

SZÖVEGES LEJEGYZÉSBŐL NYELVI ADATBÁZIS

Oravecz Csaba és Sass Bálint
{oravecz, joker}@nytud.hu

MTA Nyelvtudományi Intézet

BUSZI I. szimpózium – 2008. december 9.

- 1 BEVEZETŐ
 - (Beszélt) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

- 1 **BEVEZETŐ**
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

NEHÉZSÉG

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

CÉL

A BUSZI szöveges lejegyzésekből olyan explicit nyelvi adatbázist készíteni, amely lehetővé teszi a számítógép felhasználását a szövegek elemzésében és lekérdezésében.

CÉL

A BUSZI szöveges lejegyzésekből olyan explicit nyelvi adatbázist készíteni, amely lehetővé teszi a számítógép felhasználását a szövegek elemzésében és lekérdezésében.

⇒ up-translation

KIINDULÓPONTOK

- **tartalmi**: BUSZI szövegek ellenőrzött lejegyzett Word fájlokban tárolt változata
- **technológiai**: korábbi hasonló (pl. ÉKsz) feladatok során kidolgozott eljárások és eszközök

- 1 BEVEZETŐ
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

MIT ÉRTÜNK ITT ADATBÁZIS ALATT?

AZ INFORMÁCIÓ

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

formában van tárolva.

- 1 BEVEZETŐ
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

KÓDOLÁSI MODELLEK

RENDEZÉSI ELV ALAPJÁN

- 1 megnyilatkozások egyes elemeinek időbeli viszonyai (ELAN, EXMARaLDA, Praat stb.)
 - STMT (Single Timeline Multiple Tiers) modell: precíz leírás az időszerkezetre vonatkozóan
 - nyelvi elemzés leírásához a modellt ki kell egészíteni
- 2 megnyilatkozások egyes elemeinek hierarchikus viszonyai (TEI)
 - OHCO (Ordered Hierarchy of Content Objects) modell: precíz leírás a nyelvi elemzésre vonatkozóan
 - az időbeli viszonyok leírásához a modellt ki kell egészíteni

KÓDOLÁSI MODELLEK

RENDEZÉSI ELV ALAPJÁN

- ① megnyilatkozások egyes elemeinek időbeli viszonyai (ELAN, EXMARaLDA, Praat stb.)
 - STMT (**Single Timeline Multiple Tiers**) modell: precíz leírás az időszerkezetre vonatkozóan
 - nyelvi elemzés leírásához a modellt ki kell egészíteni
- ② megnyilatkozások egyes elemeinek hierarchikus viszonyai (TEI)
 - OHCO (**Ordered Hierarchy of Content Objects**) modell: precíz leírás a nyelvi elemzésre vonatkozóan
 - az időbeli viszonyok leírásához a modellt ki kell egészíteni

ALKALMAZHATÓ MEGOLDÁS

Elvileg bármelyik, gyakorlatilag a transkripció fókusza szerint egyik vagy másik modell mint kiindulópont.

-

3

-

-

MIÉRT NEM JÓ AZ EREDETI ÁLLAPOT?

FORMÁTUM: DOC

- nehéz feldolgozható
- nem hordozható
- nem szabványos
- nem hatékony

Nem erre való.

B7102		
Terepmunkás: Molnár Gyula 1987.12.03.		
Lejegyezte: Borbély Anna 1993.09.03.		
Ellenőrizte: Bartha Csilla 1995.05.04.		
Ellenőrizte2: Borbély Anna 2006.05.18.		
[t felolvassa az instrukciót 1a0014-1a0149]		
B7102MUN 1a0152 1993.09.03.		
t:	Hát talán <u>először</u> is azt szeretném megkérdezni, hogy ■ ö ■ mióta tanít ön ■ is<l> <u>ába</u> <n>?	1a0200
a:	Én 1957 szeptemberétől tanítok, [sóhajt] mivel hogy ö 1957-ben végeztem az Eötvös Loránd Tudományegyetem ■ <u>Bölcsész</u> ■ karán ■ a magyar-orsz szakot.	
t:	Igen.	
a:	Tet <=tehát> ötvenhétben ■ kaptam tanári oklevelet ■ és ötvenhét ö ■ augusztus közepétől vettem állományban <ba> állományba, ■ tehát '57 ö szeptember elsejétől kezdtem tanítani.	

