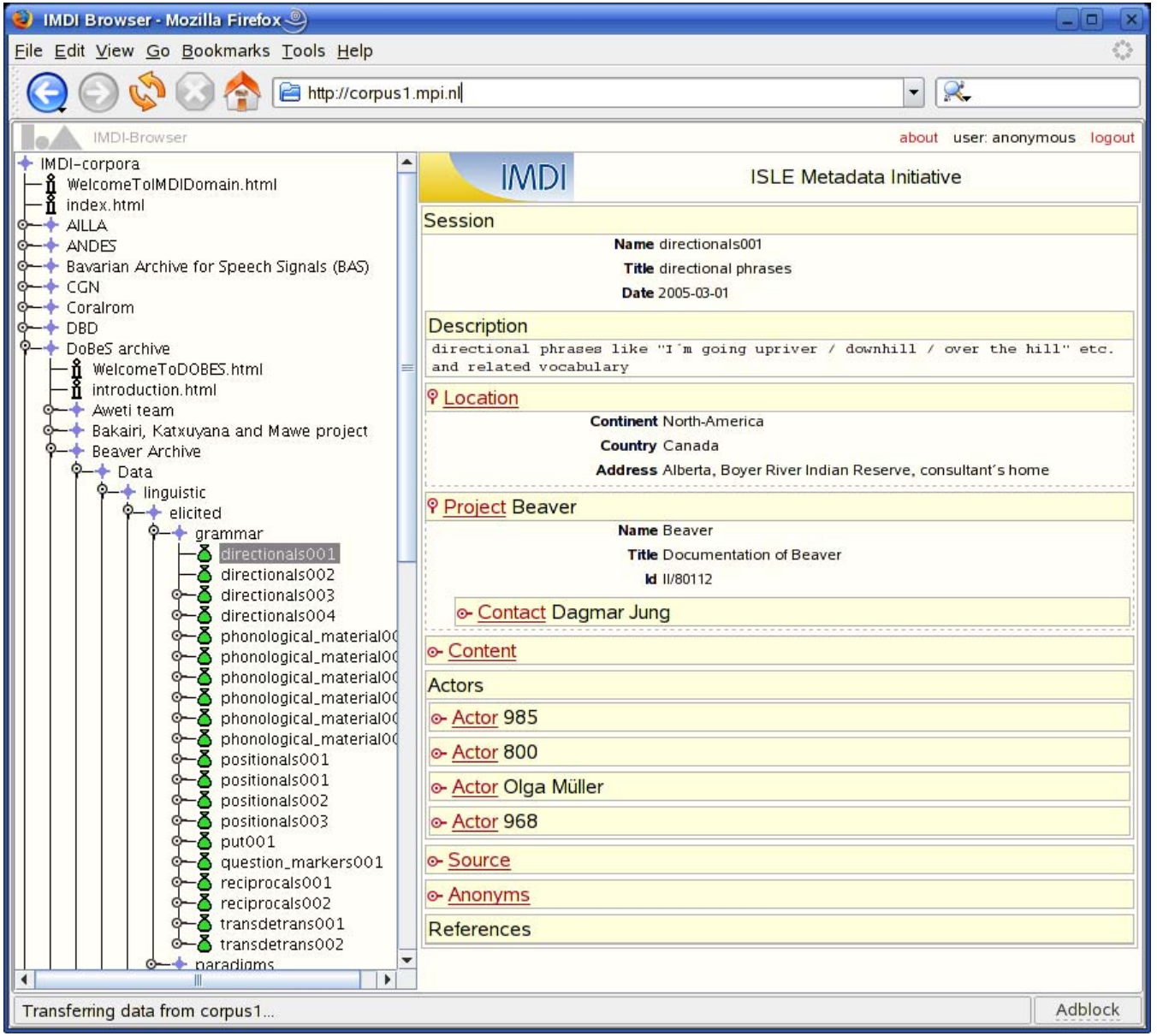
CGN: example annotation



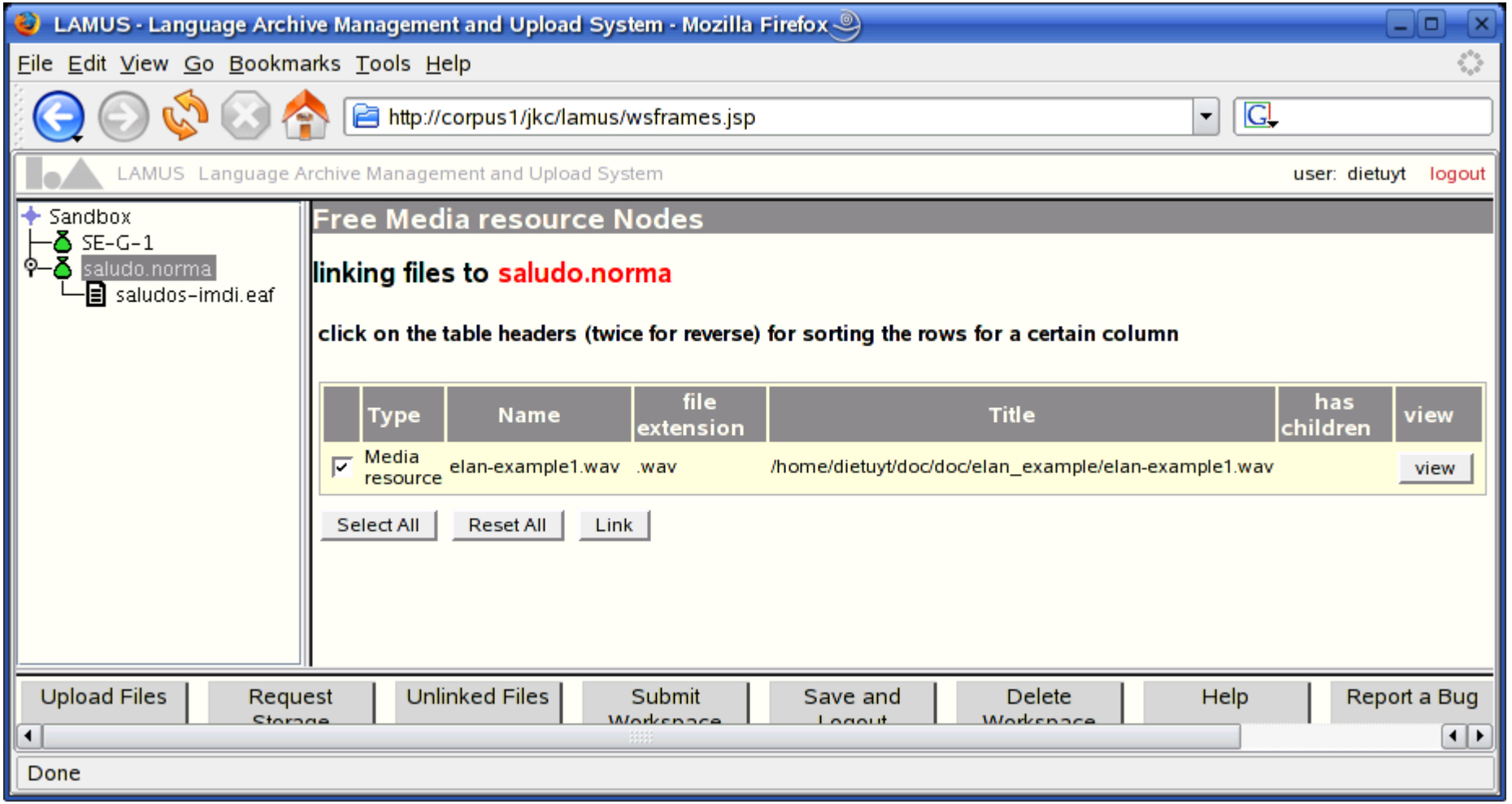
Focus on LAT apps: ARBIL



Focus on LAT apps: IMDI browser



Focus on LAT apps: LAMUS



Focus on LAT apps: Access Management

Node Authorization Overview

- view access privileges to LAT Resources 'LAMS Demo: Lams Demo Title' -

Focus overview on selected User/Group: <Pease Select>

