# Korpuszkeresés, NoSkE, Mtsz, MNSZ2, NKP

2016. június 1.
szeminárium, MTA NYTI

Sass Bálint
sass.balint@nytud.mta.hu

# Cím

**NoSkE** = korpuszkezelő rendszer (← *lényeg!*)
NoSketchEngine (régi nevén: Manatee/Bonito)
`https://nlp.fi.muni.cz/trac/noske`

**Mtsz** = Magyar történeti szövegtár
`http://clara.nytud.hu/mtsz`

**MNSZ2** = Magyar Nemzeti Szövegtár 2. kiadás
`http://mnsz.nytud.hu`

**NKP** = Nemzeti Korpuszportál
`http://corpus.nytud.hu/nkp`
= korpuszok gyűjtőoldala, innen elérhető „minden”

# 1.
# NoSkE + példa: Mtsz

# Nszt + Mtsz

A *Magyar Nyelv Nagyszótára* korpusza.
1772-2000
$\rightarrow$ bővítés: 1772-2010
= 240 év, 30 millió szövegszó

2016. március: új lekérdezőfelület a kibővített korpuszhoz

Miért?

*jelenleg:* a leggondosabban összerakott (NoSkE-s) lekérdezőfelület

*jó:* viszonylag „kicsi” (MNSZ2 = Mtsz $\times$ 26) $\rightarrow$ gyors …

# Mtsz

**Elemzetlen (!)** korpusz

– szöveg:
  Csokonai a *Földiekkel játszó* stb. éneket. 15-ben Sárosy is,

– írásjelek különválasztva (kötőjel nem!):
  Csokonai a *Földiekkel játszó* stb . éneket . 15-ben Sárosy is ,

– tokenek:
  |Csokonai|a|*Földiekkel*|*játszó*|stb|.|éneket|.|15-ben|Sárosy|is|,|

(Minden) korpusz reprezentációja: **tokenek sora**

# Token + annotáció

Alapegység: *token*

→ ezekhez lehet aztán az annotációkat hozzátenni (→ *elemzett!*):

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (0) | Csokonai | a | *Földiekkel* | *játszó* | stb | . | éneket | . |
| (1) | w | w | w | w | w | p | w | p |
| (2) | n/name | det | n | mni | abb | p | n | p |
| (3) | Csokonai | a | földi | játszik | stb | . | ének | . |
| (4) | | | title | title | | | | |

(1) szó/írásjel, (2) szófaj, (3) szótő, (4) „szövegjelleg”, bármi …

*valamint:* dokumentumhoz rendelt annotáció = *metaadat*

szó-annotáció ↔ struktúra-annotáció

# Az Mtsz felülete

egyszerű keresés: *de viszont*

*Ami látszik:*

– nagybetű/kisbetű nem számít – sőt: f
– strukturális információk (oldal, bekezdés, (vers)sor): zölddel
– találatok időrendben

*Ami nem látszik:*

– évszám katt = részletes bibliográfiai adatok
– találat katt = nagyobb kontextus

# NoSkE funkciók

– alkorpuszok – *minden metaadatból automatikusan!* (Baróti, 1808)

– mentés – *összes találat!* (sorok max. száma)

– megjelenítés – `<doc>`, `<oldal>`, `<par>`, `<ital>`, `<br>` (Ctrl!)

– rendezés – *jobb* (vesszők)

– véletlen minta

– **szűrés** – *1..1* (vessző)

– **gyaklisták** – *szóalakok, évszámok,* $1R$

– kollokációk ($\rightarrow$ *se, sem, ne, nem, nincs, nélkül*)

– **CQL = Corpus Query Language – formális lekérdezőnyelv**

$\rightarrow$ használatával tárhatjuk fel a korpuszban rejlő teljes információt!

