Keresés korpuszban

Sass Bálint
Nyelvtudományi Kutatóközpont
sass.balint@nytud.hu

Témák

NoSkE = NoSketchEngine – korpuszkezelő rendszer (← *lényeg !*)
Mtsz = Magyar történeti szövegtár – *elemzetlen*

MNSZ2 = Magyar Nemzeti Szövegtár – *elemzett*
Mazsola, **Ómagyar Korpusz**, **BUSZI** . . .

NKP = Nemzeti Korpuszportál
`http://corpus.nytud.hu/nkp` ← itt minden megtalálható

A korpuszkeresés elvei

Példák az MNSZ2 logból
Feladatok

0.
Korpusz

elemzetlen **korpusz**

pl.: **Mtsz**

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

→ valahogy felvagdossuk:
trivi: szóközök mentén, vagy
kicsit okosabb: írásjelek különválasztva

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

korpusz = **tokenek sora**

elemzett korpusz

szóalak	szó/új	szófaj	szótő	jelleg
Csokonai	w	n/name	Csokonai	–
a	w	det	a	–
Földiekkel	w	n	földi	title
játszó	w	mni	játszik	title
stb	w	abb	stb	–
.	p	p	.	–
éneket	w	n	ének	–
.	p	p	.	–

elemzett korpusz = **tokenek + annotáció** $\sim$ **táblázat**

elemzett korpusz

szóalak	szó/új	szófaj	szótő	jelleg
<s>				← struktúra
...				
Csokonai	w	n/name	Csokonai	–
a	w	det	a	–
Földiekkel	w	n	földi	title
játszó	w	mni	játszik	title
stb	w	abb	stb	–
.	p	p	.	–
éneket	w	n	ének	–
.	p	p	.	–
</s>				

elemzett korpusz = **tokenek + annotáció** $\sim$ **táblázat**

elemzett korpusz

szóalak	szó/új	szófaj	szótő	jelleg	
<s id="5">					← **struktúra-annotáció,** *metaadat*
...					
Csokonai	w	n/name	Csokonai	–	← **szó-annotáció**
a	w	det	a	–	
Földiekkel	w	n	földi	title	
játszó	w	mni	játszik	title	
stb	w	abb	stb	–	
.	p	p	.	–	
éneket	w	n	ének	–	
.	p	p	.	–	
</s>					

elemzett korpusz = **tokenek + annotáció** $\sim$ **táblázat**

1.
NoSkE + Mtsz

Nszt + Mtsz

A *Magyar Nyelv Nagyszótára* korpusza.

1772-2010 = 240 év, 30 millió szövegszó

Miért?

– gondosan összerakott (NoSkE-s) lekérdezőfelület

– *alkalmas:* viszonylag „kicsi” (MNSZ2 = Mtsz $\times$ 35) $\rightarrow$ gyors ...

NoSkE felület = az Mtsz felülete

egyszerű keresés: *de viszont* **(1. példa)**

Ami látszik:
- nagybetű/kisbetű nem számít – sőt: ſ
- 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 – struktúrák – `<oldal>`, . . . `<g>`; infó – szó sorszáma (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

CQL – reguláris kifejezések (regkif, regex) (fut)

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)`**

((9) kevesebb karakterrel? Hiba?)

CQL (Corpus Query Language) (mondat)

`[..]`	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"]`

→ **regex két szinten:** attribútumértéken belül + tokenek szintjén

((4) másképp? (6) kérdőjel belülre? Hiba?)

2. példa: tárgy + ige

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

2. példa: tárgy + ige

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

```
".+t" ".+tt"
```

2. példa: tárgy + ige

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

```
".+t" ".+tt"
```

most itt –???

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

2. 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 ...

3. példa: alanyesetű melléknév

Nincs fogodzó …

3. 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 …

3. 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!

4. példa: fog + FNI

Feladat. Készítsünk gyakorisági listát
a *fog*-tól jobbra 1, 2 vagy 3 szó távolságban lévő FNI-kből.