Rules of Node 'LAMS Demo'

- MPI0#/MPI1# -

Assigned Licenses:

Webcode (webcode.html)

Manage Node Licenses

John Doe	• Read Info Files Allow • Read Annotations Allow • Read Images Allow • Read Audio Files Allow • Read Video Files Allow	Edit
John Doe's research group	• Read Info Files Allow • Read Annotations Allow	Edit

Add new Rules

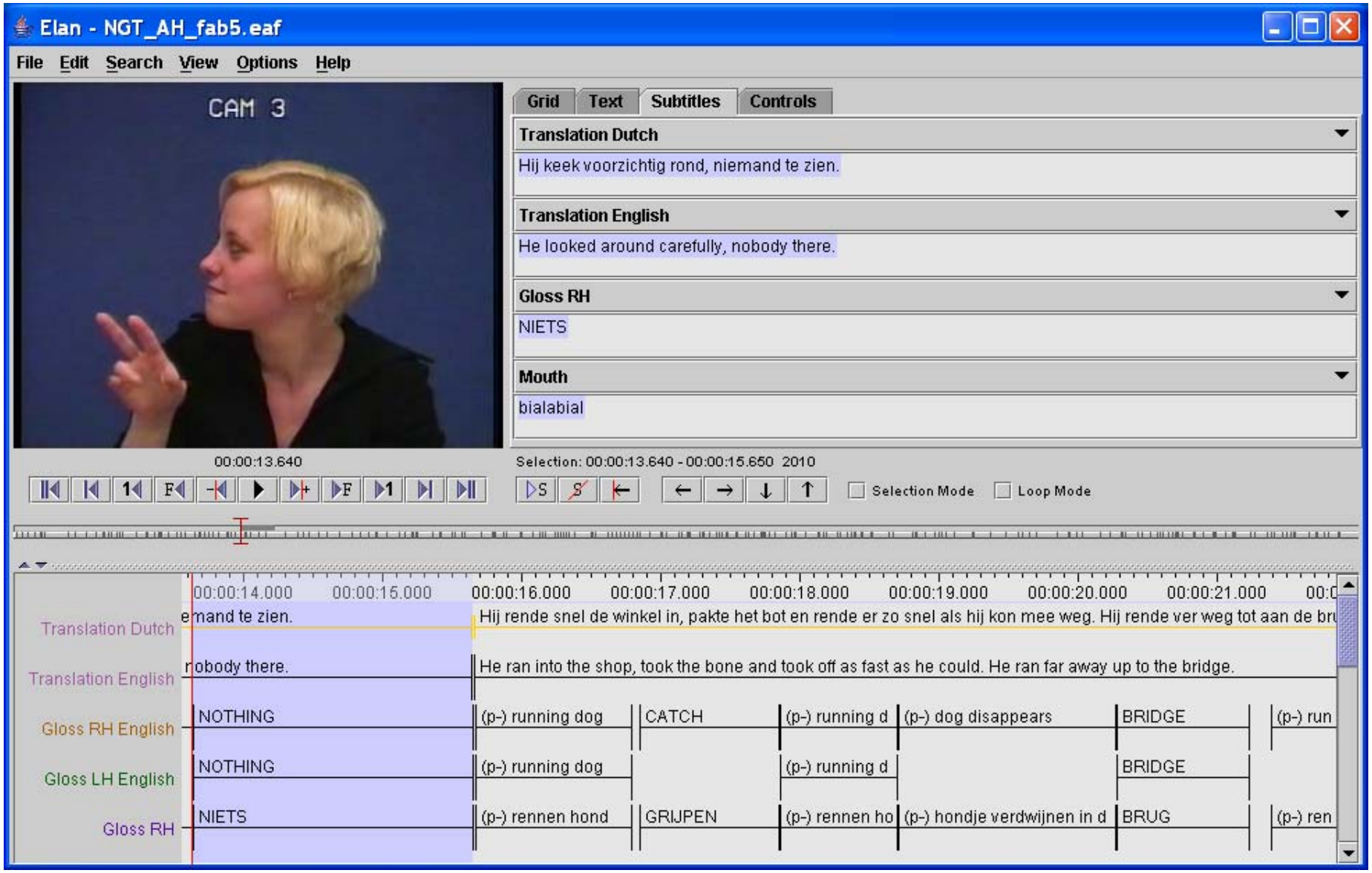
Rules of Node 'Demo'

