

A HuComTech adatbázis mint kutatási infrastruktúra

Váradi Tamás

`varadi.tamas@nytud.mta.hu`

MTA Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály





Nyers adat → nyelvi erőforrás

- annotáció
- metaadatok
- szabványok

Erőforrások → infrastruktúra

- CLARIN (2008 -)
- CESAR@META-SHARE (2011 - 2013)
- HUNCLARIN - NEKIFUT (2009 - 2014)



Nem pusztán nyers adatok halmaza

- reprezentatív minta
- annotáció = hozzáadott érték
 - XML technológia
 - szabványos kódolási útmutató TEI (Text Encoding Initiative)
- metaadatok



Nem pusztán erőforrások halmaza

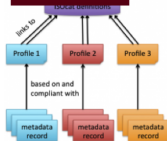
- központi gondolat: megosztás
- metaadatok és szabványos annotáció
- az erőforrásokmetaadat megfelelő gondozása
 - lehetőleg szabványos és egységes annotáció
 - egyértelmű és átlátható jogi helyzet
- interoperabilitás
- fenntarthatóság



Technology

Component Metadata

[CMDI examples](#)
[CMDI tutorial September 2012](#)
[Component Registry Documentation](#)
[Component Registry REST interface documentation](#)
[OAI-PMH for CMDI](#)
[Feedback](#)
[Content Search](#)
[Federated Identity](#)
[Persistent Identifiers](#)
[Repositories](#)
[Specification Documents](#)

 CLARIN Portal (new users start here!)
 Search language resources
 Depositing Services
 Easy access to protected resources
 Virtual Language Observatory
 Web Services


range of language resources.

Further information and examples

- For a detailed explanation, take a look at the [Frequently Asked Questions about CMDI](#)
- For a general introduction, read the [CMDI section in the CLARIN-D User Guide](#)
- [Examples and data sets](#)

... Document about metadata and metadata based on CLARIN D-User Guide

Metadata for language resources and tools exists in a multitude of formats. Often these descriptions contain specialized information for a specific research community (e.g. [TEI](#) headers for text, [IMDI](#) for multimedia collections).

To overcome this dispersion CLARIN has initiated the Component MetaData Infrastructure ([CMDI](#)). It provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile"). Both are stored and shared with other users in the [Component Registry](#) to promote reuse. Each metadata record is then expressed as an XML file, including a link to the profile on which it is based.

The CMDI approach combines architectural freedom when modeling the metadata with powerful exploration and search possibilities over a broad





The screenshot displays the Virtual Language Observatory (VLO) website interface. At the top, there is a navigation bar with the VLO logo and the text "Nyelvtudományi Intézet Magyar Tudományos Akadémia". Below this, a search bar contains the text "formal dating" and a "Search" button. The search results are displayed in a list format, showing 10 results. Each result includes a title, a description, and a "Expand" button. The results are as follows:

Item ID	Title	Description	Action
000016_F	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000016_I	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000016_2	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000016_3	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000022_F	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000022_I	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000022_2	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000022_3	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000024_F	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000024_I	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000024_2	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000024_3	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000026_F	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000026_I	Arrested formal dating (simulated job interview) - Hungarian	Arrested formal dating (simulated job interview)	Expand
000026_2	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand
000026_3	Arrested formal dating about personal topics - Hungarian	Arrested formal dating about personal topics - Hungarian	Expand

On the right side of the search results, there is a "NARROW DOWN" section with a list of filters: COLLECTION, COUNTRY, MODALITY, GENDER, SUBJECT, FORMAT, ORGANISATION, and DATA PROVIDER. Below this is a "SEARCH OPTIONS" button.



Accessibility | Contact | Log in

A Network of Excellence forging the
Multilingual Europe Technology Alliance

Search Site
META META-NET META-VISION META-SHARE META-RESEARCH News & Events LT-World Contact

Introducing META-NET

META-NET, a Network of Excellence consisting of 44 research centres from 31 countries, is dedicated to building the technological foundations of a multilingual European information society. **META-NET** is forging **META**, the Multilingual Europe Technology Alliance.

Join the online discussion!

