

SZÁMÍTÓGÉPES NYELVI ADATBÁZISOK A MAGYARÓRÁN

Sass Bálint
joker@nytud.hu

Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Korpusznyelvészeti Osztály

XVI. MANYE kongresszus
Gödöllő, 2006. április 10-12.

1 AZ E-SZÓKINCS PROJEKT

- Célkitűzés
- Keretrendszer

2 NYELVI ADATBÁZIS

- Nyelvi „erőforrások”
- Bemutató

3 ELEKTRONIKUS TANANYAG

- A tananyag elkészítése
- A tananyag szerkezete
- Példák

1 Az E-SZÓKINCS PROJEKT

- Célkitűzés
- Keretrendszer

2 NYELVI ADATBÁZIS

- Nyelvi „erőforrások”
- Bemutató

3 ELEKTRONIKUS TANANYAG

- A tananyag elkészítése
- A tananyag szerkezete
- Példák

CÉLKITŰZÉS

- Kortárs szépirodalmi nyelvi adatbázis és hozzá kapcsolódó elektronikus tananyag
- Jellemzők
 - nyelvészetileg releváns keresetőség
 - kvantitatív vizsgálatok
 - modern nyelvtechnológiai eszköztár bemutatása

KERETRENDSZER

- SDT (Sulinet Digitális Tudásbázis) kezdeményezés
- *Cél:* 12 évfolyamra kiterjedő, teljes tantervet lefedő elektronikus tananyag létrehozása
- Bárki által szerkeszthető, továbbfejleszthető
- Szaklektorálás → ún. *értékes közkinccs*
- Zárt egészet alkotó, szabadon kombinálható tananyagelemek
- Témák > foglalkozások > „lapok”

<http://sdt.sulinet.hu>

► Sulinet Digitális Tudásbázis

[illegible]

Navigation icons: back, forward, search, and other controls.

1 AZ E-SZÓKINCS PROJEKT

- Célkitűzés
- Keretrendszer

2 NYELVI ADATBÁZIS

- Nyelvi „erőforrások”
- Bemutató

3 ELEKTRONIKUS TANANYAG

- A tananyag elkészítése
- A tananyag szerkezete
- Példák

NYELVI „ERŐFORRÁSOK”

- Szövegtár
 - 20 millió szavas
 - morfológiailag elemzett korpusz

NYELVI „ERŐFORRÁSOK”

- Szövegtár
 - 20 millió szavas
 - morfológiailag elemzett korpusz
- Alaktani táblázatok
 - adott szó összes lehetséges ragozott alakja
 - igei és névszói paradigmák megállapítása
 - Elekfi László *Magyar Ragozási Szótára* alapján

NYELVI „ERŐFORRÁSOK”

- Szövegtár
 - 20 millió szavas
 - morfológiailag elemzett korpusz
- Alaktani táblázatok
 - adott szó összes lehetséges ragozott alakja
 - igei és névszói paradigmák megállapítása
 - Elekfi László *Magyar Ragozási Szótára* alapján
- Gyakorisági listák
 - szóalak / szótő / szófaj szerint
 - adott kötet / szerzői életmű / teljes korpusz tekintetében

NYELVI „ERŐFORRÁSOK”

- Szövegtár

- 20 millió szavas
- morfológiailag elemzett korpusz
<http://corpus.nytud.hu/e-szokincs/szovegtar>

- Alaktani táblázatok

- adott szó összes lehetséges ragozott alakja
- igei és névszói paradigmák megállapítása
- Elekfi László *Magyar Ragozási Szótára* alapján
<http://corpus.nytud.hu/e-szokincs/alaktan>

- Gyakorisági listák

- szóalak / szótő / szófaj szerint
- adott kötet / szerzői életmű / teljes korpusz tekintetében
<http://corpus.nytud.hu/e-szokincs/gyakorisag>

BEMUTATÓ - SZÖVEGTÁR

VMware Workstation

File Edit View VM Power Snapshot Windows Help

Home x Windows XP Home Edition x

E-Szókincs - Szövegtár - Mozilla Firefox

Eőf Szerkesztés Mőzet Ugrás Könyvjelők Eszközök Súgó

http://corpus.ny.tud.hu/e-szokincs/szovegtar/#

Hivatkozások testre... Ingyenes Hotmail Windows Windows Media

E-Szókincs - Szövegtár

Keresés indítása

1. szó: list szóra: lge ...