MIÉRT NEM JÓ AZ EREDETI ÁLLAPOT?

KÓDOLÁS:

- 1 nem egyértelmű
- 2 nem explicit
- 3 nem hatékony
- 4 ...

Nem számítógépes, hanem emberi feldolgozásra készült. (V.ö. papírszótárak használata: útmutató alapos ismerete + szócikk értelmezése)

MIÉRT NEM ELÉG JÓ AZ EMBERI FELDOLGOZÁS?

AZ EMBERI FELDOLGOZÁS

- alapos felkészülést igényel
- csak korlátozott lekérdezést tesz lehetővé
- a lekérdezés eredményét az eredeti szövegben nézegetem és értelmezem
- majdnem visszajutottunk a cédulázási módszerhez

CÉL: SZÁMÍTÓGÉPPEL SEGÍTETT ELEMZÉS

- természetesen a *szakmai elemzést* nem a számítógép végzi
- 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

-

-

-

PÉLDÁK 1.

TÖBBÉRTELMŰSÉG: <=xxx>

- általában korrekció ('nem javított nyelvbottlás'): *nem tom <=tudom>; azzz <=a>; el tu<d>om <=eltudom>*
- idegen szó leírt változat: *fon hörnzágn, <=vom Hören Sagen "hallás után mondani">; rijóba <=Rio-ba>; nyü jorkba <=New York-ba>*

PÉLDÁK 2.

IMPLICIT INFORMÁCIÓ: A MAGYAR MORFOSZINTAXIS...

készségük_□=ket
történelem_□=ből
rengeteg_□=get
összefog_□=függ
szakíró_□=ró
egyetemi_□=re
elkezdők_□=dött

A normalizált alak visszanyerésére nincs hatékony számítógépes eljárás.

00

rrreális

A jelenség teljes körű azonosítottása csak az összes típus ismeretében lehetséges. Az információra való általános keresés is minden egyes esetben az összes típusra kellene, hogy hivatkozzon.

PÉLDÁK 4.

SZABADSZÖVEGES KÓDBEVITEL

- megszoríthatlan, kézi annotáció
- elkerülhetetlen és a gépi ellenőrzés hiánya miatt rejtve maradó kódolási hibák
- ilyen mennyiségű szövegben minden lehetséges hiba elő fog fordulni, sőt még az is, ami nem lehetséges (pl. átfedő beszéd kódolása)
⇒ folyamatos javítás, karbantartás

PÉLDÁK 4.

SZABADSZÖVEGES KÓDBEVITEL

- megszorítatlan, kézi annotáció
- elkerülhetetlen és a gépi ellenőrzés hiánya miatt rejtve maradó kódolási hibák
- ilyen mennyiségű szövegben minden lehetséges hiba elő fog fordulni, sőt még az is, ami nem lehetséges (pl. átfedő beszéd kódolása)
⇒ folyamatos javítás, karbantartás

o<l>an ⇔ o<ly>an
mert<t>, most<t>
vo<l><:>t ⇔ vo<:><l>t

- 1 BEVEZETŐ
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

KIBŐVÍTETT TEI

A „KIHÍVÁS”

Más beszélt nyelvi korpuszokhoz képest a BUSZI lejegyzése igen részletes és a jelenségek olyan széles körét és típusait fedi le, melyet más adatbázisok egyáltalán nem tartalmaznak.

KIBŐVÍTETT TEI

A „KIHÍVÁS”

Más beszélt nyelvi korpuszokhoz képest a BUSZI lejegyzése igen részletes és a jelenségek olyan széles körét és típusait fedi le, melyet más adatbázisok egyáltalán nem tartalmaznak.

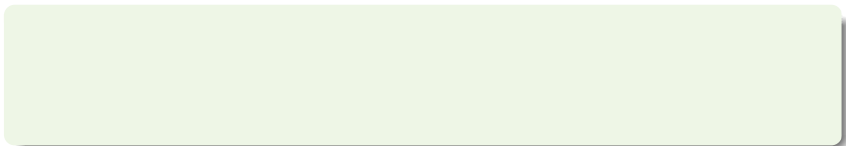
A MEGOLDÁS

Hierarchikus alapú TEI közeli XML annotáció

- a kezelendő jelenségek egységesített leírásával (egy darab típusosztályozott elem)
- az átfedő beszéd kezelésével kiegészítve.

