

# Spontán beszélt nyelvi adatbázisok építésének technológiai kérdései

Oravecz Csaba, Sass Bálint



NYELVTUDOMÁNYI INTÉZET  
MAGYAR TUDOMÁNYOS AKADÉMIA

{oravecz, joker@nytud.hu}

"The computer-readable encoding of transcriptions of spoken-language is a notoriously difficult area..."  
Thomas Schmidt (2005)

## Cél

- Spontán beszélt nyelvi hanganyagból olyan explicit nyelvi adatbázist készíteni, amely lehetővé teszi a számítógép hatékony felhasználását a szövegek **nyelvi** elemzésében és lekérdezésében.
- Számítógéppel segített elemzés:
  - számítógépes eljárásokkal minél több, minél relevánsabb adatot akarunk nyújtani az elemzéshez
  - minél gazdagabban, relevánsabban és egyértelműbben van kódolva a forrásanyag, annál használhatóbb az elemzéshez kigyűjtött adat

## Fókusz

- Fonetikai elemzések mellett magasabb szintű nyelvészeti elemzést is megengedő reprezentáció.

## Mit értünk itt adatbázis alatt?

Az információ

- explicit,
- egyértelmű,
- azonos szerkezetű,
- számítógéppel egyszerűen/hatékonyan kiolvasható/feldolgozható

formában van tárolva.

## Sztenderd kódolási modellek

- Rendezési elv alapján
  - megnyilatkozások egyes elemeinek időbeli viszonyai (ELAN, EXMARALDA, Praat stb.)
    - STMT (Single Timeline Multiple Tiers) modell: precíz leírás az időszekvenciákra vonatkozóan
    - mivel az időben szegmentált egységek nem feltétlenül esnek egybe releváns nyelvi egységekkel, a nyelvi elemzés leírásához a modellt ki kell egészíteni egy nyelvi szegmentumokat reprezentáló szinttel
  - megnyilatkozások egyes elemeinek hierarchikus viszonyai (TEI)
    - OHCO (Ordered Hierarchy of Content Objects) modell: precíz leírás a nyelvi elemzésre vonatkozóan
    - mivel a nyelvi szegmentumok nem feltétlenül esnek egybe a releváns temporális egységekkel, az időbeli viszonyok leírásához a modellt ki kell egészíteni az időszekvenciát reprezentáló szinttel
- Alkalmazható megoldás
  - Elvileg bármelyik, gyakorlatilag a transkripció fókusz szerint egyik vagy másik modell mint kiindulópont, kiegészítve a megfelelő reprezentációs szinttel.

## A lejegyzés gyakori problémái

- Szabadszöveges kódbevitel valamilyen általános szövegszerkesztővel:
  - megszorítatlan, kézi annotáció
  - elkerülhetetlen és a gépi ellenőrzés hiánya miatt rejtve maradó kódolási hibák
- Nem elsősorban számítógépes, hanem emberi feldolgozásra készült lejegyzés
  - kódolási többértelműség
  - nem explicit annotáció
  - horgonyzási pont és hatókör problémák

## A kódolási modell

- **Előfeltételek:**
  - a későbbi alkalmazásokra tekintettel kidolgozott egyértelmű és explicit annotációs útmutató
  - a transkripció készítése során a folyamatos formális ellenőrzést biztosító fejlesztői környezet
- **Példafeladat:**
  - részletes és a beszélt nyelvi jelenségek széles körét tartalmazó adatbázis előállítás, a tárolt információt szolgáltató felhasználói felülettel
- **A javasolt megoldás:**
  - hierarchikus TEI alapú XML annotáció
  - a kezelendő jelenségek egységesített leírásával,
  - az átfedő beszéd kezelésével kiegészítve az időszekvenciát reprezentáló szinttel.

## Miért XML?

- szabványos formátum, készen kapott feldolgozó modulokkal
- hasonló fejlesztések eredményei használhatók
- hordozható adatbázist eredményez.

## Adatbáziskezelő rendszer

- Elvárások:
  - robusztus működés
  - nagy kifejező erejű lekérdező nyelv
- Felhasználói felület
  - a jelenségek széles körét lefedő menürendszer (a lekérdező nyelv összes lehetőségét nem lehet menürendszerrel megragadni)
  - többszempontú megjelenítési funkciók

