

Korpuszlekérdezők evolúciója

Sass Bálint

`sass.balint@nytud.mta.hu`

MTA Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály

2015. november 10.



- korpusz?
- lekérdező?

- korpusz = sok szöveg
- lekérdező?



- korpusz = sok szöveg
- lekérdező?



- korpusz = sok szöveg
- lekérdező = szavakat keresünk benne



78.o:



53.o:



- korpusz = sok szöveg
- lekérdező = szavakat keresünk benne



78.o:



53.o:



- korpusz = sok szöveg
- lekérdező = szavakat keresünk benne
- + nyelvészetileg releváns kérdéseket fogunk feltenni
- + számítógépes rendszerek



78.o:



53.o:



- korpusz = sok szöveg
- lekérdező = szavakat keresünk benne
 - + nyelvészetileg releváns kérdéseket fogunk feltenni
 - + számítógépes rendszerek

Nyelvészeti szövegkeresők fejlődése

1230: Biblia és konkordancia

Lk 14,12–14; 18,28–30; Róm. 15,26,27; 2 Kor. 8,2–4; Jel. 2,9; — ünkenes sz. lisd Evángélium tanítók.
Szentidő: Mt. 5,5; 11,29; 21,5; 2 Kor. 10,1; Gal. 5,23; 6,1; Ef. 4,2; Kol. 3,12; 1 Tim. 6,11; Tit. 3,2; Jak. 1,21; 5,13; 1 Pét. 3,4.
Szentháromság: Mt. 3,16–17; 28,19; Mk. 1,10–11; Jn 14,15–17; Róm. 1,1–4; 8,11; 2 Kor. 1,21,22; 13,13; Ef. 1,17; 1 Pét. 1,2; 1 Ján. 5,7.
Szentlélek: istensége: Ap. Csel. 5,3,4; 1 Kor. 2,6–15; 12,4–11; — sz. igazság tanítói: Jn 14,15–17,26; 15,26,27; 16,7–15; — megszentelő: Róm. 15,16–19; 1 Kor. 6,11; — ajándékai: 1 Kor. 12,1–11; 14,1–25; — élet a Lélek szerint: Gal. 5,16–25; — lélek és test ellentéte: Róm. 7,14–25; 8,1–11; — bűnök a Sz. ellen: Mt. 12,31,32; Mk. 3,28,29; — a Sz. megjelenései: Mt. 3,16; Jn. 1,32,33; Ap. Csel. 2,1–41.
Szenvedés: nevelő hatása: Zsid. 12,4–13; — Krisztushoz tesz hasonlósá: 2 Kor. 4,8–12; Fil. 3,10; 1 Pét. 4,12–19; — vigasztalás a sz.-ben: 1 Pét. 3,14,17; 2 Kor. 5,7; 4,8–5,10; — sz. a testvérekért: Kol. 1,24–29; 2 Tim. 1,8; — türelmes elviselése: Jak. 5,10; — Jézus megővődött szenvedését: Mt. 16,21–23; 17,22,23; 20,17–19; Mk. 8,31–33; 9,30–32; 10,32–34; Lk. 9,22,43–45; 19,31–34.
Szeretet: parancsa: Mt. 22,34–40; Mk. 12,28–34; Lk. 10,25–28; Jn. 13,34,35; 15,9–12; Gal. 5,13,14; Jak. 2,8; 1 Ján. 4,21; 5,2,3; 2 Ján. 4–6; — himnuszai: 1 Kor. 13,1–13; — intés a sz.-re: 2 Ján. 4–6; — felebaráti szeretet: Mt. 25,34–46; Lk. 10,30–37; Jn. 13,33–35; 15,12–17; Róm. 12,9–21; 13,8–10; Gal. 5,13–15; 1 Ján. 3,11–24; 4,7–21; 1 Pét. 1,22–25.
Szolgálat: apostoli, tanítói: Róm. 5,7; 12,4; 1 Kor. 4,1; 12,5; 2 Kor. 3,6–11; 5,18; Ef. 3,9;