- MPI0# -

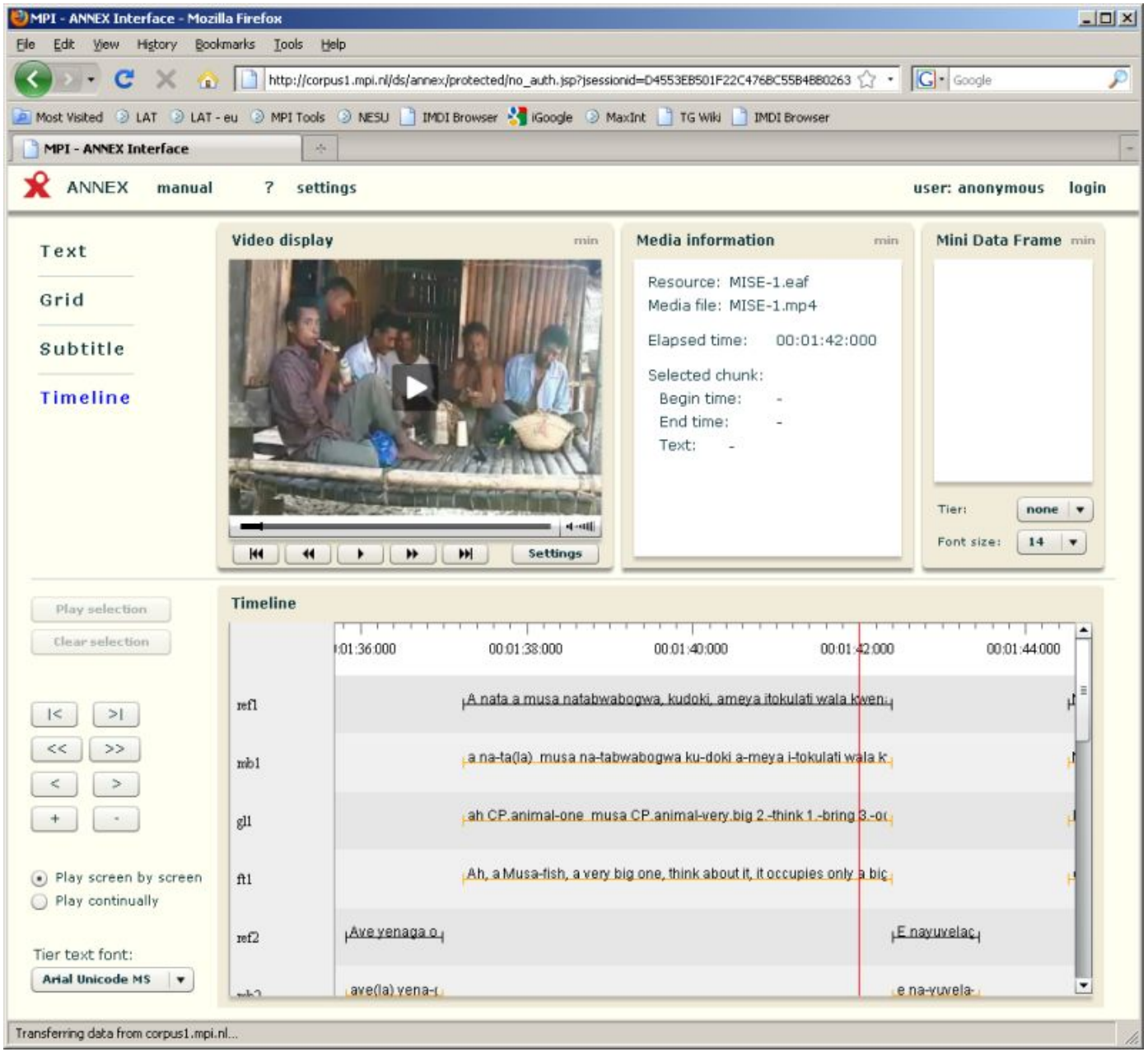
There are no licenses assigned to Node 'Demo'

There are no privileges directly assigned to Node 'Demo'