- 1 BEVEZETŐ
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

ELEMZÉS



<w

>

szeretném

</w>

ELEMZÉS

- egyértelműsített morfológiai elemzés, szótő

```
<w lemma="szeret" msd="V.TFel"
```

```
>
```

```
szeretném
```

```
</w>
```

ELEMZÉS

- egyértelműsített morfológiai elemzés, szótő
- *regularizált* szótő CV váza, magánhangzók BNF alakban

```
<w lemma="szeret" msd="V.TFel"  
skel="CNCNC"
```

```
>  
szeretném  
</w>
```


ELEMZÉS

- egyértelműsített morfológiai elemzés, szótő
- *regularizált* szótő CV váza, magánhangzók BNF alakban
- *elhangzott* szóalak fonetikai reprezentációja

```
<w lemma="szeret" msd="V.TFel"  
skel="CNCNC"  
phon="Seretném"  
>  
szeretném  
</w>
```

BEMENET

```
<div type="modul" id="B7102MUN.1">
<head>B7102MUN 1a0152 1993.09.03.</head>
<u id="B7102.1" who="tm.1" n="1a0200">
  __W-B7102-1-1__ __W-B7102-1-2__
  __W-B7102-1-3__ __W-B7102-1-4__
  __W-B7102-1-5__ __W-B7102-1-6__
  __C-B7102-1-7__ __W-B7102-1-8__
  <pause/> <vocal desc="o_hesitation" iterated="n"/>
  <pause/> __W- B7102-1-9__ __W-B7102-1-10__
  __W-B7102-1-11__ <pause/>
  <annot
resp="enc.1" type="l_drop_prevow ba_ban" reg="iskolában">
  __W-B7102-1-12__</annot> __C-B7102-1-13__</u>
```

KIMENET

```

<div type="modul" id="B7102MUN.1">
<head>B7102MUN 1a0152 1993.09.03.</head>
<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>

```

TANULSÁGOK

- *A „géppel olvasható” formátumra alakítás a nyelvtechnológia hírhedt rabszolgamunkája — tételt most nem sikerült megcáfolni. Nem is lehet, ha*

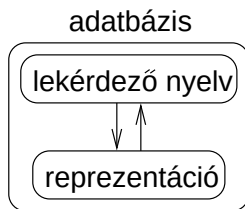
TANULSÁGOK

- *A „géppel olvasható” formátumra alakítás a nyelvtechnológia hírhedt rabszolgamunkája* — tételt most nem sikerült megcáfolni. Nem is lehet, ha
 - az adatbázis eredetileg nem erre a célra készült,

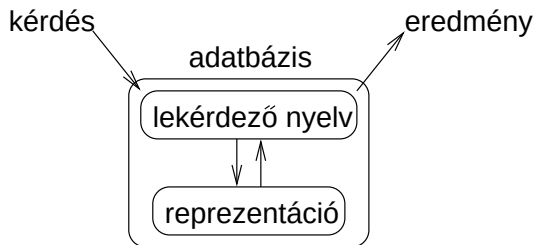
TANULSÁGOK

- *A „géppel olvasható” formátumra alakítás a nyelvtechnológia hírhedt rabszolgamunkája* — tételt most nem sikerült megcáfolni. Nem is lehet, ha
 - az adatbázis eredetileg nem erre a célra készült,
 - a triviálisnál részletesebb leíró információt tartalmaz.
- Nyelvi adatbázis készítésekor ma már ne a(z) anyanyelvi olvasó és értelmező) felhasználó legyen a célközönség. (V.ö. egy klasszikus papírszótár ma már melléktermék.)