4. példa: fog + FNI

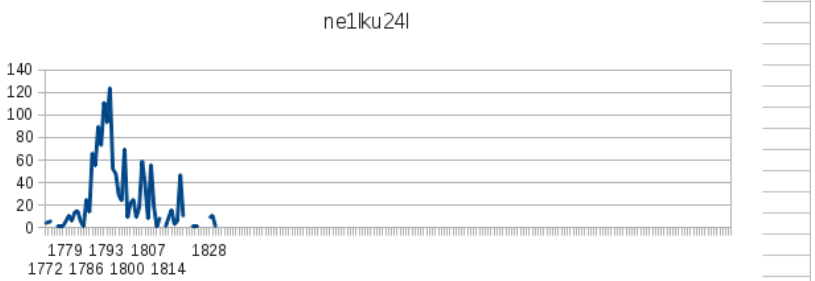
Feladat. Készítsünk gyakorisági listát
a *fog*-tól jobbra 1, 2 vagy 3 szó távolságban lévő FNI-kből.

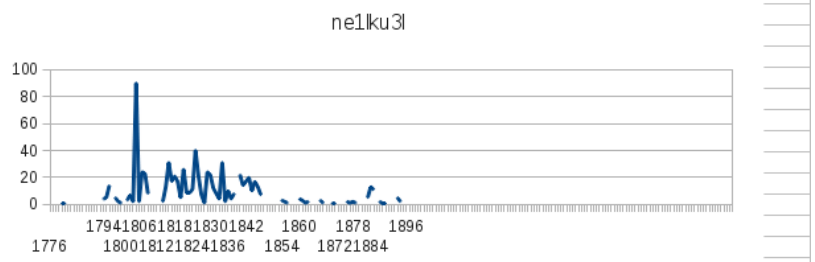
Ez a jó sorrend:
FNI (`" . *ni "`) + szűrés:-3..-1 *fog*

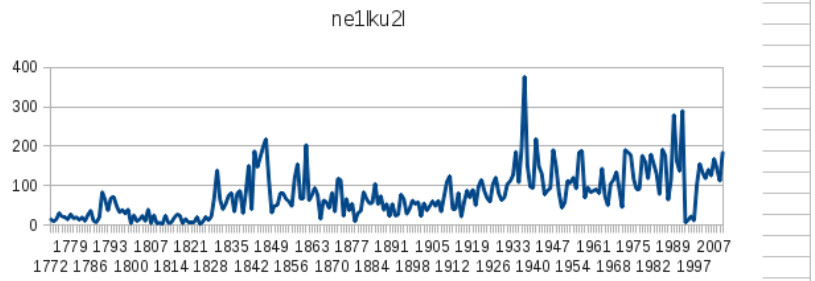
5. példa: *nélkül* helyesírása

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

MNSZ2

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

1,04 milliárd szövegszó (= Mtsz $\times$ 35) – v2.0.5

méretéből adódóan sok esetben lassú (gateway timeout! `"m.*"`)
ami gyors: szóalak, szótő, CQL $\leftrightarrow$ egyszerű keresést ne!
kisbetű/nagybetű eltér:
`[word="nem"]` $\leftrightarrow$ `[word="[Nn]em"]` $\leftrightarrow$ `[word="(?i)nem"]`

metaadatok kevésbé kidolgozottak

viszont: **elemzett!** = plusz attribútumok

(vö: Mtsz megjelenítés $\leftrightarrow$ MNSZ2 megjelenítés, reg?)

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

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

példa: `[lemma="szép"]` – *példa:* `[lemma_cv="CBCCNC"]`

(az attribútumoknak megfelelően vannak újabb gyaklista-típusok is, ana...)

MNSZ2 – leendő attribútumok

form	lemma	xpostag	compound	prev	previd	prevpos
kapaszkodik	kapaszkodik	[/V][...]	kapaszkodik			
→						
kapaszkodik	kapaszkodik	[/V][...]	kapaszkodik			
eloldozódott	eloldozódik	[/V][...]	el#oldozódik			
→						
eloldozódott	eloldozódik	**[/Prev]**[/V][...]	el#oldozódik	**pfx**		
tér	tér	[/V][...]	tér			
vissza	vissza	[/Prev]	vissza			
→						
tér	**visszatér**	**[/Prev]**[/V][...]	**vissza#tér**	**sep**	**7**	**+1**
vissza	∅	[/Prev]	∅	**conn**	**7**	
haza	haza	[/Prev]	haza			
akarok	akar	[/V][...]	akar			
menni	megy	[/V][Inf]	megy			
→						
haza	∅	[/Prev]	∅	**conn**	**26**	
akarok	akar	[/V][...]	akar			
menni	**hazamegy**	**[/Prev]**[/V][Inf]	**haza#megy**	**sep**	**26**	**-2**

igekötős példák

1. igekötős ige összes találata:
   CQL: `[lemma="előjön"]`

2. elvált alakjai:
   CQL: `[lemma="előjön" & prev="sep"]`

3. összes igekötős ige:
   CQL: `[xpostag="\[/Prev\]\[/V\].*"]`

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 is lehet szűrni!

