

A HuComTech adatbázis a Nyelvtudományi Intézet nyelvi archívumában

Oravecz Csaba

oravecz.csaba@nytud.mta.hu

MTA Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály





Előzmények

- digitálisan hozzáférhető nyelvi adatok mennyiségének ugrásszerű növekedése (hangzó anyagok, multimodalitás)
- ad hoc tárolási megoldások → adattemetők

Kívánalmak

- változatos megjelenési formájú nyelvi adatok hosszútávú megőrzése
- a gyűjtött adatok szervezett és biztonságos tárolása
- hatékony és szabványos hozzáférés biztosítása



Megoldás

- kutatási infrastruktúrába illeszkedő adatszolgáltató központ
- metaadatok sztenderdizált módon történő rögzítése
 - **IMDI**: ISLE Meta Data Initiative
(ISLE: International Standard for Language Engineering)
 - **CMDI**: Component MetaData infrastructure
- kereshetőség, hozzáférés szabályozhatósága



HASRIL IMDI Metadata Domain

RESEARCH INSTITUTE FOR LINGUISTICS

The IMDI Metadata Domain allows you to browse and search in the whole domain of linked IMDI metadata descriptions as they are registered at the IMDI portal at the Research Institute 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](#).

How to explore the archive:

On the left hand side, you see the whole hierarchy of linked IMDI corpora, which you can explore by clicking on the small circles or by double-clicking the names. If you click on an item with the **right mouse button** (or control-click with a single button Mac mouse), you will get a contextual menu with additional functionality, depending on what kind of item you have selected:

- view node:** shows the content of a file. For Metadata files, you will get the content of the metadata description. For media files, you will get to see the actual resource. For EAF, Shoebox/Toolbox and CHAT annotation files, the ANNEX annotation viewer will be launched. If access to a resource is restricted, you will get an authentication window in which you need to type your user name and password (only once).
- create bookmark:** gives you a page with several links on it which you can bookmark, so you can easily return to that resource or corpus node or send it to someone else.
- metadata search:** gives you a metadata search interface. All metadata files under the selected part of the hierarchy will be searched. There are two types of search: a simple "google-like" search, and a more advanced search for searching within selected metadata



The screenshot shows the IMDI Browser interface. On the left is a navigation tree with the following structure:

- HASRIL Language Archive Server
 - WelcomeToHASRIL.html[0]
 - Demo
 - External Resources
 - Hungarian Multimodal Corpus [0]
 - HuComTech Corpus[1]
 - Resources [0]
 - HASRIL Resources

The main content area displays the following information:

hide accessibility of resources about manual register user: anonymous login logout

metadata search annotation content search manage access rights request resource access

View Node Create bookmark Download

Imdi Information

IMDI ISLE Metadata Initiative

Corpus

Name Hungarian Multimodal Corpus
Title HuComTech Hungarian Multimodal Corpus

Description

The HuComTech multimodal corpus contains 50 hours of recorded conversations divided into two major parts: a simulated job interview and a guided dialogue about personal topics. The participants are university students (54 females, 67 males) mostly involving the same interviewer in both scenarios. The language of the recordings is Hungarian. The corpus was created for various purposes: improve and verify communication models, provide training material for recognition software, investigate the individual features of spontaneous

Catalogue

Name HuComTech Corpus
Title HuComTech Hungarian Multimodal Corpus
Id

DocumentLanguages

SubjectLanguages

[Location](#)



IMDI Browser

hide accessibility of resources about manual register user: anonymous login logout

- HASRIL Language Archive Server
 - WelcomeToHASRIL.html[0]
 - Demo
 - External Resources
 - Hungarian Multimodal Corpus [0]
 - HuComTech Corpus [1]**
 - Resources [0]
 - HASRIL Resources

Imdi Information

IMDI ISLE Metadata Initiative

Catalogue

Name HuComTech Corpus
Title HuComTech Hungarian Multimodal Corpus
Id

DocumentLanguages

SubjectLanguages

Location

ContentType
Multimodal

Format

Quality

SmallestAnnotationUnit
Unspecified

Applications
natural language processing, speech recognition, discourse analysis, gesture recognition, prosody, topic detection and tracking

Date
2013

Project HuComTech

Publisher
University of Debrecen, Dept. of General and Applied Linguistics

The screenshot shows the MIDI Browser interface. On the left is a tree view of resources under 'Hungarian Multimodal Corpus' and 'HuComTech Corpus'. The selected resource is '008mc20_l_a.eaf'. On the right, the 'Resource Information' panel displays the following details:

- URL: https://clarin.nytud.hu/corpora/media-archive/donated_corpora/HuComTech/Resources/Annotations/008mc20_l_a.eaf
- Handle URI: hd:1839/00-0000-0000-001B-AB5B-7
- Internal Node ID: MPI271#
- Node type: Written Resource
- Format: text/x-eaf+xml
- File size: unknown
- Last modified: unknown
- MDS Check sum: unknown

The 'Access' section contains the following text:

You can access the resource itself (if you have the privileges), by clicking on the the selected resource icon with the right mouse-button and selecting the "view" option or use your browser (plugin) via this [link](#). You will need to authenticate yourself (again) for downloading the resource when it is protected.