elemzett korpusznál is hasznos, de *elemzetlennél nagyon kell!*

az így megfogalmazott kérdésre alkalmazható az összes fenti funkció

# Pozíciók szűréshez és gyaklistához

keresett kifejezés: *viszont*

| | Ám | de | viſzont | hallá | , | hogy | majd | a | ’ | Trójai | vérből |
|---|---|---|---|---|---|---|---|---|---|---|---|
| szűrés ablak | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| gyaklista pozíció | 2L | 1L | [Node] | 1R | 2R | 3R | 4R | 5R | 6R | 7R | 8R |

szűrés ablak (lehet több token):
-1..1 = de viſzont hallá
1..3 = hallá , hogy
1..1 = hallá

gyaklista pozíció (itt csak 1 token!):
1L = de
1R = hallá

# Pozíciók szűréshez és gyaklistához – advanced

keresett kifejezés: *de viszont* (← többszavas!)

| | Ám | de | viſzont | hallá | , | hogy | majd | a | ’ | Trójai | vérből |
|---|---|---|---|---|---|---|---|---|---|---|---|
| szűrés ablak eleje | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| szűrés ablak vége | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| gyaklista pozíció | 1L | [....Node....] | | 1R | 2R | 3R | 4R | 5R | 6R | 7R | 8R |

**(!)** A szűrés ablak végét a találat *végéhez* viszonyítja! → így: -1 = 1L és 1 = 1R

szűrés ablak (lehet több token):
-1..1 = Ám de viſzont hallá
1..3 = viſzont hallá , hogy
1..1 = viſzont hallá **(!)**
2..1 = hallá **(!)**

gyaklista pozíció (itt csak 1 token!):
1L = de
1R = hallá

(beállítás a szűrésnél: "első" + "találati szót beleértve"!)

# Többszavas lekérdezés vagy szűrés? – advanced

*Ha többszavasra keresünk:*
annak a részeiből nem tudunk gyaklistát készíteni (*Node*).
De az egészből és a hozzá képest vett $n$-edik szóból igen.

*Ha egy szóra keresünk + szűrés:*
csak az első szóhoz képest $n$-edik szóból tudunk gyaklistát készíteni.
Az itt-ott megjelenő „szűrésből kijött” szavakból nem.

Mindig végig kell gondolni: éppen melyik megközelítés a hasznos.

*Lehetőség:* többszavast így felépíteni: egy szó + 1..1, 2..2 szűrés
$\rightarrow$ és akkor lehet gyaklistát csinálni a részeiből.

# CQL – reguláris kifejezések (regkif)

Bizonyos tulajdonságú karaktersorozatok megadására.

*Speciális jelentésű karakterek:*

| | |
|---|---|
| . | tetszőleges karakter |
| * | a megelőző karakterből 0 vagy több |
| + | a megelőző karakterből 1 vagy több |
| ? | a megelőző karakterből 0 vagy 1 |
| [ab] | ’a’ vagy ’b’ karakter |
| [^ab] | nem ’a’ és nem is ’b’ karakter |
| r\|s | ’r’ vagy ’s’ reguláris kifejezés |
| (..) | egybefoglalás |
| \ | a követő karakter „escape”-elése |

| | | |
|---|---|---|
| (1) alma | (4) **nélk[üűúű]l** | (7) alma\|almá.* |
| (2) tejf.l | (5) .* | (8) \. |
| (3) mondjá(to)?k | (6) .*bb | (9) **([Aa]\|[Aa]z\|Ee]gy)** |

(Kevesebb karakterrel?)

# CQL (Corpus Query Language)

| | |
|---|---|
| [..] | egy tokenre vonatkozó megkötések |
| [..]*op* | egy tokenre vonatkozó operátorok: *op* = * ? + {n,m} |
| x="y" | *x* attrib értéke legyen *y* – Mtsz: csak 1 attrib van, a *word* |
| x!="y" | *x* attrib értéke *ne* legyen *y* |
| & | *és* kapcsolat megkötések között |
| <s> | strukturális elem: mondat eleje |

(1) [] []
(2) [word="majd"]
(3) "majd"
(4) [word!="a.*"]
(5) []{0,5}
(6) <s> [word="[Nn]em"] [word="kellett"] [word="volna]? [word=".*ni"]

*Regkif 2 szinten:* attribútumértéken belül + tokenek szintjén

((4) másképp? (6) kérdőjel belülre?)