Igektő: ragozár: idő/mód: szám: személy:

TETSZ TETSZ múlt idő többes TETSZ

önmagában.

Egy 50 elemű véletlen minta 10 szó kontextusban.

A szövegtár kivül csak a keresett szót az alapértelmezett kód is jelenjen meg. Adatbázisunk felbukkanó ablakban.

Rendezés: az 1. szót követő szó szerint. Bibliográfiai adatok: felbukkanó ablakban.

Szerző: Károlyi László Cím: TETSZŐLEGESE

A szövegtár keresését a CQP program működteti.

http://corpus.ny.tud.hu/e-szokincs/szovegtar/#

Start E-Szókincs - Szövegtár...

HU 17:47

BEMUTATÓ - SZÖVEGTÁR

1. szófaj:

igekötő: ragozás: idő/mód: szám: személy:

Egy elemű véletlen minta szó kontextusban.

jelenjen meg. ☐ Attribútumok fe

BEMUTATÓ - SZÖVEGTÁR

VMware Workstation

File Edit View VM Power Snapshot Windows Help

Home x Windows XP Home Edition x

E-Szókincs - Szövegtár - Eredmény - Mozilla Firefox

Eőj Szerkesztés Méret Ugrás Könyvtárak Eszközök Súgó

http://corpus.nyut.hu/cgi-bin/e-szolincs/szovegtar/corpus_hun.cgi Ugrás

Hivatkozások testre... Ingyenes Hozzátel Windows Windows Media

Lekérdezés: a: [lemma = "lát" & msd = "(Pre\\.)?V\\.[TI]?Mt[123]?"] :: a.div_author = ".*Kálnoky László.*";
Találatok száma: 21 db 1,44 db / millió szó
Keresési idő: 7,38s

- szörnyalakká, / s az ezerlábú, százfejú csoda / gomolygását ott **láttuk** ige tárgyas múlt t/1 az egen / sok-sok é
- szörnyalakká, / s az ezerlábú, százfejú csoda / gomolygását ott **láttuk** ige tárgyas múlt t/1 az egen / sok-sok é
- szörnyalakká, / s az ezerlábú, százfejú csoda / gomolygását ott **láttuk** ige tárgyas múlt t/1 az egen / sok-sok é
- , és megmutatom / otthon a gyermekeimnek. Ugyanis még sohasem **láttak** ige múlt t/3 banánt életükben."
- vállalkozás. / Azonkívül mindketten csak a távoli tábornú fel-fellobbanó / lángjainál **láttuk** ige tárgyas múlt t/1 bizonytalanul egymé
- vállalkozás. / Azonkívül mindketten csak a távoli tábornú fel-fellobbanó / lángjainál **láttuk** ige tárgyas múlt t/1 bizonytalanul egymé
- nem meglejtette át. / Ki tudja, micsoda nyomorúságos szalmavackon **láttak** ige múlt t/3 feküdni utolsó / láto
- magasban / Pedig amit felmutatnak arra föl kell / tekinteni Ázt sem **láttak** ige tárgyas múlt t/3 hogy a köztéri szob
- , s nem tudtuk, mi a gyertya. / Nem **láttuk** ige tárgyas múlt t/1 meg a gazda marká
- kerepel inkább fejében. / Kocsink egy falun át rohant, **láttuk** ige tárgyas múlt t/1 tükrét egy messzi t
- lánnyról és egy francia fiúról, akik a valóságban / sohasem **láttak** ige tárgyas múlt t/1 ugyan egymást, de l
- räärő ideje, eltávoztott, / ahogyan jött. Többé sosem **láttuk** ige tárgyas múlt t/1 viszont. / Hadd tegy
- , meg-megáll. / Mindig megújulnak a minták, / de mintha **láttuk** ige tárgyas múlt t/1 volna valamennyit /
- elnéptelenedett táborban / az üres sátorba, de a sátrakhoz közeledve **láttuk** ige tárgyas múlt t/1 , hogy a táborparan
- elnéptelenedett táborban / az üres sátorba, de a sátrakhoz közeledve **láttuk** ige tárgyas múlt t/1 , hogy a táborparan
- az emlékeket. / Nem vigyáztunk e tolvaj koldusokra, / nem **láttuk** ige tárgyas múlt t/1 , hogy kérés kezük
- az emlékeket. / Nem vigyáztunk e tolvaj koldusokra, / nem **láttuk** ige tárgyas múlt t/1 , hogy kérés kezük
- az emlékeket. / Nem vigyáztunk e tolvaj koldusokra, / nem **láttuk** ige tárgyas múlt t/1 , hogy kérés kezük
- mint a begyúlt motor, / zúgott és kábult elfolyón / s **láttuk** ige tárgyas múlt t/1 , hogy kotrógép kot
- a halott. / S hogy végre megpihentek földaloid, / mert **láttak** ige tárgyas múlt t/3 , hogy nincs egy fale
- a halott. / S hogy végre megpihentek földaloid, / mert **láttak** ige tárgyas múlt t/3 , hogy nincs egy fale