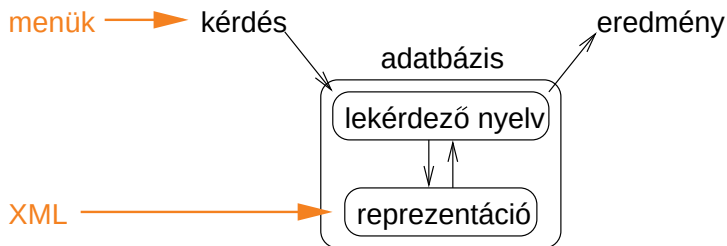
KORPUSZKEZELŐ RENDSZER MODELLJE



KORPUSZKEZELŐ RENDSZER MODELLJE



KORPUSZKEZELŐ RENDSZER MODELLJE



A VÁLASZTOTT ESZKÖZ

Emdros (2004) – <http://www.emdros.org>

Elv: a korpusz egységek (ún. **monádok**) sorozata

Ilyen egységek *bármilyen* kombinációját össze lehet vonni egy nagyobb egységbe:

- folyamatos sorozat
- kihagyás
- átfedés

PÉLDA

fontosnak tartot [P] ták az olyan

A VÁLASZTOTT ESZKÖZ

Emdros (2004) – <http://www.emdros.org>

Elv: a korpusz egységek (ún. **monádok**) sorozata

Ilyen egységek *bármilyen* kombinációját össze lehet vonni egy nagyobb egységbe:

- folyamatos sorozat
- kihagyás
- átfedés

PÉLDA

	fontosnak	tartot	[P]	ták	az	olyan
monádok:	1	2	3	4	5	6

A VÁLASZTOTT ESZKÖZ

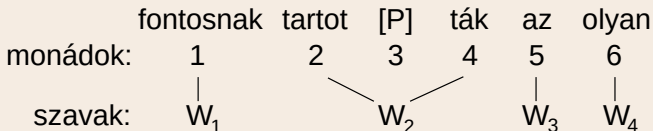
Emdros (2004) – <http://www.emdros.org>

Elv: a korpusz egységek (ún. **monádok**) sorozata

Ilyen egységek *bármilyen* kombinációját
össze lehet vonni egy nagyobb egységbe:

- folyamatos sorozat
- kihagyás
- átfedés

PÉLDA



LEKÉRDEZŐ NYELV ↔ MENÜK

Nagy kifejező erejű lekérdező nyelv

PÉLDA

```
[  
  [W FOCUS msd ~ '^Pre\.V\.' AND NOT (msd ~ 'V\.IN') ]  
] OR [  
  [W FOCUS msd ~ '^V\.' AND NOT (msd ~ 'V\.IN') ]  
  .. BETWEEN 0 AND 1  
  [W FOCUS msd = 'Pre']  
] OR [  
  [W FOCUS msd = 'Pre']  
  .. BETWEEN 1 AND 1  
  [W FOCUS msd ~ '^V\.' AND NOT (msd ~ 'V\.IN') ]  
]
```

A lekérdező nyelv összes lehetőségét
nem lehet menürendszerrel megragadni.

EGYSZERŰ LEKÉRDEZÉS

Keressük meg a ‘fontosnak’ szó előfordulásait a korpuszban!

[1] B7102 / MUN / 78 / ak

[hesit_length_n]_nnnamost azért nem jele nem jelentett nehézséget , [t_drop_final]_mer mint mondtam eléggé válogatott [o_hesitation] gyerekek [o_hesitation] jöttek ide . [P] És ezeknek a gyerekek a nagy része [P] [o_hesitation] úgy jött ide , mint ahogy most a gyerekek gimnáziumba mennek , [P] hogy [o_hesitation] [P] [o_hesitation] [P] nevezetesen azzal a szándékkal , [P] hogy tovább [o_hesitation] szeretnének tanulni majd egyetemen . [P] Tehát ők [o_hesitation] [P] **fontosnak** tartot[P]ták az olyan tárgyak tanulását is , [P] amelyek majd az egyetemen , vagy főiskolán [hesit_length_m]_nemmm lesznek [o_hesitation] [o_hesitation] már [hesit_length_m]_nemmm nem léteznek , és és [o_hesitation] és [d_drop_final]_maj nem [P] már [o_hesitation] már nem élnek , [t_drop_final]_min nevezetesen a magyar nyelv és irodalom [P] tehát az általános műveltségű [o_hesitation] műveltségűkhöz hozzátartozik . És [P] hát [o_hesitation] vő voltak annyira éreimsek , hogy [P] hogy [o_hesitation] [P] megértették azt , hogy [o_hesitation] [P] [o_hesitation] [o_hesitation] nagyon fontos az anyanyelv ismerete , [P] mert anélkül nem lehet fejlődni és tanulni . [P] Nagyon fontos az anyanyelv ismerete és idegen nyelv ismerete is nagyon fontos . [P] A természetesen a [P] [hesit_length_a]_aaa [o_hesitation] [hesit_length_o3]_első helyre teszem a matematikát , [P] a matematika , a fizika és a műszaki tárgyak ismeretén kívül és ezek mellett .