kiss ottó: lepereg

messzire távoli távoli senkije
távoli semmibe csillaga rózsafa
kellene hallani zongora belseje
gyermeki nagymama tartani kellene

messzire mondani mennyire bökdösi
kezdeni kellene mennyire belseje
mennyire mondani mennyire holmira
messzire hordani hajnali városi

csuklani dallama dallama nápolyi
tartani nénire lakhelye semmire
mennyire hajdani hajdani démoni
fölveszi mesteri gyűlöli majdani

„Automatikus” versírás

részletes keresés / fonetikai tulajdonságok használata

szóalak =
`{con}(({lng}|{sht}{con}){con}{sht}{con}{sht}`

NoSkE – parancssoros hozzáférés

```
corpquery
```

```
corpquery
  /home/corpora/MNSZ2
  '[lemma="aszfalt"]'
  -a word,lemma,msd
  -c 3
```

MNSZ2: `clara.nytud.hu`

Eredmény:

```
#162523 jólesően /jólesően/HA  csoszogott /csoszog/IGE.Me3
az /az/DET < aszfalton /aszfalt/FN.SUP > . /./SPUNCT
</p></s><s><p> A /a/DET  madár /madár/FN.NOM
```

3.
Mazsola

Mazsola

igék bővítményszerkezetének vizsgálatára

reprezentáció:
A lány vállat vont. → `ige=`von `alany=`lány `tárgy=`váll

felület …

példák:

- *eszik -t*
- *hagy -t*
- *hideg hátán* – „kifordított” keresés: igére
- *erőt vesz rajta vmi* – csináljuk meg jobban!:)

4.

Ómagyar, BUSZI

Ómagyar Korpusz

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

másik korpuszkezelő rendszer: *Emdros*

Ómagyar Korpusz – Emdros

másik korpuszkezelő rendszer: *Emdros* (`emdros.org`)
saját lekérdezőnyelv: MQL – infó: MQL Query Guide

példák:

– *jonh* – normalizált eleje

```
[W FOCUS w_4 ~ '^4\(\(jonh']
```

  – hasonlít a CQL-re – `[..]` az egy egység
  – több egységet egymás után lehet tenni (beágyazni is lehet!)
  – ~ operátor = regex illesztés
  – kódokat próbalekérdezésekből lehet kitalálni: `w_6e`; *nem* → `Adv`

– *nem* – gyaklista

BUSZI

Budapesti Szociolingvisztikai Interjú

270000 szó

részletesen lejegyzett *beszélt nyelvi* korpusz

gazdag annotáció

Emdros

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

(külön papíron regisztrálni szükséges)

5.
A korpuszkeresés elvei

A korpuszkeresés elvei #1

1. Nyelvi példákat korpuszból!
   Korpusz = élő, valódi nyelvhasználat.
2. Minden találat kell!
3. Ne bízzunk vakon az annotációban!
4. „Alap” korpuszkészlet.
5. Nézzük meg a korpuszban!

Nyelvi példákat korpuszból 1.

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

Nyelvi példákat korpuszból 1.

konstruált példa $\leftrightarrow$ *élő példa:*

két ló húzza a szekeret (ÉKSz)
mint a hogy húzza a vetőgépet a ló, és a jármot az ökör (Mtsz)

a Győr-Moson-Sopron megyeiek tettek bele rendkívül sok pénzt (MNSZ2)
olcsó az alma, rendkívül sok termett (0...)

Nyelvi példákat korpuszból 1.

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

Honnan a példa?

Naná: korpuszból kerestem ki. Hogyan?

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

Korpusz = élő, valódi nyelvhasználat.
Nyelvi példáinkat vegyük korpuszból!

„Minden találat kell!” elv 2.

két alapvető cél: példák (pontosság) ↔ statisztikai vizsgálat (fedés)

a korpuszlekérdezők célja: hogy a felhasználó az *összes* találatot megkapja arra a kérdésre, amire a felület használata közben *gondolt*.:)

másképp: magas fedés kell! ↔ alacsony pontosság nem annyira gond

- *hogy* esetén: *hoyg* (3174)?
- *tejföl* (2566) → visszaadjuk a *tejfel*-t is (479 = 16%)?
- *bokor* → *bokrok*?
- ómagyar: *majd* → *maÿd* biztosan kell. Kérdés: *majdan*?