4,11; — alázatos: Mt. 20,26–28; Lk. 17,10; 1 Pét. 4,11; 5,2,3; — burgói: Mt. 25,21; Lk. 12,37; Ap. Csel. 20,24; 1 Tim. 4,6; — a hívek istentisztelete: Mt. 4,10; 6,24; Róm. 7,6; Ef. 6,6; Kol. 3,24; 1 Tessz. 1,9; Zsid. 9,14; 12,28; Jel. 22,3; — testvéri sz.: Gal. 5,13; Zsid. 6,10; 1 Pét. 4,10.
Szombat: Mt. 12,1–13; 24,20; 28,1; Mk. 1,21; 2,23–28; 3,1–5; 15,42; 16,1; Lk. 6,1–10; 13,10–17; 14,1–6; 23,54–56; Jn. 5,1–15; 7,14–24; 9,14–16; 19,31; Ap. Csel. 13,14,27,42; 15,21; 16,13; 17,2; 18,4; Kol. 2,16.
Szomorúság: Mk. 10,22; Lk. 18,21; 22,45; 24,17; Jn. 16,6,20; 2 Kor. 7,9–11; 1 Tessz. 4,13; Zsid. 12,11.
Tanúság: Mt. 3,11,12; 8,4; 10,18; 15,19; 18,16; 19,18; 24,14; 26,59–61; Mk. 1,7,8; 6,11; 10,19; 14,55–59; Lk. 3,15–18; 18,20; 22,71; Jn. 1,7,15,19–34; 5,31–7,9; 9,18; 19,35; Ap. Csel. 22,18; 1 Kor. 1,6; 2,1; 2 Kor. 1,12; 2 Tessz. 1,10; 2 Tim. 1,8; Tit. 1,13; Zsid. 3,5; 11,2,5; 1 Ján. 5,6–11; 3 Ján. 12; Jel. 1,9; 6,9; 11,7; 12,11; 19,10; 20,4.
Talentumok: Mt. 25,14–30; Lk. 19,11–28.
Templomadó: Mt. 17,24–27.
Templom megisztitása: Mt. 21,12,13; Mk. 11,15–17; Lk. 19,45,46; Jn. 2,13–27.
Tisztaság: Mt. 5,8; 2 Kor. 6,16; Gal. 5,16–19,24; 1 Tessz. 4,3–5,7,8; 1 Pét. 2,11.
Tökéletesség: Mt. 5,48; 19,16–22; Mk. 10,17–22; Lk. 18,18–21; Róm. 12,2; Fil. 3,12,15; Kol. 1,28; 3,14; 4,12; Jak. 1,4,25; 3,2; 1 Ján. 2,5.
Tűzkorona: Mt. 27,27–30; Mk. 15,16–19; Jn. 19,2,3.
Tűrelem: Mt. 18,26,29; Róm. 2,4; 8,25; 15,4,5; 2 Kor. 6,4; Gal. 5,22; Ef. 4,2; Kol. 1,11; 3,12; 1 Tessz. 5,14; 2 Tessz. 3,5; 1 Tim. 6,11; 2 Tim. 4,2; Tit. 2,2; Zsid. 6,12; Jel. 2,19; 3,10.

konkordancia:
a Bibliához készült
„tárgymutató”,
hol szerepel az adott szó.

az elsőt 1230-ban készítették
domonkos szerzetesek

tartalmi, fogalmi
– nem nyelvészeti

1998:

Google!

1998:

The Google logo is displayed in its classic multi-colored font (blue, red, yellow, green, blue) with a slight 3D effect. It is centered between two thin horizontal lines.

Ugyanaz! :) Sok szövegben szavakat keresünk.

Korpuszlekérdező = nyelvészeti szövegkereső

Nyelvi tudású keresőt szeretnénk, ami nyelvi adatot szolgáltat a nyelvészeti kutatás, a magyar nyelv vizsgálata számára.