[2] B7114 / VAL / 613 / ak

Hát mit mondjak , ha az ember nem [o_hesitation] [P] [o_hesitation] teljes mértékben vagy százszázalékban a [P] [hesit_length_v]_vvvallás alapján hívó , [P] akkor az első [o_hesitation] nem tudom egy vagy kettőt aligha fogja [o_hesitation] [m_hesitation] **fontosnak** tartani , [P] de mindazt amim ami emberi és humánus benne

[3] B7207 / VAL / 251 / ak

Én **fontosnak** tartom (>> (azt) >>) .

[4] B7301 / VAL / 289 / ak

<((A vallást <)) nem tartom **fontosnak** .

[5] B7512 / QQQ / 961 / ak

<((Hát ilyen <)) szereléseket , azt tartom **fontosnak** , de azt el is mondtam .

[6] B7515 / NYE / 642 / ak

Hát **fontosnak** fontos , [P] persze .

EGYSZERŰ LEKÉRDEZÉS

er mint mondtam eléggé válogatott [o_hesitation] gyerekek [o_hesitation] gyerekek gimnáziumba mennek , [P] hogy [o_hesitation] [P] [o_hesitation] [P] Tehát ők [o_hesitation] [P] fontosnak tartot[P]ták az olyan tárgyak tanu[esitation] már [hesit_length_m]_nemmm nem léteznek , és és [o_hesitation] nyelv és irodalom [P] tehát az általános műveltségű [o_hesitation] műve[P] megértették azt , hogy [o_hesitation] [P] [o_hesitation] [o_hesitation] nyelv ismerete és idegen nyelv ismerete is nagyon fontos . [P] A termész[matikát] , [P] a matematika , a fizika és a műszaki tárgyak ismeretén kívu[vagy százszázalékban a [P] [hesit_length_v]_vvvállás alapján hívő , [P] fontosnak tartani , [P] de mindazt amim ami emberi és humánus benne

- 1 BEVEZETŐ
 - (Beszért) nyelvi adatbázis
- 2 KITEKINTÉS
- 3 AZ ADATBÁZIS KIALAKÍTÁSA
 - A lejegyzés problémái
 - A kódolási modell
- 4 HOZZÁADOTT ÉRTÉK
 - Nyelvi elemzés
- 5 KORPUSZKEZELŐ RENDSZER
- 6 FELHASZNÁLÓI FELÜLET

A FELÜLET FELÉPÍTÉSE

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

--

Megjegyzés:

Mehet

Törlés

v0.5.6 – 2008.11.03. – O. Cs. | S. B. | Emdros

← Jelenség: 'I' kiesés pótlónyúlás nélkül ... ▾ = ▾

Kontextus: a teljes megszólalás ▼

Prezentáció: gyakorisági lista ▼

Interjú:

Modul: mind

Szerep: adatközlő ▼

Terepmunkás:

A FELÜLET FELÉPÍTÉSE

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

Megjegyzés:

Mehet

v0.5.6 – 2008.11.03. – ☐ Cs. | ☐ S. B. | ☐ Emdros

Jelenség:	1' kiesés pótlónyúlás nélkül ...	=	
Kontextus:	a teljes megszólalás		
Prezentáció:	gyakorisági lista		
Interjú:	mind		
Modul:	mind		
Szerep:	adatközlő		
Terepmunkás:	mind		

jelenségek

A FELÜLET FELÉPÍTÉSE

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

Megjegyzés:

Mehet

Törölés

v0.5.6 – 2008.11.03. – [O. Cs.](#) | [S. B.](#) | [Emlős](#)

Jelenség:

'l' kiesés pótlónyúlás nélkül ...