Kész

Start E-Szókincs - Szövegtár... E-Szókincs - Szövegtár... HU 17:48

BEMUTATÓ - SZÖVEGTÁR

obanó / lángjainál **láttuk** *ige tárgyas múlt t/1* bizonytalanul
obanó / lángjainál **láttuk** *ige tárgyas múlt t/1* bizonytalanul
os szalmavackon **láttak** *ige múlt t/3* feküdni utols
tekinteni Azt sem **látták** *ige tárgyas múlt t/3* hogy a közté
a gyertya. / Nem **láttuk** *ige tárgyas múlt t/1* meg a gazda
y falun át rohant, **láttuk** *ige tárgyas múlt t/1* tükrét egy m
ágban / sohasem **látták** *ige tárgyas múlt t/3* ugyan egymá
ött. Többé sosem **láttuk** *ige tárgyas múlt t/1* viszont. / Ha
nták, / de mintha **láttuk** *ige tárgyas múlt t/1* volna valami
rakhoz közeledve **láttuk** *ige tárgyas múlt t/1* , hogy a tábo
rakhoz közeledve **láttuk** *ige tárgyas múlt t/1* , hogy a tábo
oldusokra, / nem **láttuk** *ige tárgyas múlt t/1* , hogy kérge

BEMUTATÓ - ALAKTANI TÁBLÁZATOK

VMware Workstation

File Edit View VM Power Snapshot Windows Help

Home x Windows XP Home Edition x

Alaktani táblázatok - Mozilla Firefox

Egy Szerkesztés Nyomtató Ugrás Kinyitójelzők Esetlövő Szűrés

http://corpus.nyit.hu/cgi-bin/e-szolincs/alaktan?id=37856;cat=MIFF%2F0.PL%2F0.POS%2F0

Hivatkozások tesztelése... Ingyenes Hotmail Windows Windows Media

lát

Toldalékolási típus: a felszólító mód más tövaltozatból, mely

tő: lát	tő: látó	tő: látók	tő: látóké
ragok	ragok	ragok	ragok
alanyeset	alanyeset	alanyeset	alanyeset
tárgyeset	tárgyeset	tárgyeset	tárgyeset
-nAk	-nAk	-nAk	-nAk
-vAl	-vAl	-vAl	-vAl
-ért	-ért	-ért	-ért
-n	-n	-n	-n
-rA	-rA	-rA	-rA
-rÖl	-rÖl	-rÖl	-rÖl
-bAn	-bAn	-bAn	-bAn
-bÖl	-bÖl	-bÖl	-bÖl
-bA	-bA	-bA	-bA
-nÁl	-nÁl	-nÁl	-nÁl
-hOz	-hOz	-hOz	-hOz
-tÖl	-tÖl	-tÖl	-tÖl
-ig	-ig	-ig	-ig
többesszám	többesszám	többesszám	többesszám
többesszám	többesszám	többesszám	többesszám
birtokos személyragok	birtokos személyragok	birtokos személyragok	birtokos személyragok
E/1	E/1	E/1	E/1
E/2	E/2	E/2	E/2
E/3	E/3	E/3	E/3
T/1	T/1	T/1	T/1
T/2	T/2	T/2	T/2
T/3	T/3	T/3	T/3