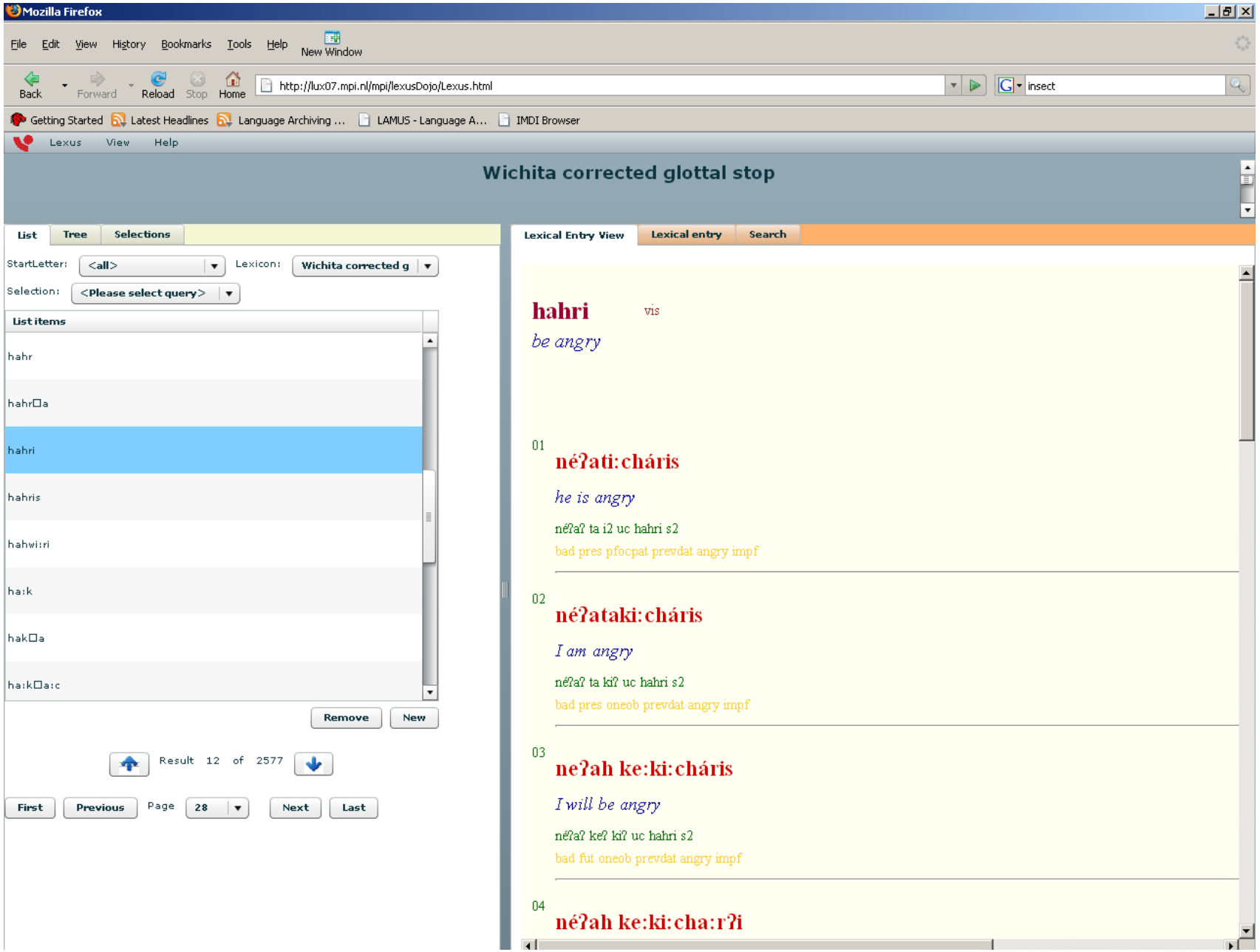
Focus on LAT apps: ELAN



Focus on LAT apps: ANNEX



Focus on LAT apps: LEXUS



Focus on LAT apps: VICOS

ViCoS – Visualising Conceptual Spaces

http://lux16.mpi.nl/mpi/vicos/ViCoS.html

ViCoS – Visualising Conceptual ... | http://lux16.mp...ojo/Lexus.html | ViCoS Editor and Navigator | ViCoS Image Manager

ViCoS - Max Planck Institute For Psycholinguistics - http://www.mpi.nl

Lexicon: Yeli_Dnye | Data Categories: lexeme | Help | RT Mgr

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z te search

Word list view

- chaa póopu
- chuu kîgha
- ch:oo
- chuu
- chile
- cheete
- chii lêê
- chîmo
- chalê
- châám:a
- chu
- chaa pyu
- chêêlî
- chimi chapî
- chala
- chili
- chaa

Welcome | tpile tp:oo | ch:oo | te

Lexical entry view for ch:oo (Yeli_Dnye)

Lexical Entry	description	description	image	description	image	exeme	description
[ch:oo]	small rafter	hitching pole for canoes		fish sp. (Orangespine unicornfish, Naso literatus)	null	ch:oo	tree type

Source

Clear

Relation Type: is_a_kind_of

Save

Target: [tpile tp:oo]