Mi benne a nyelvi tudás? Annotált.

Mitől nyelvészeti? Nyelvészetileg releváns kérdésekre válaszol.

Nem csak adott szavakat, hanem
nyelvészeti szempontok szerint megadott
szóosztályokat lehet keresni.

- körülültük, felszededegettük, elsimítottuk, végigcsináltuk
- forrásokban, fellegekben, falvakban, fejekben
- cél, csal, csaj, csel, dzsal

+ adott tulajdonságú szavak sorozatai

mi? + milyen gyakran? milyen környezetben?



a Magyar Nemzeti Szövegtár

első változata

187 millió szó

szépirodalom, hivatalos, tudományos,

sajtó, internetes fórumok

hazai és határon túli anyag

MNSZ1



100m

2005: MNSZ1 – ember

Nyelvváltozat: magyarországi, szlovákiai, kárpátaljai, erdélyi, vajdasági

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes

Lekérdezés: [lemma = "ember"] ;

Lekérdezett korpusz mérete: 181328691 szó

Találatok száma: 262868 db 1449,68 db / millió szó

Keresési idő: 0,7s

- | | | |
|-----|---|--|
| 1. | magam. Nem egy szimpatikus ember <i>N.NOM</i> | a Simicska, az tutti |
| 2. | meg) leginkább a náthás ember <i>N.NOM</i> | á betűihez lehet hasonlítani. |
| 3. | már magában is vagy százezer embert <i>N.ACC</i> | akadályozott meg abban, hogy |
| 4. | az, hogy nem szabad emberek <i>N.PL.NOM</i> | akaratának ellenszegülni, hanem az |
| 5. | a döntés, ezek az emberek <i>N.PL.NOM</i> | csak végrehajtották a kötelességüket " |
| 6. | el, és mintegy hatvanmillió ember <i>N.NOM</i> | használja. A rekordot nem |
| 7. | hisz nem azért vesz az ember <i>N.NOM</i> | ingatlant drága pénzért, hogy |
| 8. | Ivan Karamazov mondaná. Az embereket <i>N.PL.ACC</i> | korántsem az érdekli, hogyan |
| 9. | , még nem szüntettem meg ember <i>N.NOM</i> | lenni: gondolkoztam. Viselkedésem |
| 10. | különböző munkatáborokban, ebből 56 ember <i>N.NOM</i> | maradt úgy oda, hogy |
| 11. | s hogy az itt élő emberek <i>N.PL.NOM</i> | minden gondolata, babonáik, |
| 12. | Lottó a matematikához nem értő emberek <i>N.PL.NOM</i> | önkéntes adója. Nem én |
| 13. | igaz. De bizonyára minden ember <i>N.NOM</i> | számára megvan a horog, |
| 14. | hogy a zsidó állam 400 embert <i>N.ACC</i> | , a szélsőséges Hamász és |
| 15. | ilyet? Azon önmagukba fordult emberek <i>N.PL.NOM</i> | , akiket a vezetőik, |
| 16. | hogy Krisztus a földön nem ember <i>N.NOM</i> | , hanem tökéletes istenség volt |
| 17. | Holst, mielőtt legorombitod az embereket <i>N.PL.ACC</i> | , hogy nem olvasnak, |
| 18. | kisded tisztikarral, középén az ember <i>N.NOM</i> | - vagy állatvonta, a |
| 19. | buta, sot bunko muveletlen emberek <i>N.PL.NOM</i> | . Konyorgom 250 millió emberrel |
| 20. | végződött azzal, hogy az ember <i>N.NOM</i> | ... Tradoni nem " akarta |

TUDOMÁNY ÜNNEPE



2005: MNSZ1 – „fut után”

Nyelvváltozat: magyarországi, szlovákiai, kárpátaljai, erdélyi, vajdasági

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes

Lekérdezés: [lemma = "fut"] [{}{0,5} [word = "után"] ;