Join META!

Participate in **META-FORUM 2011** (June 27/28)!

Europe's rich and diverse linguistic heritage must be the multicoloured fabric from which its web is made, rather than hindering the free flow of knowledge and thought.

Members Countries			
META	258	42	List Join
META-NET	44	31	List -

RECENT BLOG POSTS

- More than half EU internet surfers use foreign language when online May 11, 2011
- META-NET Vision Paper Updated Apr 25, 2011
- META-FORUM 2011: Registration now open Apr 21, 2011
- More...

UPCOMING EVENTS

- International Workshop on Ontology and Semantic web for Manufacturing (OSEMA 2011)
- 8th Extended Semantic Web Conference (ESWC 2011)
- 5th INTERNATIONAL CONFERENCE ON LANGUAGE AND AUTOMATA THEORY AND APPLICATIONS (LATA 2011)
- More...

META on

Accessibility | Contact | Log In



A Network of Excellence forging the
Multilingual Europe Technology Alliance

META META-NET META-VISION META-SHARE META-RESEARCH News & Events LT-World Contact

Search Site English



CESAR
Central and South-east europeAn Resources


CESAR – Summary – Consortium



The CESAR kick-off meeting in Luxembourg, 2011-02-10

Coordinator: **Tamás Váradi**
Research Institute for Linguistics
Hungarian Academy of Sciences












About the partners
 META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning it into as largely distributed infrastructure as possible.

Select network's node
 Please select one of the following META-SHARE network nodes to proceed. For an explanation of the differences between META-SHARE Managing Nodes and other META-SHARE nodes, you can visit [this page](#).



















META-SHARE Managing Nodes


 <small>CNR - National Research Council of Italy</small>	 <small>DFKI - Deutsches Forschungszentrum für Künstliche Intelligenz</small>	 <small>ELIA - Evaluation of Language resources Distribution Agency</small>	 <small>FBK - Fondazione Bruno Kessler</small>	 <small>IEA - Institute for Language and Speech Processing</small>	 <small>ICAR - Institute of Computer Science, Polish Academy of Sciences</small>
--	---	---	--	--	--



Meta-Nord

Other META-SHARE Nodes

 <small>Bulgarian Academy of Sciences</small>	 <small>Phonetic Language Semantics & Pragmatics</small>	 <small>Institute for Religion Language, Pragmatics, Academy of Sciences</small>	 <small>ELIAC Institute of Linguistics, Slovak Academy of Sciences</small>	 <small>Lithuanian Language Institute</small>	 <small>National Library of Hungary</small>
 <small>Research Institute for Linguistics, Hungarian Academy of Sciences</small>	 <small>Russian Academy of Sciences Center for Artificial Intelligence</small>	 <small>University of Bologna</small>	 <small>University of Copenhagen</small>	 <small>University of Göttingen</small>	 <small>University of Helsinki</small>
 <small>University of London</small>	 <small>University of Leeds</small>	 <small>University of Turku</small>	 <small>University of the Basque Country</small>	 <small>University of Jyväskylä</small>	 <small>Universitat Politècnica de Catalunya</small>



HSE - Higher School of Economics



Home
Profile
META-SHARE

[Home Resource](#) | [Community](#) | [Recommendations](#) | [Statistics](#)

HuComTech Multimodal Corpus and Database

HuComTech

<https://hucotech.uniobk.hu/hucotech/>

DOI ID [10.21203/rs.3.rs-1000000/v1](#)

The HuComTech multimodal corpus consists of about 50 hours of video and audio recordings of 111 formal dialogues (stimulus job interview and 111 informal but guided dialogues. The language of the recordings is Hungarian. The participants were university students aged 19-22 female 56 and male 45. The corpus was annotated for video (facial expressions, instances of eyeblinks, gaze, headshakes, handshakes, touchmarker and posture) and audio (interruptions, discourse, prosody and textual transcription). Its unique features in a wider comparison include its special attention to pragmatics focusing on a comparative study of the annotated vs. multimodal features of communication (as compared to multimodality alone) as well as the study of the syntax and prosody of spoken language with respect to the above wide range of multimodal characteristics. The data can be queried in ELAN and in our web-based SQL database.