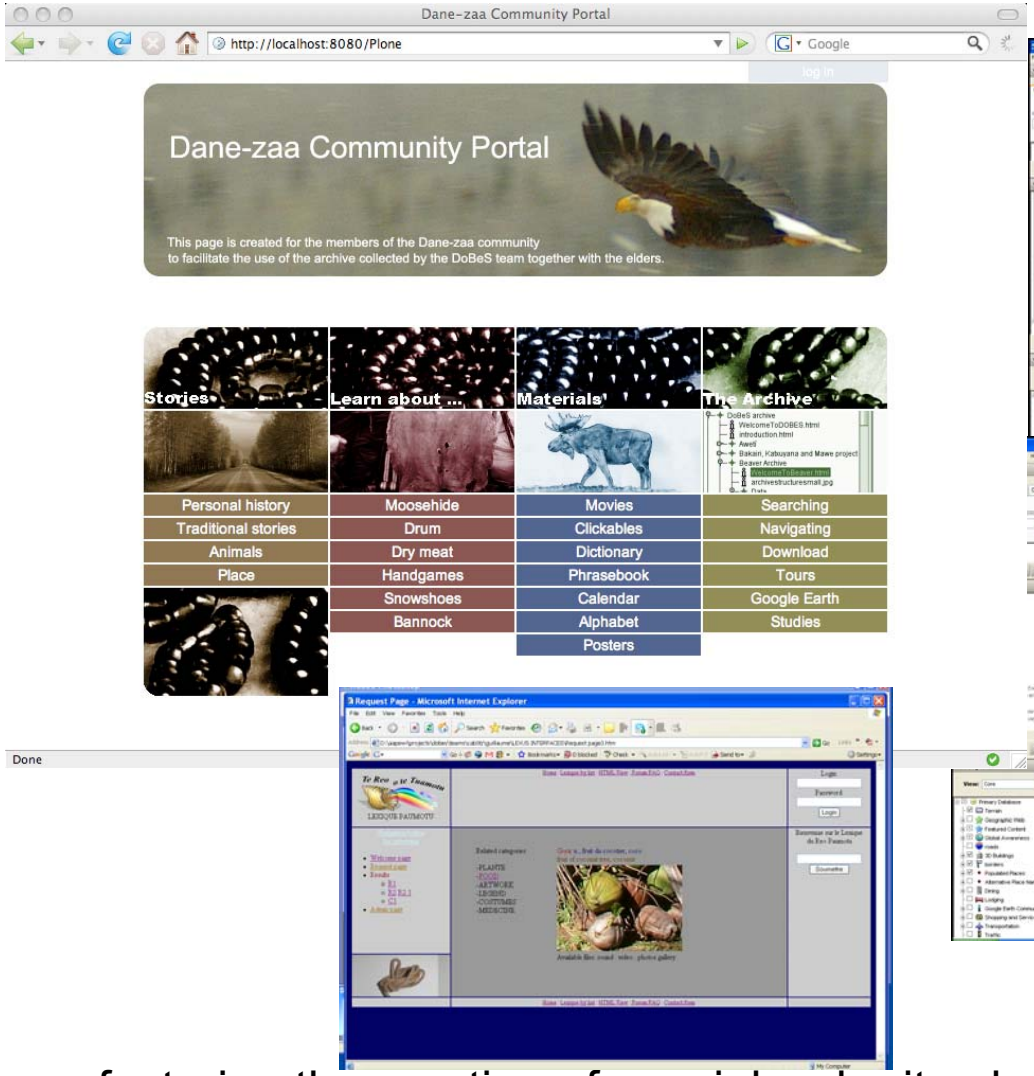
Clear

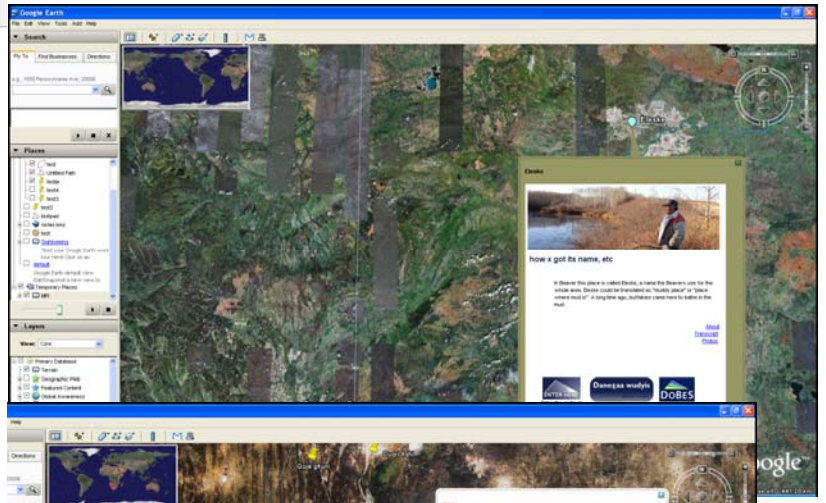
Add to Knowledge Space | Show in Knowledge Space | Edit in Image Mapper | Edit in Text Mapper