Mit szeretne a felhasználó?

Legyen külön kapcsoló minden jelenségre?
e/ö, helyesírási hibák, régies alak, ragozott alak ...

„Minden találat kell!” elv 2.

Megoldás lenne elvben: **normalizálás**

$\sim$ vö: kitalálni, amit a felhasználó látni szeretne.

A normalizálás arra szolgál, hogy a lekérdezésre vetítse
az összes olyan korpusz-tokent, ami rá illik/illeszthető.

Hogy találjuk ki mit szeretne a felhasználó?

ötlet: „nyelvészetileg” releváns-e az adott különbség vagy nem?
$\rightarrow$ Ha nem, akkor normalizáljuk = azonos alakra hozzuk!
De el lehet-e ezt dönteni?

Az *eredeti* felszíni alak biztosan meghagyandó.

A nem tökéletes annotáció elve 3.

Annotáció és fedés

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

Ne bízzunk vakon a korpusz annotációjában, tartalmazhat hibákat.
Tudatosítsuk, hogy konkrétan mennyire bízhatunk benne.

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

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

> **Ne várjuk, hogy a korpusz annotációja tökéletes lesz.**
> **Ne várjuk, hogy pont az aktuális kutatási kérdésünket fogja automatikusan megválaszolni.**

Használjuk a meglévő annotációt kreatívan!

Nemzeti Korpuszportál (NKP)

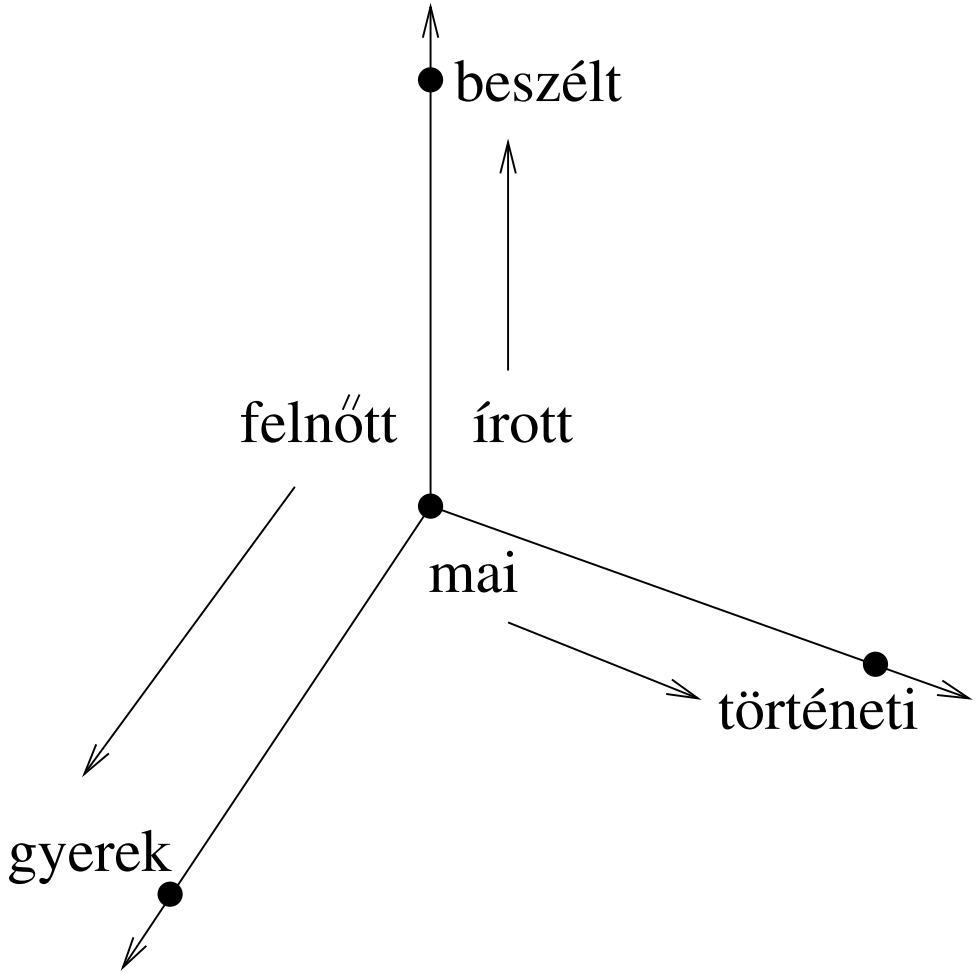
Együtt, egy helyen minél több meglévő…
- magyar nyelvű, online lekérdezhető korpusz
- korpuszlekérdező funkció