Lekérdezett korpusz mérete: 181328691 szó

Találatok száma: 295 db 1,63 db / millió szó

Keresési idő: 12,89s

- | | | |
|-----|---|--|
| 1. | home-videózás céljából, és most fut V.e3 | a befektetett energiája után NU. Tán így. Vigasztaljon |
| 2. | kell kezdeni - van lábuk fussanak V.Pt3 | a busz után NU! Aztán jöhetnek az öt |
| 3. | részen dalolt a gazda, futottam V.Me1 | a gazda után NU. Ugrottam, bukfenceztem is |
| 4. | azért feltápaszkodott, s újra futott V.Me3 | a királyfi után NU. Mikor még egyszer körbefutották |
| 5. | árak újra estek, mindenki futott V.Me3 | a pénze után NU, ami már rég nem |
| 6. | önkormányzatnak, amely március óta fut V.e3 | a pénze után NU. Az alpolgármester cáfolta a |
| 7. | tagjukból. A MIÉP meg fut V.e3 | a pénze után NU. Jelenleg a Röck Szilárd |
| 8. | - hitből vagy kényszerűségből - fut V.e3 | a pénze után NU. Kellemetlen a dolog Kiszelynek |
| 9. | ? Hányan vannak, akik futnak V.t3 | a pénzük és az álmaik után NU? Erre a bíróságokon, |
| 10. | Tudósítónktól 2001._augusztus_9., csütörtök Futottunk V.Mt | a pénzünk után NU - mondja az IT-Com ügyvezetője |
| 11. | őket Segesvárra. Éva hosszan futott V.Me3 | a vonat után NU. Egyik kezével integetett, |
| 12. | , a síri éjbe, fuss V.Pe2 | az eltűnt cél után NU! (Kálnoky László fordítása |
| 13. | következménye is volt, hogy futottunk V.Mt2 | az eredmény után NU. Sőt zsinórban három norvég |
| 14. | Nyirfa-ügy két főszereplője magas karriert futott V.Me3 | be a rendszerváltás után NU. Nikolits István a Népszava |
| 15. | Jöjjön velünk. - Ne fuss V.Pe2 | délibáb után NU. - Semmi okod kételkedni |
| 16. | emberek ilyenkor. Hát így futunk V.t1 | egymás után NU. Kiabálok Kukuj után. |
| 17. | ilyen elszántan, ilyen szemérmetlenül futunk V.t1 | életünk után NU. Velünk szemközt a már |
| 18. | számlát kell ellenőriztetnem, s futottam V.Me1 | Judit után NU. Éppen letette a kagylót |
| 19. | fiúra. - Én nem futok V.e1 | senki után NU, tőlem jöhet akárki, |
| 20. | szemközt oldalon megjelent Pista, futott V.Me3 | , egy szánkót húzott maga után NU. Átintett a fiúnak. |

az MNSZ1 anyaga – *más szempontból*
igék és bővítmények vizsgálata
a szórend „fölött” általánosítunk



az MNSZ1 anyaga – *más szempontból*
igék és bővítmények vizsgálata
a szórend „fölött” általánosítunk

- hagy ?-t = *mi a tárgy?*

az MNSZ1 anyaga – *más szempontból*
igék és bővítmények vizsgálata
a szórend „fölött” általánosítunk

- hagy ?-t = *mi a tárgy?*
→ kívánnivalót, nyomot, kétséget, üzenetet

az MNSZ1 anyaga – *más szempontból*
igék és bővítmények vizsgálata
a szórend „fölött” általánosítunk

- hagy ?-t = *mi a tárgy?*
→ kívánnivalót, nyomot, kétséget, üzenetet
- ? hideg hátán = *mi az ige?*

az MNSZ1 anyaga – *más szempontból*
igék és bővítmények vizsgálata
a szórend „fölött” általánosítunk