[Back](#) | [Download](#) | [Edit Resource](#)

Distribution

Availability
Available - Restricted Use
Start date: 01/29/2013

License
All-CC-BY-NC-ND

Restrictions: Academic - Non-Commercial Use, No Derivatives
Distribution Access/Methods: Web Executable
Licenses: [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#)

Distribution rights holders:
Department of General and Applied Linguistics, University of Debrecen

DOI Holder:
Department of General and Applied Linguistics, University of Debrecen

Contact Person:
László Hunyadi

audio %

Monolingual audio corpus

Languages
Hungarian

Linguistic
Linguistic type: Monolingual

Size
50 hours

Modality

Spoken Language

Metadata

Created: 01/08/2013
Last Updated: 05/10/2014
Source: COAR

Metadata Creator:
[László Hunyadi](#)

People who looked at this resource also viewed the following:

- [DISEM-C: Corpus of Hungarian School Metalinguage - Interview Corpus](#)
- [Hungary](#)
- [H-cvlib](#)
- [HuNE2016: Automatically generated NE tagged corpus for Hungarian](#)





MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS



The Language Archive





Home
Team
Projects
Tools
Resources
Events
Forums
Contact

TLA Tools
Other Tools



The Language Archive

The Language Archive:

- Data archive about languages worldwide
- Data creation, management and exploration tools
- Archiving and software expertise for the Digital Humanities

A unit of the **Max Planck Institute for Psycholinguistics**.
Funded by **BBAW, KNAW** (in cooperation with **DANS**) and **MPS**.

The Language Archive > Tools > TLA Tools

TLA Tools

	AMS	a tool to grant and deny the access to a (part of a) corpus.
	Annex	the Annotation Exploration tool in the MPI web-based framework for archive exploration and enrichment
	Arbil	Arbil is an application for arranging research material and associated metadata into a format appropriate for archiving.
	Archive Structure Viewer	access the MPI archive through the web
	Browser/Catalogue	access the MPI archive through the web
	ELAN	Multimedia Annotator
	Imex	a tool for exploring images
	ISOcat	a web-based Data Category Registry
	Kin-Graph Kinship Archiver	A flexible kinship application producing publishable quality diagrams
	LAMJUS	Language Archive Management and Upload System
	Lexus	a web-based lexicon tool
	Trova	the annotation search tool in the MPI web-based archive framework
	Vicos	a tool that allows users to complement lexical spaces (as created by LEXUS) with ontological spaces.

Workflow of listed tools:















Search

Quick links

TLA tools

[Access the Archive](#)

You are not logged in.

[Login/Register](#)

The Language Archive

Max Planck Institute for Psycholinguistics


Street address:
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

Mailing address:
P.O. Box 310
6500 AH Nijmegen
The Netherlands

+31 24 3521 011
(Reception)

Email: [please use the contact form](#) or the [support forums](#)






**National Research
Infrastructure Register**

Bejelentkezett felhasználó adatai

User name: **Sass Bálint** User status: **Adatgazda** [Logout](#)
 Research infrastructure: **Hungarian CLARIN network (Network)**, 26 member(s)
 Type of the RI: **Research infrastructure**

Kutatási Infrastruktúrák listája

Hungarian CLARIN network (Network, 26 member(s))



Short name of the RI:	HunCLARIN
Organization:	Research Institute for Linguistics of HAS
Contact person:	Váradai Tamás
Website of the RI:	http://-
Area of science:	<p>Clinical medicine : Other clinical medicine Biotechnology : Environmental biotechnology Linguistics and literary scholarship : Linguistics Philosophical and historical sciences : Psychology Engineering sciences : Automation and computing, Electronic devices and technologies Mathematical sciences : Operational research, Information technology and computer science</p>
Status of the RI:	Evaluated
A KI publikus:	Publikált
A KI munkacsoportjai:	MCS#3

HunCLARIN is the part of the European CLARIN network. RIL HAS is one of the founding members of CLARIN, and participates in management committee. CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling humanities.



Köszönöm a figyelmet!