General Accessibility: ● This resource is accessible to registered users of the archive
Currently accessible to user clarin_test@nytud.mta.hu: yes

Applicable License Agreement(s)
No licenses required for this resource.

For more information about this resource see the metadata in the containing session.

IMDI Browser MPI - ANNEX Interface user: clarin_test@nytud.mta.hu logout

Annex 1.5.41597 manual ? embed Show tooltips Compact Spacious

Text

Grid

Subtitle

Waveform

Timeline

Combined

Video display min

00:05:31

Full Buffer

Information min

General Session Technical

Resource: 008mc20_1_a.eaf
 Media file: 008mc20_1_C3.mp4
 Elapsed time: 00:00:08:726

Selected chunk:
 Begin time: -
 End time: -
 Text: -

Mini Data Frame min

%s %s +igen. %s V kicsit másabb volt, mint Kingánál, **így jobban izgultam egy** %s ((tehát.hogy--)) ((még)) a kérdések is, hogy --%s uhum. (b) %s uhum. ((jó.)) %s %s %a legjobb élményem, mikor %o Pesten voltunk az ismerősökkel, és két napot eltöltöttünk. %s bejártuk egész Budapestet, %s felmentünk a Gellért-hegyre, a Citadellára, mentünk a Duna-parton, nagyon jó volt. %s %o az akkor nem volt. %s %s debreceni vagyok. %o mostanában szoktunk menni, meg régebben is

Tier: **A_speaker_text**

Font size: 14

Play selection

Clear selection

Create bookmark

<| >|

<< >>

< >

+ -

Play screen by screen

Play continually

Waveform and Timeline

0:00:05:000 0:00:05:500 0:00:06:000 0:00:06:500 0:00:07:000 0:00:07:500 0:00:08:000 0:00:08:500 0:00:09:000 0:00:09:500 0:00:10:000 0:00

Audio channel

A JP	SL	V	HC	SC	SC	SL	V	HC
A_emotional	SL	N	N	N	N	SL	N	P
A_discourse	SL	T K	T	K	K	SL	K T	T
A_speaker_text	%s	V	kicsit másabb volt,	mint Kingánál,	így jobban izgultam egy	%s	(((tehát.ho ((még)) a két	
A_agent_text	újini magad.	%s				%s		
SegErrors								
SynErrors								

IMDI Browser

hide accessibility of resources about manual register user: anonymous login logout

[metadata search](#)
[annotation content search](#)
[manage access rights](#)
[request resource access](#)

[View Node](#)
[Create bookmark](#)
[Download](#)

Imdi Information

IMDI ISLE Metadata Initiative

Corpus

Name Hungarian Multimodal Corpus
Title HuComTech Hungarian Multimodal Corpus

Description

The HuComTech multimodal corpus contains 50 hours of recorded conversations divided into two major parts: a simulated job interview and a guided dialogue about personal topics. The participants are university students (54 females, 67 males) mostly involving the same interviewer in both scenarios. The language of the recordings is Hungarian. The corpus was created for various purposes: improve and verify communication models, provide training material for recognition software, investigate the individual features of spontaneous

Catalogue

Name HuComTech Corpus
Title HuComTech Hungarian Multimodal Corpus
Id

DocumentLanguages

SubjectLanguages

[Location](#)

IMDI Browser Trova Search Application MPI - ANNEX Interface

Trova 1.5.39358 [help](#) user: clarin_corpman@nytud.mta.hu [re-login](#) [logout](#)

Substring Search Single Layer Search Multiple Layer Search

Types: EAF (1)

Domain: 008mc20_F_a_v.eaf

History:

Mode: Sort tier lists: alphabetically by number of matches

Time:

 Found 0 hits in 0 annotations

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="button" value="All Tiers [18]"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="button" value="Must be in same file"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="button" value="All Tiers [18]"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="button" value="Must be in same file"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="button" value="All Tiers [18]"/>

Action: Page:

Context Size: Font: Show Info Balloons



<http://clarin.nytud.hu>



Köszönöm a figyelmet!