- hagy ?-t = *mi a tárgy?*
→ kívánnivalót, nyomot, kétséget, üzenetet
- ? hideg hátán = *mi az ige?*
→ végigfut, futkos, futkározik

Budapesti Szociolingvisztikai Interjú
270000 szó
részletesen lejegyzett
beszélt nyelvi korpusz
gazdag annotáció



Budapesti Szociolingvisztikai Interjú

270000 szó

részletesen lejegyzett

beszélt nyelvi korpusz

gazdag annotáció

- ... bizonyos dógokban □ mmm tát, hogy ööö
lustább annál, mint amilyennek elkép*zel*tem, ...

Budapesti Szociolingvisztikai Interjú

270000 szó

részletesen lejegyzett

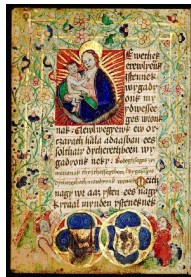
beszélt nyelvi korpusz

gazdag annotáció

- ... bizonyos dógokban □ mmm tát, hogy ööö
lustább annál, mint amilyennek elkép*zel*tem, ...
→ Majnem mindig kiesik a *d*.

2013: Ómagyar Korpusz

az összes **ómagyar kódex** szövege
2 millió szó
egységes forma, kódolás, annotáció
speciális karakterek: ý, ŷ ...
ómagyar morfológia



2013: Ómagyar Korpusz

az összes **ómagyar kódex** szövege

2 millió szó

egységes forma, kódolás, annotáció

speciális karakterek: ý, ŷ ...

ómagyar morfológia

- szórendi vizsgálat: *nem futott ki* $\leftrightarrow$ *ki nem futott*

2013: Ómagyar Korpusz

az összes **ómagyar kódex** szövege

2 millió szó

egységes forma, kódolás, annotáció

speciális karakterek: ý, ŷ ...

ómagyar morfológia

- szórendi vizsgálat: *nem futott ki* $\leftrightarrow$ *ki nem futott*
→ „Ver touaba **kŷ nem futott**” – egyenes szórend!

*összetett szavak, fonológiai jegyek, képzők
szűrés, gyakorisági listák, kollokációvizsgálat*

*összetett szavak, fonológiai jegyek, képzők
szűrés, gyakorisági listák, kollokációvizsgálat*

- megállapítás, fantazmagória, ötlet, marhaság, fantáziálás, dolog, érvelés, állítás, vád

*összetett szavak, fonológiai jegyek, képzők
szűrés, gyakorisági listák, kollokációvizsgálat*

- megállapítás, fantazmagória, ötlet, marhaság, fantáziálás, dolog, érvelés, állítás, vád
→ légbőlkapott



2014: MNSZ2: *farmerban* vagy *farmerben* (1/4)



„magas **hangrendű** szóhoz magas hangrendű toldalék,
mély és vegyes hangrendűhöz pedig
mély hangrendű toldalék kapcsolódik.” (Wikipedia)

- mély / hátulképzett: öőüű
- *neutrális*: eéíí
- magas / előlképzett: aáoóuú

bánatban – örömben

neutrális: „átmegy rajta a harmónia”: Sanyival – Bözsivel

kérdés: farmerban vagy farmerben?

a neutrális *e* miatt *farmerban*-t várunk

lekérdezések:

{con}{bck}{con}{con}{neu}{con}ban

{con}{bck}{con}{con}{neu}{con}ben

lekérdezések:

{con}{bck}{con}{con}{neu}{con}ban = 11103 db

{con}{bck}{con}{con}{neu}{con}ben = 15020 db

→	-ban	185 féle szó	csak -ban:	163 féle szó
			mindkettő:	22 féle szó
	-ben	267 féle szó	csak -ben:	245 féle szó

csak *-ban*:

kastély, Madrid, szomszéd, pótlék, kasmír, szándék

→ ezt vártuk

csak *-ben*:

két csoport látszik

- háttér, napfény, tárgyév, módszer, raktér, nagyrész, gyógyszer, csapvíz
- garden, tajpej, harlem, tandem