Kész

Start E-Szókincs - Szövegt... Alaktani táblázatok - ... HU 17:54

BEMUTATÓ - ALAKTANI TÁBLÁZATOK

tő: látó

ragok

alanyeset	látó
tárgyeset	látóf
-nAk	látónak
-vAl	látóval
-ért	látóért
-n	látón
-rA	látóra
-rÓl	látóról
-bAn	látóban
-bÓl	látóból
-bA	látóba
-nÁl	látónál
-hOz	látóhoz
-tÓl	látótól
-ig	látóig

többesszám

többesszám látók >

birtokos személyragok

E/1 látóm >

tő: látók

ragok

alanyeset	látók
tárgyeset	látókat
-nAk	látóknak
-vAl	látókkal
-ért	látókért
-n	látókon
-rA	látókra
-rÓl	látókról
-bAn	látókban
-bÓl	látókból
-bA	látókba
-nÁl	látóknál
-hOz	látókhoz
-tÓl	látóktól
-ig	látókig

birtokjelek

birtokjel látóké >

birtokjel látókéi >
többesítővel

tő: látóké

ragok

alanyeset	látóké
tárgyeset	látókét
-nAk	látókénak
-vAl	látókéval
-ért	látókéért
-n	látókén
-rA	látókéra
-rÓl	látókéről
-bAn	látókéban
-bÓl	látókéből
-bA	látókéba
-nÁl	látókénál
-hOz	látókéhez
-tÓl	látókétől
-ig	látókéig

BEMUTATÓ - GYAKORISÁGI LISTÁK

VMware Workstation

File Edit View VM Power Snapshot Windows Help

Home x Windows XP Home Edition x

E-Szókincs - Gyakorisági listák - Mozilla Firefox

Élő Szerkesztés Nézet Ugrás Könyvjelzők Eszközök Súgó

http://corpus.nyud.hu/cgi-bin/e-szokincs/gyakorisag?mode=pdwork=0&.cgifields=work

Hivatkozások testre... Ingyenes Hotmail Windows Windows Media

E-Szókincs - Gyakorisági listák

szófaik **gyakorisági listája**

Kiválasztott művek:

- összes
- Agh István összes
- Agh István - Ahogy a vers mibennünk**
- Agh István - A képzelet emléke
- Agh István - A madár visszajár

OK

A kiválasztott mű összesen 105323 szövegszót tartalmaz

Helyezés	Szófaj	Abszolút gyakoriság	Relatív gyakoriság (%)
1	főnév	31515	29,92
2	ige	14563	13,83
3	névelő	12768	12,12
4	melléknév	10767	10,22
5	határozószó	9142	8,68
6	kötőszó	8096	7,69
7	névmás	6646	6,31
8	ismeretlen	3231	3,07

Kész

Start E-Szókincs - Szövegt... E-Szókincs - Gyakor...