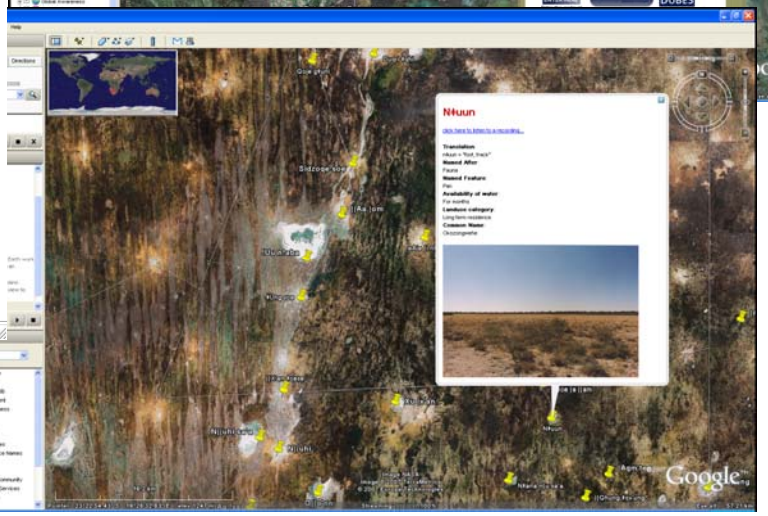
ViCoS Status Line

Done

New: Portals & Special GIS Stuff

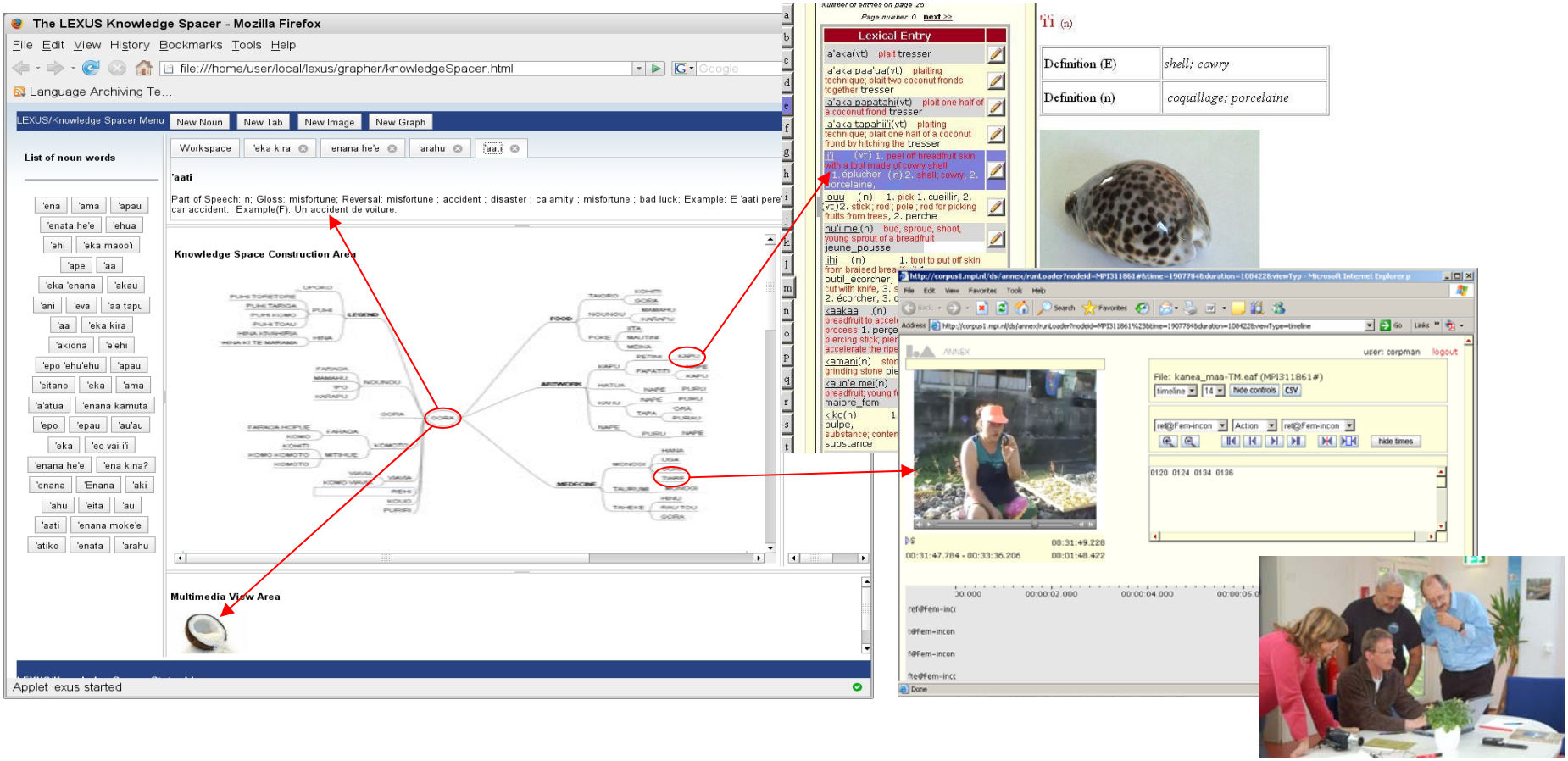






- fostering the creation of special web-sites by REST interfaces and templates
- fostering the GIS presentation by special converters

New: Cultural Documentation



- collaborative knowledge space with culturally relevant concepts for semantic navigation where concepts are center points for all sorts of information
- genealogy view to come next

Future evolutions: CLARIN

- Building the next-generation eHumanities infrastructure:
  - (No new data,) but make the existing data interoperable
- Some pillars:
  - Component-based metadata
  - Data Category Registries
  - Persistent Identifiers + Virtual Collections
  - Federated Identity management
  - Simplified license templates

Future evolutions: AVATECH

- Advancing Video/Audio Technology in Humanities Research
- MPI Halle + MPI Nijmegen + Fraunhofer institute
- Integrate audio/video recognizers in LAT framework

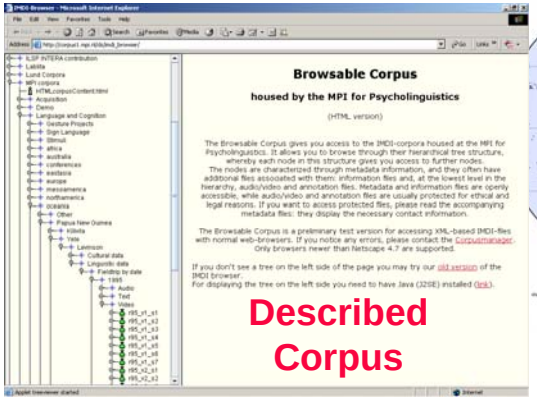
LAT

Questions?