# 1. példa: tárgy + ige

*Feladat.* Keressünk ilyet: *tárgyesetű szó + múltidejű E/3 ige!*

# 1. példa: tárgy + ige

*Feladat.* Keressünk ilyet: *tárgyesetű szó + múltidejű E/3 ige!*

```
".*t"  ".*tt"
```

# 1. példa: tárgy + ige

*Feladat.* Keressünk ilyet: *tárgyesetű szó + múltidejű E/3 ige!*

```
".*t" ".*tt"
```

*most itt – ???*

```
".*t" [word=".*tt" & word!="(itt|alatt)"]
```

# 1. példa: tárgy + ige

1. CQL: `".*t" ".*tt"`
2. Gyakoriságok / szóalakok
3. $p \rightarrow$ erőt vett
4. Milyen szó jön utána? $\rightarrow$ Gyakoriságok: 1R
5. $p \rightarrow$ rajta
6. Rendezés / jobb $\rightarrow$ hogy *mi* vesz erőt rajta

$\rightarrow$ félelem, féltékenység, habozás, kacagás, kishitűség, kíváncsiság …

# 2. példa: alanyesetű melléknév

Nincs fogodzó …

# 2. példa: alanyesetű melléknév

Nincs fogodzó . . . *csak a kontextusban!*

*-bAn* = leggyakoribb esetrag: `".*b[ae]n"` → főnevek
(esetleg: -rA, -vAl ↔ nem jó: -t, -nAk)

**1L gyaklista** → nem valami jó . . .

szűrés: -2..-2 `"([Aa]z?|[Ee]gy)"`

1L gyaklista → egész jó
(1-2 birtokos: ember, világ, nm-k . . . kizárni hogy lehetne?)

- szomszéd – nem főnév, melléknév!
- mult – helyesírási hibás!

# 3. példa: honnan a Csokonais példa?

Csokonai a *Földiekkel játszó* stb . éneket . 15-ben Sárosy is ,

*Naná:* korpuszból kerestem ki. Hogyan?

```
"stb" "\."
```

*konstruált példa ↔ élő példa:*

két ló húzza a szekeret
mint a hogy húzza a vetőgépet a ló, és a jármot az ökör

a Győr-Moson-Sopron megyeiek tettek bele rendkívül sok pénzt
olcsó az alma, rendkívül sok termett

# 3. példa: honnan a Csokonais példa?

Csokonai a *Földiekkel játszó* stb . éneket . 15-ben Sárosy is ,

*Naná:* korpuszból kerestem ki. Hogyan?

```
"stb" "\."
```

*konstruált példa* $\leftrightarrow$ *élő példa:*

két ló húzza a szekeret
mint a hogy húzza a vetőgépet a ló, és a jármot az ökör

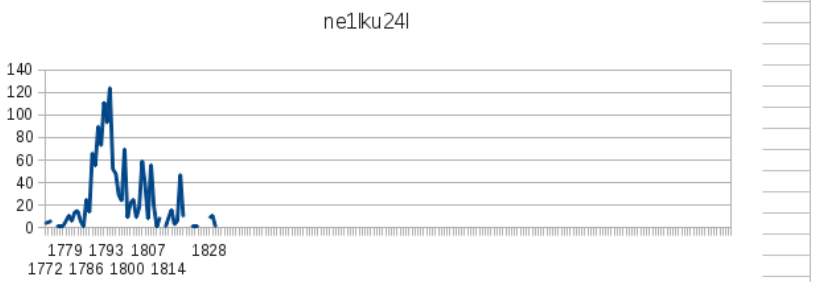
a Győr-Moson-Sopron megyeiek tettek bele rendkívül sok pénzt
olcsó az alma, rendkívül sok termett (0!)