HU 18:06

A GYAKORISÁGI TÁBLÁZATOK HASZNÁLATA

- Tamási Áron és Kálnoky László szavai
 - Tamási Áronnál: a kötőszavak (*mint, ha, vagy* stb.) sokkal ritkábbak
 - leggyakoribb igék: *van, mond* ↔ *van, tud*
 - Kálnoky Lászlónál: nagyobb szókincs
- Nemes Nagy Ágnes két művének leggyakoribb főnevei
 - Szó és szótlanság: *vers, szó, költő, kép, költészet*
 - Az aranyecset: *kéz, szem, ecset, ember, nap*

1 AZ E-SZÓKINCS PROJEKT

- Célkitűzés
- Keretrendszer

2 NYELVI ADATBÁZIS

- Nyelvi „erőforrások”
- Bemutató

3 ELEKTRONIKUS TANANYAG

- A tananyag elkészítése
- A tananyag szerkezete
- Példák

A TANANYAG ELKÉSZÍTÉSE

- Több mint 24 tanórányi anyag
- Gyakorló magyarszakos pedagógusok munkája
- Magyarázatok, feladatok, tesztsorok

<http://sdt.sulinet.hu>

► Sulinet Digitális Tudásbázis (2006. tavaszától ??)

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása
- Fejezetek az alaktan és a mondattan köréből

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása
- Fejezetek az alaktan és a mondatan köréből
- Hangalak és jelentés

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása
- Fejezetek az alaktan és a mondatkörtől
- Hangalak és jelentés
- Nyelvváltozás

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása
- Fejezetek az alaktan és a mondatan köréből
- Hangalak és jelentés
- Nyelvváltozás
- Nyelvhelyesség

A TANANYAG SZERKEZETE

- Az adatbázisok használatának leírása
- Fejezetek az alaktan és a mondatban köréből
- Hangalak és jelentés
- Nyelvváltozás
- Nyelvhelyesség
- Helyesírás

TIPIKUS FELADAT

ÖSSZETETT MONDATOK – SZÖVEGTÁRHASZNÁLAT – 9. ÉVF

- 1 Keressen ki a szövegtárból összetett példamondatokat az alábbi módon!
- 2 Legyen az első keresett szó valamelyik kötőszó!
pl.: *hogy, és, is, vagy, akár, de, bár, ám, noha, mert, ezért, így, emiatt, mint, mintha* stb.
- 3 A példamondatok közül válassza ki azokat, amelyekben valóban tagmondatokat köt össze a beírt kötőszó!
Határozza meg az összetétel fajtáját!
- 4 Figyelje meg, hogy a példamondatokban gyakran csak szavakat vagy szó szerkezeteket köt össze az illető kötőszó, tehát ha kötőszavakra keresünk rá, akkor nem találunk automatikusan összetett mondatokat.

TIPIKUS FELADAT

BIRTOKVISZONY – ALAKTANI TÁBLÁZATOK HASZNÁLATA – 9. ÉVF

- 1 Az alaktani táblázatokból keresse ki különböző névszók birtokos alakjait.
pl.: *szó, fül, ház, olvasó, apa*
- 2 Ezekből válassza külön a ragos és a jeles alakokat!
- 3 Jelölje meg közülük az egyalakú illetve a többalakú toldalékokat!
- 4 Van-e a birtokviszonyt kifejező toldalékok között többjelentésű?

KÖZREMŰKÖDTEK

- Pedagógusok
 - Csábi Domonkos
 - Hamza Ágnes
 - Kulin Ágnes
 - Szűcs Gábor
 - Tilless Béláné
- Számítógépes nyelvészek
 - Váradi Tamás (osztályvezető)
 - Nagy Viktor
 - Oravecz Csaba
 - Sass Bálint

ÖSSZEGZÉS

- A bemutatott nyelvi erőforrások az interneten keresztül használhatók.
`http://corpus.nytud.hu/e-szokincs`
- Az SDT-be integrált tananyag 2006. tavaszától (??) lesz elérhető. `http://sdt.sulinet.hu`
- A tapasztalatokról várjuk visszajelzéseiket.
`joker@nytud.hu`

ÖSSZEGZÉS

- A bemutatott nyelvi erőforrások az interneten keresztül használhatók.
`http://corpus.nytud.hu/e-szokincs`
- Az SDT-be integrált tananyag 2006. tavaszától (??) lesz elérhető. `http://sdt.sulinet.hu`
- A tapasztalatokról várjuk visszajelzéseiket.
`joker@nytud.hu`

Köszönöm a figyelmet!