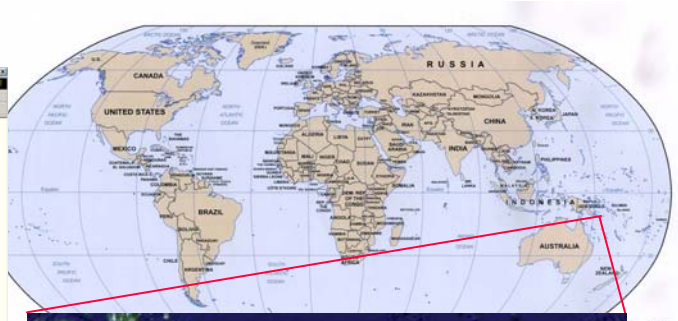
LAT

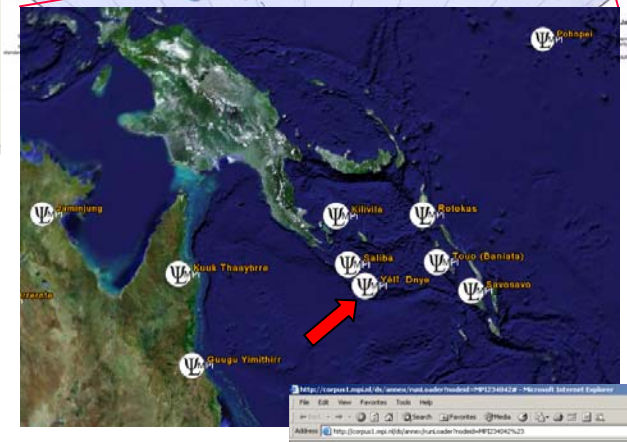
Backup Slides

Archive Content:
Yélî Dnye (Rossel Island)



