

Kutatási adatok az MTA Nyelvtudományi Intézetében

`linginst@nytud.mta.hu`

Oravecz Csaba
Nyelvtechnológiai Kutatócsoport
`www.nytud.hu`

Kutatási adatok ws, 2015. május 14.



Társadalomtudomány természettudományos módszerekkel

- több mint 100 kutató
- változatos kutatási területek
(nyelvtörténet [...] → fonetika [...] → számítógépes nyelvészet)
- változatos adatok
(kézirat [...] → hangfelvétel [...] → nemzetközi szabványokhoz illeszkedő XML annotált adatbázis)



- közel 30 féle gyűjtemény, adatbázis
- nyelvi adatok:
 - több százezer kéziratos cédula és gyűjteményes füzetek
 - több száz órányi hang- és videófelvétel
 - több millió lexikális tételt és
 - több milliárd szövegszót tartalmazó adatbázisok
- metaadatok:
 - kéziratos fejléc
 - külön táblázat
 - SQL adatbázis
 - T(ext) E(ncoding) I(nitiative) XML fejléc
 - szabványos leíró séma (ISLE Meta Data Initiative – IMDI)



Tárolás

- papírdobozban
- CD-n, saját számítógépen (fájlonként)
- szerveren (megosztással, könyvtárstruktúrában)
- szerveren adatbázisban
- repozitóriumban



Formátum

- papírcédula
- doc, xls, txt, xml
- wav, mp3, mp4

Kezelés

- archiválási protokoll nem jellemző, szabályozás nincs



Elérés

- saját felhasználás
- kutatócsoporton belül
- külön megállapodás alapján
- dedikált felhasználói felületen regisztráció után
- szabadon elérhető

Megosztás

- nincs általános szabályozás
- adatvédelem, kutatói hozzáállás



The screenshot shows a web browser window with the URL https://tla.nyttud.hu/ids/fmdi_browser/. The browser interface includes navigation buttons (Back, Forward, Reload, Stop), a search bar, and a print button. The page content is divided into two main sections.

Left Panel (File Listing): A tree view showing a directory structure. The selected file is `Kurtyamova Anna_Gavrilovna1`. Other files include `JeprinakKurtyamova_Vera_Nykolajevna1`, `KovaljovaPinszeva_Maria_Alekszandrovna_20`, and various `ValgAA` files.

Right Panel (Media File Details): A detailed view of the selected file, `Kurtyamova Anna_Gavrilovna1`.

- Date:** Unspecified
- Location:** Unspecified
- Project:** Unspecified
- Content:** Unspecified
- Actors:**
 - Actor:** Ruttkay-Miklán Eszter
 - Actor:** Gavrilovna
- MediaFile:**
 - Type:** audio
 - Format:** WMA
 - Size:** 11698KB
 - Quality:** Unspecified
- RecordingConditions:** Unspecified
- TimePosition:**
 - Start:** Unspecified
 - End:** Unspecified
- Access:** Unspecified
- Keys:**
 - Place:** Ovgort, Surtiskár Járás, Jamali Nyenyec Autonóm Község, Országoság year
 - Topic:** 2011
 - Processing:** Női tisztálanság, füstöléssel tisztítás, szénrel arc bekénesése, menopausa; szemétkézelés.
 - Duration:** Magyar nyelvű tartalomeírás időköddal
 - Informant:** 00:12:24

IMDI Browser | MPI - ANNEX Interface

Annex 1.5.41597 manual ? embed Show tooltips Compact Spacious user: clarin_test@nytud.mta.hu logout

Text

Grid

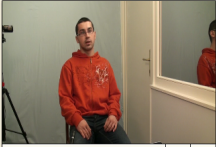
Subtitle

Waveform

Timeline

Combined

Video display min



019/523 Full Buffer

Information min

General Session Technical

Resource: 008mc20_F_a_v.eaf
 Media file: 008mc20_F_C2.mp4
 Elapsed time: 00:00:18:798

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

Mini Data Frame min

%s %s jó napot kívánok, [SURNAME] [NAME].
 köszönöm. %s %s %s %s még nem volt előző
 munkahelyem, pályakezdő vagyok, most jöttem az
 egyetemről. %s %s %s %s diák munkában több helyen is
 voltam. %s nyomdában dolgoztam, mint %s
 rakodómunkás, {b} %s és %s %a vendéglátásban
 dolgoztam emellett. %s %s %s %s úgy hallottam,
 önként egy fiatalos, megbízható csapatot keresnek, és
 ennek szeretnék tagja lenni. %s %s %s pályakezdőként

Tab: A_speaker_text
 Font size: 14

Play selection

Clear selection

Create bookmark

|< >|

<< >>

< >

+ -

Play screen by screen

Play continually

Timeline

A_IP			SL HE HC	
A_emotional			SL LR LN	
A_discourse			SL LT LK	
A_speaker_text			%s:%s	még nem volt előző munkahelyem, pályakezdő vag
A_agent_text	mesélne nekem kezdésnek az előző munkahelyeiről?		%s	
SegEmos				
SymErrors				
V_facialExpres			natural+moderate	
V_gazeClass			forwards	
V_cynsmwCla				
V_headhifCla				



Európai elosztott kutatási infrastruktúrák

- CLARIN (www.clarin.eu)
- META-NET kiválósági hálózat META-SHARE disztribúciós rendszer magyar csomópont

Együttműködés

- Max Planck Institute, Nijmegen



Back Forward Reload Stop http://metashare.nyud.hu/repository/search/ Search Print

Home Bookmarks Most Visited

META-SHARE Register Login

Browse Resources Community Documentation Statistics

Resource Type
Media Type
Availability
Licence
Restrictions of Use
Foreseen Use
Use in NLP Specific
Linguality Type
Multilinguality Type
Modality Type
MIME Type
Conformance to Standards/Best Practices
Language Variety

Resource Type:
Corpus:
Lexical/Conceptual:
Tool/Service:

BEA Hungarian spontaneous speech database 0 118
Hungarian

CHSM-IG: Corpus of Hungarian School Metalanguage - Interview Corpus 0 195
Hungarian

HHC: Hungarian historical corpus 0 47
Hungarian

Ht-online 0 36
Hungarian

HuComTech Multimodal Corpus and Database 0 35
Hungarian

hunalign 1 35

Hungarian Consize Dictionary 1 32
Hungarian

Hungarian Kindergarten Language Corpus 0 55
Hungarian



- nagyon változatos, eltérő kezelést igénylő adatok
- ad-hoc, kialakított szabályozás és infrastruktúra nélküli tárolás és archiválás
- szabványos metaadatolás ritka



- nemzetközi kapcsolatok, részvétel európai hálózatokban
- helyben üzemelő gyűjteménykezelő keretrendszer



Köszönöm a figyelmet