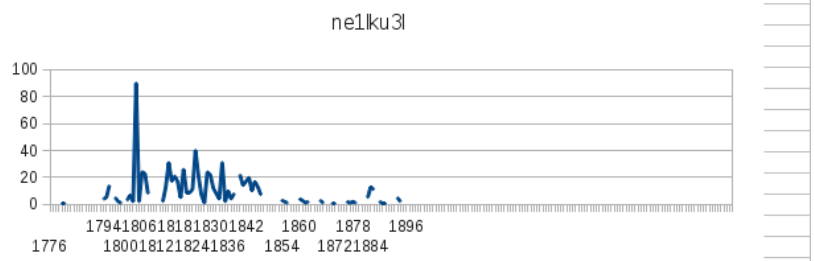
**Korpusz = élő, valódi nyelvhasználat.**
**Nyelvi példákat korpuszból!**

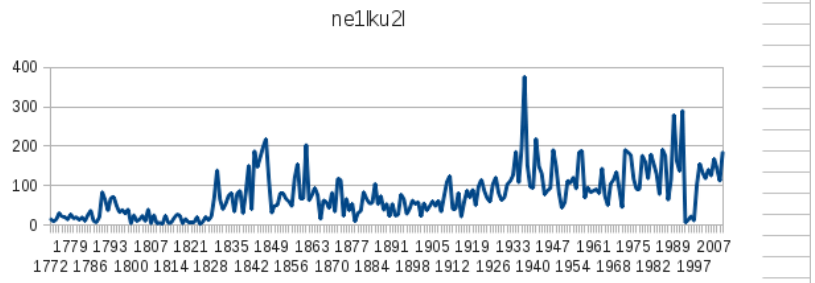
# 4. példa: *nélkül* helyesírása

diakrón vizsgálat

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | ne1lku24l | | ne1lku3l | | ne1lku2l | |
| 2 | 1772 | 4 | | | 1772 | 16 |
| 3 | 1773 | 5 | | | 1773 | 9 |
| 4 | 1774 | 6 | | | 1774 | 14 |
| 5 | | | | | 1775 | 31 |
| 6 | 1776 | 1 | 1776 | 1 | 1776 | 22 |
| 7 | 1777 | 1 | | | 1777 | 20 |
| 8 | 1778 | 1 | | | 1778 | 14 |
| 9 | 1779 | 6 | | | 1779 | 28 |
| 10 | 1780 | 11 | 1780 | 1 | 1780 | 17 |
| 11 | 1781 | 6 | 1781 | 1 | 1781 | 20 |
| 12 | 1782 | 14 | | | 1782 | 13 |
| 13 | 1783 | 15 | | | 1783 | 20 |
| 14 | 1784 | 6 | | | 1784 | 10 |
| 15 | 1785 | 1 | 1785 | 1 | 1785 | 26 |
| 16 | 1786 | 25 | | | 1786 | 37 |
| 17 | 1787 | 14 | | | 1787 | 10 |
| 18 | 1788 | 66 | | | 1788 | 7 |
| 19 | 1789 | 55 | | | 1789 | 20 |
| 20 | 1790 | 90 | 1790 | 5 | 1790 | 84 |
| 21 | 1791 | 73 | | | 1791 | 64 |
| 22 | 1792 | 111 | 1792 | 1 | 1792 | 37 |
| 23 | 1793 | 93 | | | 1793 | 69 |
| 24 | 1794 | 124 | 1794 | 4 | 1794 | 72 |
| 25 | 1795 | 52 | 1795 | 5 | 1795 | 49 |
| 26 | 1796 | 48 | 1796 | 14 | 1796 | 32 |
| 27 | 1797 | 29 | | | 1797 | 39 |
| 28 | 1798 | 24 | 1798 | 5 | 1798 | 29 |
| 29 | 1799 | 70 | 1799 | 2 | 1799 | 40 |
| 30 | 1800 | 9 | 1800 | 1 | 1800 | 4 |
| 31 | 1801 | 22 | | | 1801 | 26 |
| 32 | 1802 | 25 | 1802 | 3 | 1802 | 10 |
| 33 | 1803 | 9 | 1803 | 7 | 1803 | 14 |
| 34 | 1804 | 18 | 1804 | 2 | 1804 | 24 |
| 35 | 1805 | 59 | 1805 | 90 | 1805 | 10 |
| 36 | 1806 | 38 | 1806 | 2 | 1806 | 40 |
| 37 | 1807 | 8 | 1807 | 24 | 1807 | 4 |
| 38 | 1808 | 56 | 1808 | 23 | 1808 | 26 |
| 39 | 1809 | 19 | 1809 | 8 | 1809 | 4 |
| 40 | 1810 | 1 | | | 1810 | 6 |
| 41 | 1811 | 9 | | | 1811 | 3 |
| 42 | | | 1812 | 1 | 1812 | 25 |
| 43 | 1813 | 1 | | | 1813 | 5 |
| 44 | 1814 | 9 | 1814 | 2 | 1814 | 6 |
| 45 | 1815 | 16 | 1815 | 13 | 1815 | 18 |
| 46 | 1816 | 3 | 1816 | 31 | 1816 | 27 |



# 2.
# MNSZ2 + „Minden találat kell!”

# MNSZ2

A „mai magyar írott köznyelv reprezentatív korpusza” kíván lenni.