... és mi van középen?

csak *-ban*:

kastély, Madrid, szomszéd, pótlék, kasmír, szándék

→ ezt vártuk

csak *-ben*:

két csoport látszik

- *összetett szavak*: háttér, napfény, tárgyév, módszer, raktér, nagyrész, gyógyszer, csapvíz (itt ezt vártuk)
- garden, tajpej, harlem, tandem

... és mi van középen?

csak *-ban*:

kastély, Madrid, szomszéd, pótlék, kasmír, szándék

→ ezt vártuk

csak *-ben*:

két csoport látszik

- *összetett szavak*: háttér, napfény, tárgyév, módszer, raktér, nagyrész, gyógyszer, csapvíz (itt ezt vártuk)
- *idegen szavak*: garden, tajpej, harlem, tandem (!)

... és mi van középen?

2014: MNSZ2: *farmerban* vagy *farmerben* (4/4)

	ban	ben	ban	ben
honvéd	160	15	11x	
farmer	155	156	1x	1x
gundel	18	19	1x	1x
koffer	16	29		2x (!)
bunker	62	214		3x (!)
gallen	4	23		6x (!)
carmen	3	25		8x (!)
hamlet	6	78		13x (!)
dzsungel	30	569		19x (!)

2014: MNSZ2: *farmerban* vagy *farmerben* (4/4)

	ban	ben	ban	ben
honvéd	160	15	11x	
farmer	155	156	1x	1x
gundel	18	19	1x	1x
koffer	16	29		2x (!)
bunker	62	214		3x (!)
gallen	4	23		6x (!)
carmen	3	25		8x (!)
hamlet	6	78		13x (!)
dzsungel	30	569		19x (!)

esetleg: -ban/-ben arány ~ „idegenség”?

kérdés: **az** *öcsém* $\leftrightarrow$ *öcsém*

```
[msd=".*IGE.*" & msd!=".*_OKEP.*"] [msd="DET"] [lemma="(öcs|húg|báty|nővér)" &  
msd=".*PSe1.*"]
```

= 481

```
[msd=".*IGE.*" & msd!=".*_OKEP.*"] [lemma="(öcs|húg|báty|nővér)" & msd=".*PSe1.*"]
```

= 219

→ az esetek kb. negyede névelőtörléses

A korpuszlekérdezők fejlődésének területei az elmúlt 15 évben:

- 1 egyre nagyobb teljesítményű számítógépek
 - nagyobb korpuszok, gyorsabb lekérdezés
- 2 egyre jobb, fejlődő korpuszkezelő rendszerek
 - ezért nem egységes a használt rendszer
- 3 egyre jobb, fejlődő elemzőeszközök
 - ezeknek köszönhető az annotáció

„Nemzeti Korpuszportál”

- minden meglévő korpusz együtt, egy helyen
- minden meglévő korpuszlekérdező funkció együtt, egy helyen
minden funkció azokra a korpuszokra alkalmazva, amire lehetséges
- *valamint*: bárki hozzáadhatta a saját korpuszát, hozzárendelhesen funkciókat, és így közzétehesse

MNSZ1 és MNSZ2: <http://mnsz.nytud.hu>
Mazsola: <http://corpus.nytud.hu/mazsola>
BUSZI: <http://buszi.nytud.hu>
Ómagyar korpusz: <http://omagyarkorpusz.nytud.hu>

MNSZ1 és MNSZ2: <http://mnsz.nytud.hu>
Mazsola: <http://corpus.nytud.hu/mazsola>
BUSZI: <http://buszi.nytud.hu>
Ómagyar korpusz: <http://omagyarkorpusz.nytud.hu>

Köszönöm a figyelmet!

Sass Bálint

sass.balint@nytud.mta.hu