=

Kontextus:

a teljes megszólalás

Prezentáció:

gyakorisági lista

Interjú:

mind

Modul:

mind

Szerep:

adatközlő

Terepmunkás:

mind

lekérdezés ablak

A FELÜLET FELÉPÍTÉSE

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

Megjegyzés:

Mehet

Törlés

v0.5.6 – 2008.11.03. – O. Cs. | S. B. | Endre

← Jelenség: 'I' kiesés pótlónyúlás nélkül ... ▾ | = ▾

Kontextus: a teljes megszólalás ▾

Prezentáció: gyakorisági lista ▾

Interjú: mind ▾

Modul: mind ▾

Szerep: adatközlő ▾

Terepmunkás: mind ▾

megjelenítés

A FELÜLET FELÉPÍTÉSE

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

Megjegyzés:

Mehet Törés v0.5.6 – 2008.11.03. – O. Cs. | S. B. | Emdros

Jelenség:	1. kiesés pótlónyúlás nélkül ...	=	=
Kontextus:	a teljes megszólalás		
Prezentáció:	gyakorisági lista		
Interjú:	mind		
Modul:	mind		
Szerep:	adatközlő		
Terepmunkás:	mind		

alkorpuszok

ELŐZŐ PÉLDA

BUSZI lekérdező (használat)

Adjon meg egy lekérdezést [\(Gide\)](#) ... válasszon az alábbi lehetőségek közül

Megjegyzés:

Mehet Törölés

v0.5.6 – 2008.11.03. – [O. Cs.](#) | [S. B.](#) | [Emlős](#)

← Jelenség: egy szó ... =

↳ Szóalak: fontosnak

Szótő:

Elemzés:

szótő CV-váz (teljes):

szóalak fonó-váz: (teljes): OK

Kontextus: a teljes megszólalás

Prezentáció: gyakorisági lista

ELŐZŐ PÉLDA

BUSZI lekérdező (használata)

Adjon meg egy lekérdezést (Guide) ... válasszon az alábbi lehetőségek közül

```
[W FOCUS surface = 'fontosnak']
```

← Jelenség: egy szó ... =

Kontextus: a teljes megszólalás ▾

Prezentáció: gyakorisági lista ▼

Megjegyzés:

Interjú:

Modul: mind

Szerep: adatközlő ▼

Terepmunkás:

Mehet

Törlés

v0.5.6 – 2008.11.03. – O. Cs. | S. B. | Emdros

ELŐZŐ PÉLDA

[1] B7102 / MUN / 78 / sk

[hesit_length_n]_nnamost azért nem jele nemjelentett nehézséget, [t_drop_final]_mer mint mondtam eléggé változott [o_hesitation] gyerekek [o_hesitation] jöttek ide. [P] És ezeknek a gyerekeknek a nagy része [P] [o_hesitation] úgy jött ide, mint ahogy most a gyerekek gimnáziumba mennek, [P] hogy [o_hesitation] [P] [o_hesitation] [P] nevezetesen azzal a szándékkal, [P] hogy tovább [hesit_length] szerénnek tanulni majd egyetemen. [P] Tehátők [o_hesitation] [P] **fontosnak** tartot[P]ták az olyan tárgyak tanulását is, [P] amelyek majd az egyetemen, vagy főiskolán [hesit_length_m]_nemmm lesznek [o_hesitation] [o_hesitation] már [hesit_length_m]_nemmm nem léteznek, és és [o_hesitation] és [d_drop_final]_maj nem [P] már [o_hesitation] már nem élnek, [t_drop_final]_min nevezetesen a magyar nyelv és irodalom [P] tehát az általános műveltségű [o_hesitation] műveltségűkhöz hozzátartozik. És [P] hát [o_hesitation] volt voltak annyira érelmesek, hogy [P] hogy [o_hesitation] [P] megértették azt, hogy [o_hesitation] [P] [o_hesitation] [o_hesitation] nagyon fontos az anyanyelv ismerete, [P] mert anélkül nem lehet fejlődni és tanulni. [P] Nagyon fontos az anyanyelv ismerete és idegen nyelv ismerete is nagyon fontos. [P] A természetesen a [P] [hesit_length_a]_aaa [o_hesitation] [hesit_length_o3] elsődől helyre teszem a matematikát, [P] a matematika, a fizika és a műszaki tárgyak ismeretén kívül és ezek mellett.