2016. február: 785 millió szövegszó (= Mtsz $\times$ 26) – v2.0.3

méretéből adódóan sok esetben lassú (gateway timeout! `"m.*"`)
ami gyors: szóalak, szótő, CQL $\leftrightarrow$ egyszerű keresést ne!

struktúrák és metaadatok kevésbé kidolgozottak

*viszont:* **elemzett!** = plusz attribútumok

(vö: Mtsz megjelenítés $\leftrightarrow$ MNSZ2 megjelenítés)

# MNSZ2 – attribútumok

| | | |
|---|---|---|
| (1) | word | **szépet** |
| (2) | lemma | szép |
| (3) | msd | MN.ACC |
| (4) | ana | compound=n;;hyphenated=n;;stem=szép::MN;; morphemes=et::ACC;;mboundary=szép+et |
| (5) | word_cv | CNCNC |
| (6) | word_syll | 2 |
| (7) | lemma_cv | CNC |
| (8) | lemma_syll | 1 |
| (9) | word_phon | Sépet |
| (10) | lemma_phon | Sép |

*Példa:* `"szé.*"`

**Mind ugyanúgy használható, mint az Mtsz-ben a *word*!**

(az attribútumoknak megfelelően vannak újabb gyaklista-típusok is)

# MNSZ2 – részletes keresés

plusz szolgáltatás

kattingatással állítjuk össze a kívánt lekérdezést
$\rightarrow$ a háttérben persze CQL lesz belőle

Az elemzésnek köszönhetően …

*morfológia:*
– körülültük, felszedegettük, elsimítottuk, végigcsináltuk, …

*fonológia:*
– cél, csal, csaj, csel, dzsal, …

Részletes kereséssel lehet szűrni!

# „Minden találat kell!” elv (4/4)

## Annotáció és fedés

*gond:* ha hibás az annotáció → csökken a fedés (*pl.: barát* WSD)

Nem szabad vakon bízni a korpusz annotációjában.
Tudatosítsuk, hogy konkrétan mennyire bízhatunk benne.
*előfordulhat:* nem kellően jó a korpusz-annotáció, amire építünk.

El kell gondolkodni azon, hogy adott kérdésre az annotáció választ tud-e adni.
Ha embernek is nehéz eldöntenie, akkor a géptől ne várjuk.
Adott esetben akár hagyjuk figyelmen kívül az annotációt!

*pl.: elkészített* – melléknévi igenév *vs.* múlt idejű ige

**Ne várjuk, hogy a korpusz annotációja tökéletes lesz.**
**Azt végképp ne, hogy pont az aktuális kutatási kérdésünket fogja automatikusan megválaszolni. Ha mégis, örüljünk!**

# 4.
# Feladatok

# Feladatok

1. igemódosítót tartalmazó tagadó mondatok (Gugán Kati)
   „*nem* + egyesével néhány igekötő”
2. ilyen mondatot szeretnék:
   *Ursula szemére vetette Ralphnak a kétszínűségét.* (mi kinek a micsodája?)
3. Mik a *munka* tipikus jelzői?
4. „mindig alsó nyelvállású kötőhangzóval jár: *-abb/-ebb*, ennek csak az amúgy is kivételes, mert nem nyitó *nagy* melléknév áll ellen: *nagyobb*.” (nyest.hu)
   → Ellenőrizzük!
5. Mennyire jó a *szomszéd* fn/mn annotációja az MNSZ2-ben?
6. Igekötős ige összes (nem elváló és elváló) alakjának keresése
7. Ikes feltételes ragozás (*aludnám, aludnék, aludna*) diakrón változása