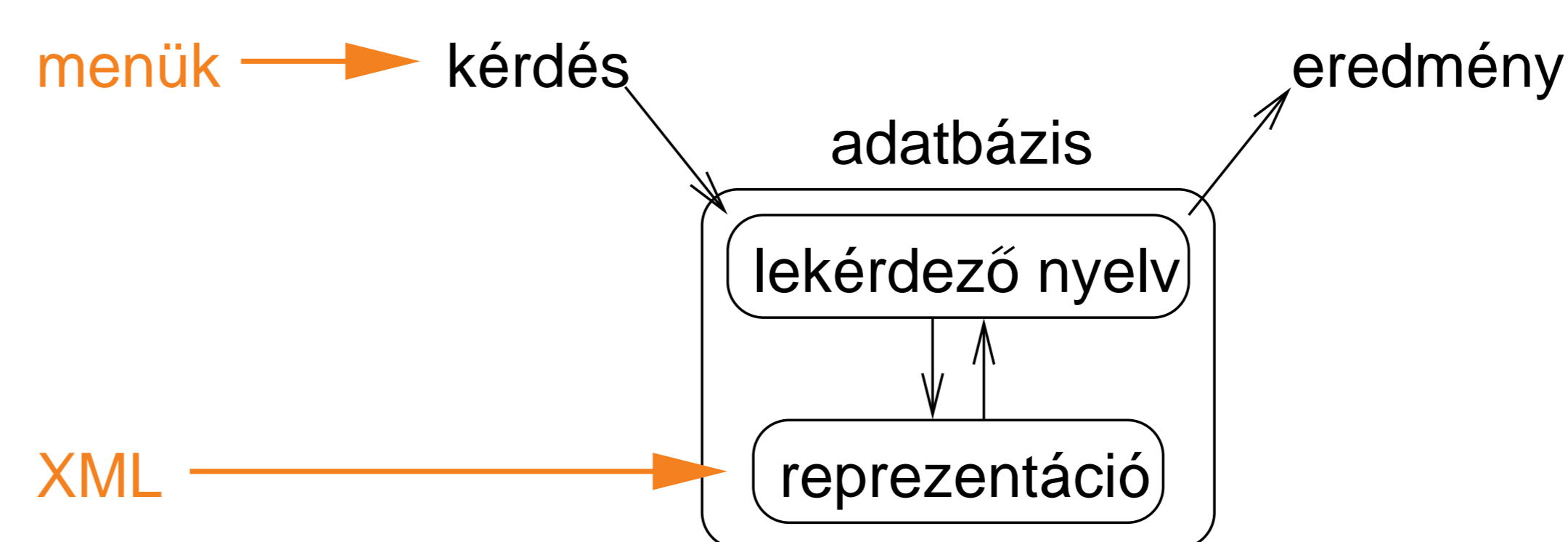
## Nyelvi elemzés mint hozzáadott érték

- ha a transkripció egyértelmű normalizált nyelvi elemeket tartalmaz, a sztenderd nyelvfeldolgozó alapeszközök (morfológiai elemző, egyértelműsítő, részleges szerkezeti elemző stb.) közvetlenül alkalmazhatók
- elemzés:
  - egyértelműsített morfológiai elemzés, szótó
  - szótó CV váza, magánhangzók BNF alakban
  - szóalak egyszerűsített fonetikai reprezentációja

## Egy adatbázis belülről

```
<u id="B7102.1" who="tm.1" n="1a0200">
...
<w lemma="hogy" msd="Con" ctag="C">hogy</w>
<pause/>
<vocal desc="o_hesitation" iterated="n"/> <pause/>
<w lemma="mióta" msd="Adv" phon="mióta" skel="CNBCB">
mióta</w>
<w lemma="tanít" msd="V.e3" phon="tanít" skel="CBCNC">
tanít</w>
<w lemma="ön" msd="N.NOM" phon="ön" skel="FC">
ön</w>
<pause/>
<annot
resp="enc.1" type="l_drop ba_ban" reg="iskolában">
<w
lemma="iskola" msd="N.INE" phon="isába" skel="NCCBCB">
isába
</w>
</annot>
<c lemma="?" msd="SPUNCT" ctag="SPUNCT">?</c>
</u>
```

## Az adatbáziskezelő rendszer modellje



## Egy működő felhasználói felület

**BUSZI lekérdező (használat)** Adjon meg egy lekérdezt... válasszon az alábbi lehetőségek közül

Jelenség:

Kontextus:

Prezentáció:

Interjú:

Modul:

Szerep:

Terepmunkás:

Megjegyzés:

Mehet  v0.7.2 - 2009.08.27. - O. Cs. | S. B. | Emdros

2009-10-15 16:30:15  
Lekérdezés: [Annot FOCUS typ ~ 'vh\_violation']  
Lekérdezés lókusz-jelöléssel: [Annot FOCUS who ~ '^a-d]k' and typ ~ 'vh\_violation']  
Találati szavak száma: 7 (korrigálás nélkül) - Futási idő: 3s

[1] b7313 / HUH / 392 / ak  
<(( Huh < )) ? ! Én nem [d\_drop\_prevow dictionary]\_tom , de ebből a két lányból én nem nézek ki semmit , mert [P] ezek két [hesit\_length\_n]\_lyennn vihogó , [L\_drop\_final dictionary]\_szoa [hesit\_length\_n]\_lyennn [hesit\_length\_s]\_gyerekess , hogy is mondjam ezt ? [P] [L\_drop\_final dictionary]\_szoa még a [P] [L\_drop\_final dictionary]\_gyerekessné is rosszabb [L\_drop\_iv L\_drop\_final]\_vaaho , [L\_drop\_final]\_mer [vh\_violation unspec]\_olyanokan [P] tudnak nevetni , [P] ami , [P] ami tényleg egy butaság , [L\_drop\_final dictionary]\_szoa egy a naivitásukon [ba\_ban]\_lényegébe és és [L\_drop\_final dictionary]\_szoa nem , nem tudom milyenek lesznek belőlük ? Lusták , nagyon lusták , [L\_drop\_final dictionary]\_szoa [P] nem szeretnek fizikai munkát csinálni , [L\_drop\_final dictionary]\_szoa nem szeretik azt , hogy [o\_hesitation] körülöttük rend van .

[2] b7313 / CSA / 573 / ak  
És még műteni se lehet , mert [P] ez egy pont egy olyan csont , [P] [ly\_drop\_iv]\_oan mint a [vh\_violation]\_gyiknek a fa gyiknak a farkja , hogy [P] hogy [o\_hesitation] levágnák , és utána kinő és még hegyesebb lesz .

[3] b7313 / HVE / 962 / ak