[2] B7114 / VAL / 613 / sk

Hát mit mondjak, ha az ember nem [o_hesitation] [P] [o_hesitation] teljes mértékben vagy százszázalékban a [P] [hesit_length_v] vvvállás alapján hívó, [P] akkor az első [o_hesitation] nem tudom egy vagy kettőt aligha fogja [o_hesitation] [m_hesitation] **fontosnak** tartani, [P] de mindazt amim ami emberi és humánus benne

[31 B7207 / VAL / 251 / sk]

Én fontosnak tartom ((> (azt)))> .

[4] B7301 / VAL / 289 / nk

<((A vallást <)) nem tartom fontosnak .

[5] B7512 / QQQ / 961 / ak

<((Hát ilyen <)) szereléseket , azt tartom **fontosnak** , de azt el is mondtam .

[6] B7515 / NYE / 642 / nk

Hát fontosnak fontos, [P] persze.

JELENSÉGEK

- Egységes keret:

$$1 \text{ jelenség} = 1 \text{ monád}$$

pl.: szünet, ööö, *l*-kiesést tartalmazó szó, *-bAn* helyett *-bA* ...

- Lekérdezés:*
mindig adott tulajdonságú monádokat keresünk.
- A kódolt jelenségeken kívül ide tartozik: *egy szó* „jelenség”
pl.: *asztal* összes alakja, *t*-végű szó, *mgh*-val kezdődő szó ...
- Jelenségek *sorozata*
→ *kihagyás* „jelenség”
pl.: *hogya* + egy szó kihagyás + ööö

JELENSÉGEK LISTÁJA

- egy szó, kihagyás
- pozícióval bíró jelenségek = kiesések (*l*, *t*, *d*, *ly*)
pozíciók: szóvégi, V•V, C•V, V•C, C•C
- pozíció nélküli jelenségek
pl.: *-bAn* helyett *-bA*, nem állítmányi *-e* ...
- önálló monádértékű elemek
pl.: szünet, megjegyzés, hezitáció ...
- egyéb: teljes megszólalás

SZÓALAPÚ – JELENSÉGALAPÚ

A korpusz *szóalapú* (monádalapú)

→ a lekérdezések adott tulajdonságú monádokat adnak vissza.

Gond:

A pozícióval bíró jelenségekből **több is lehet egy szóban**.
Nyilván szeretnénk tudni az ilyenek összesített számát.

A megoldás: *korrigálás*

Gvermekeink ? *IF*

1. *Journal of Management Studies*, 1997, 34, 1, 1-14.

Gvermekeink 2 [P]

100

[thesis length 4] L

Aztán a lányom

11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 10

MEGJELENÍTÉSI FUNKCIÓK

- (rendezett) konkordancia: a találati megszólalások listája
- gyakorisági lista
- összesítő táblázat

ALKORPUSZOK

3 független dimenzió:

- 1 interjú
- 2 modul
- 3 szerep: adatközlő vagy terepmunkás

ALKORPUSZOK

3 független dimenzió:

- 1 interjú – kvóta, terepmunkás személye
- 2 modul
- 3 szerep: adatközlő vagy terepmunkás

PÉLDÁK – DEMÓ

- egy szó: *fontosnak*
- egy jelenség: *l-kiesés*
- *hogya* + egy szó kihagyás + *ööö*
- egy szó összes alakja: szótövet adjuk meg
pl.: *mert, miért, vállalat* ...
- lekérdezőnyelv ismerete szükséges: *errő a témáró*

ÖSSZEFOGLALÁS

- Elkészült a 270000 szavas BUSZI korpusz modern korpuszkezelő rendszerbe integrált változata a hozzá tartozó lekérdezőfelülettel.
- Az **Emdros** rendszer alkalmas más hasonló részletességű annotációt tartalmazó korpusz hatékony kezelésére is.
- Hozzáférhetőség – forduljanak Kontra Miklóshoz.

ÖSSZEFOGLALÁS

- Elkészült a 270000 szavas BUSZI korpusz modern korpuszkezelő rendszerbe integrált változata a hozzá tartozó lekérdezőfelülettel.
- Az **Emdros** rendszer alkalmas más hasonló részletességű annotációt tartalmazó korpusz hatékony kezelésére is.
- Hozzáférhetőség – forduljanak Kontra Miklóshoz.

Köszönjük a figyelmet!