Described Corpus



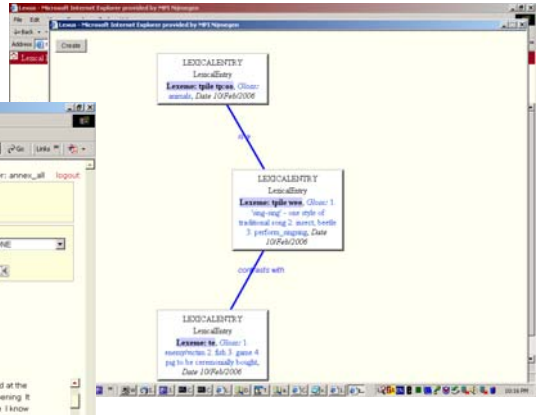


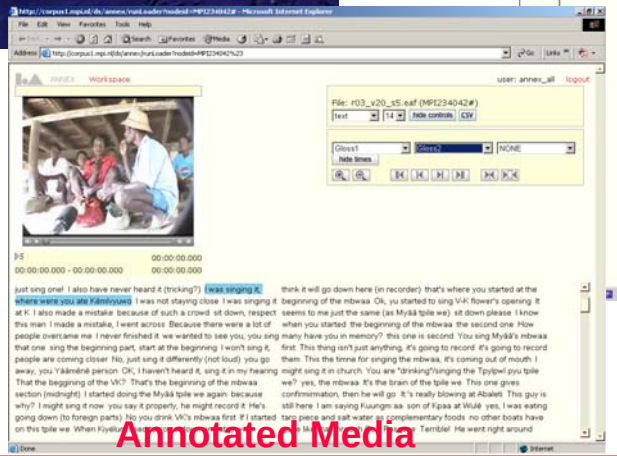
Multimedia Lexicon

Video Clips

Photos

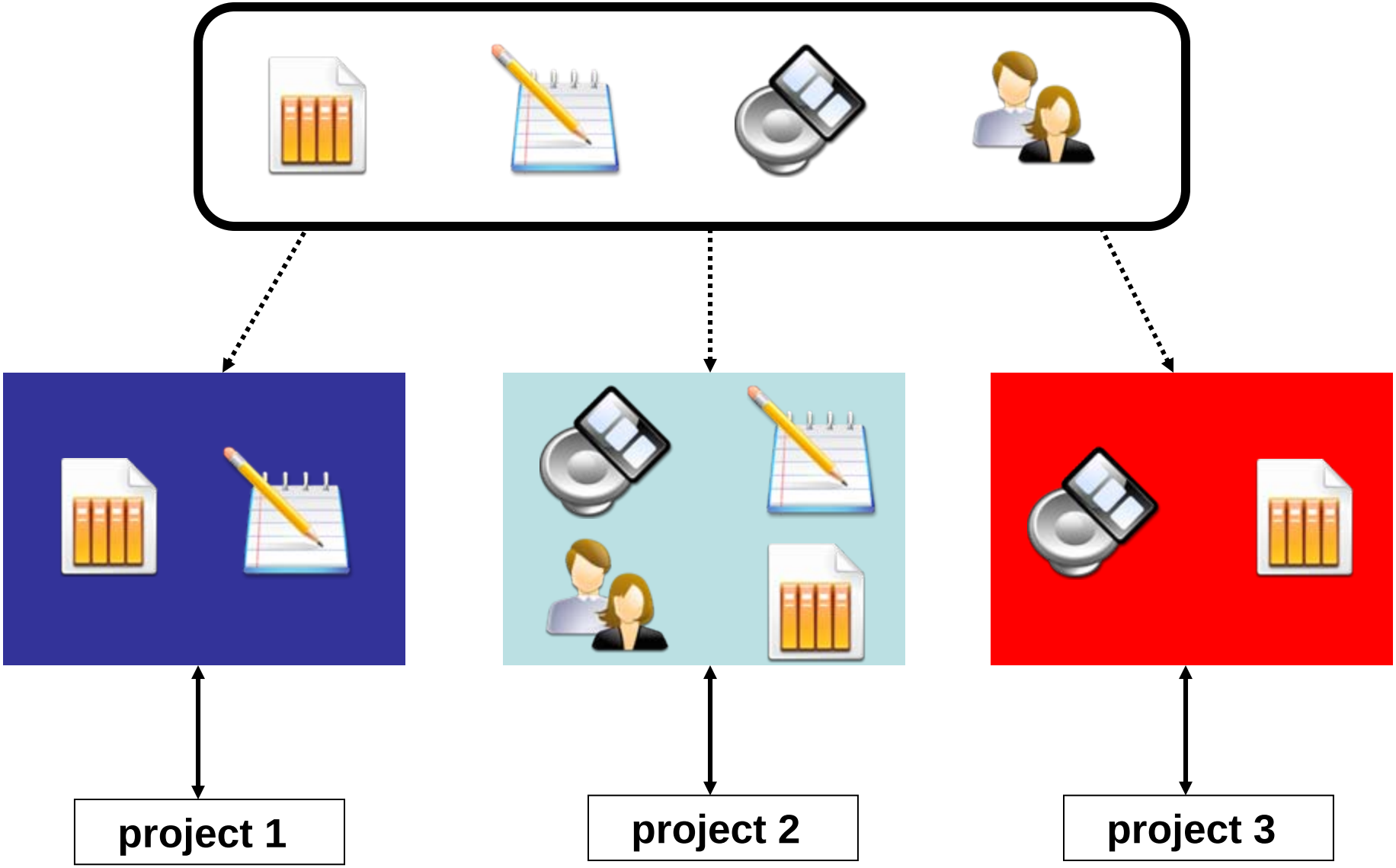
Typed Relations within the Lexicon



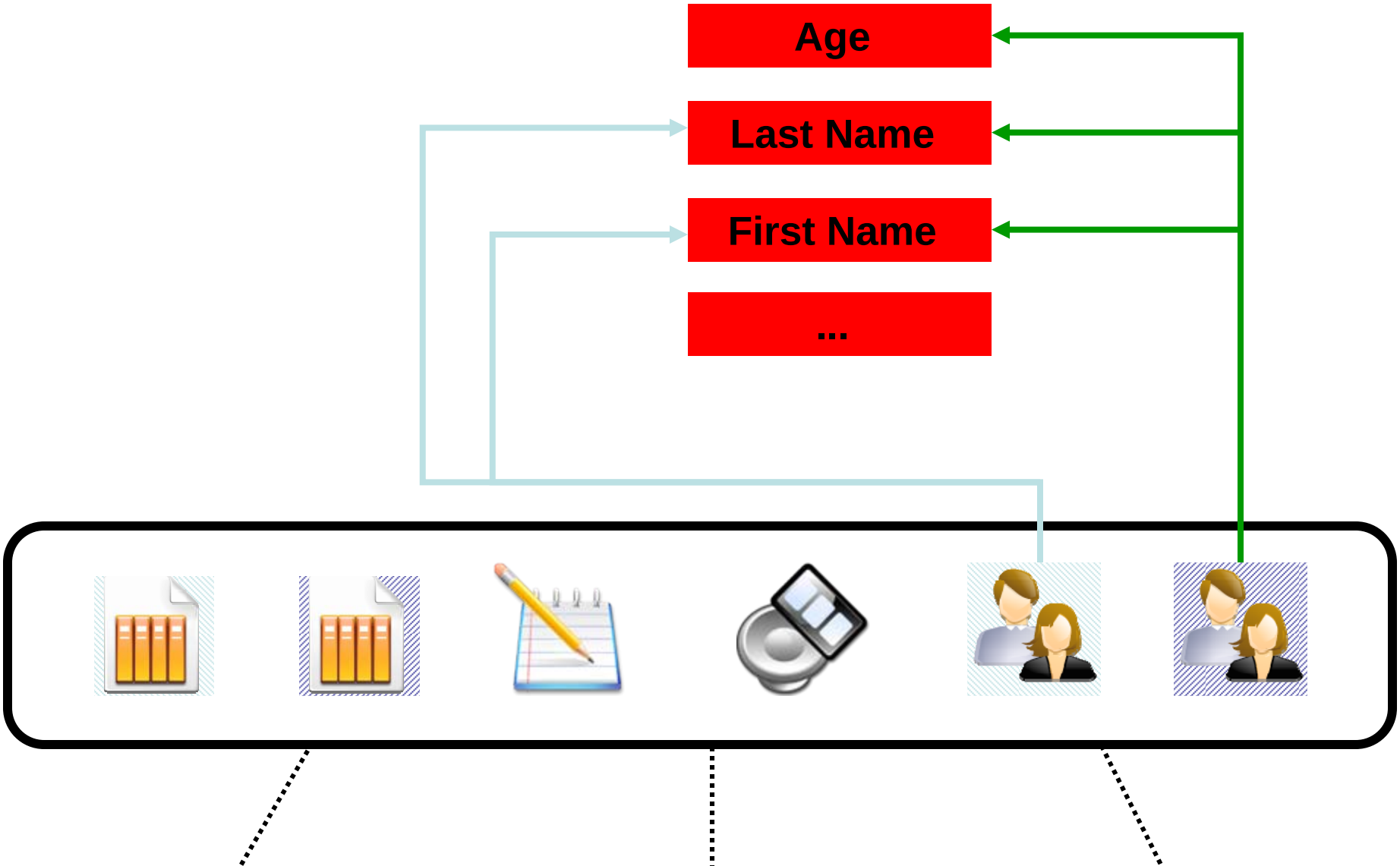


Annotated Media

Component-based metadata



Data Categories



References

- LAT: http://www.lat-mpi.eu
- MPI Nijmegen: http://www.mpi.nl
- CLARIN: http://www.clarin.eu
- CGN: http://lands.let.kun.nl/cgn/

LAT

Language Archiving Technology

http://www.lat-mpi.eu