http://corpus.nytud.hu/nkp

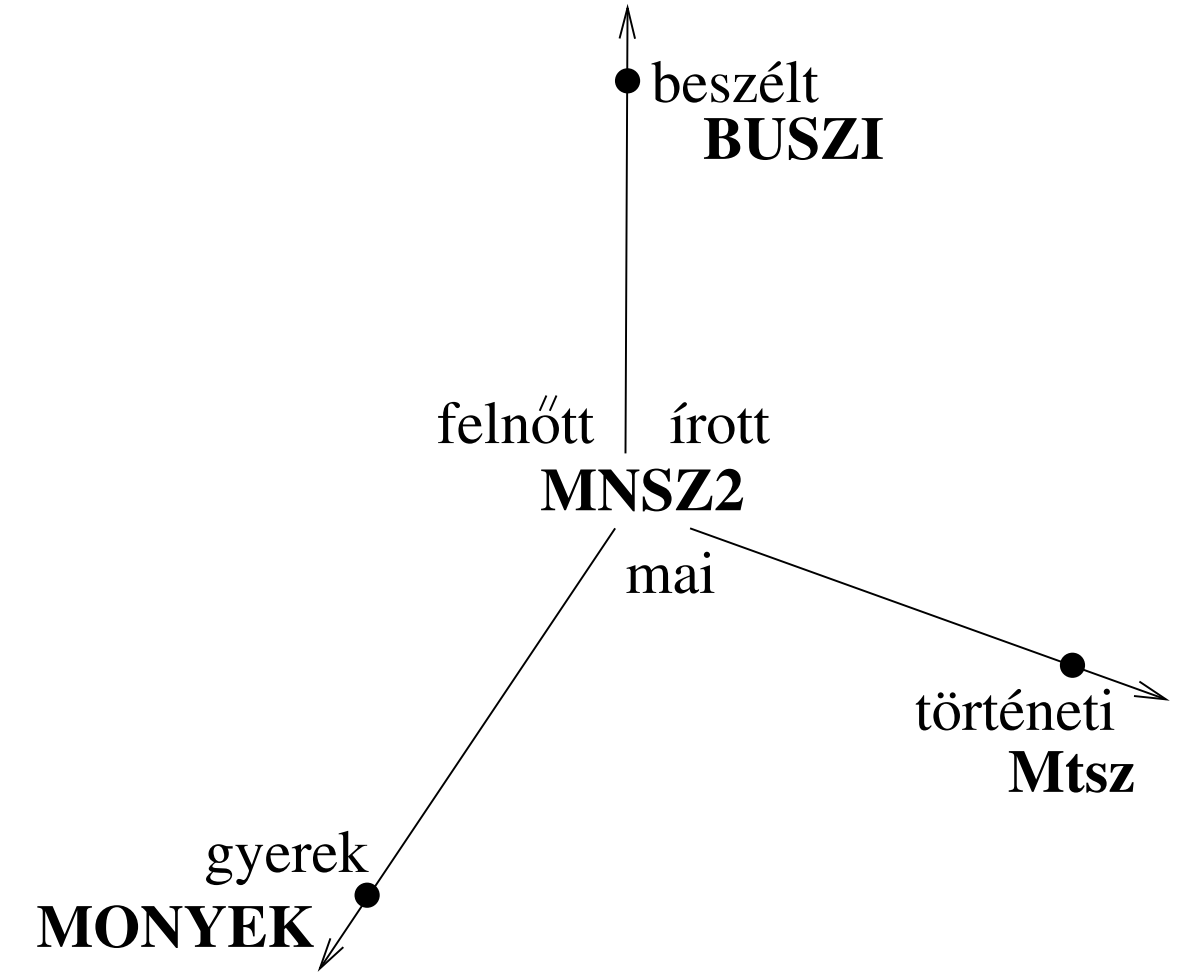
Cél: a korpuszok népszerűsítése a szakma és a nagyközönség felé

Távlati cél: egységesítés, automatizálás

„Alap” korpuszkészlet 4.



„Alap” korpuszkészlet 4.



Nézzük meg a korpuszban! 5.

A korpuszok a nyelvi adatok forrásaként arra szolgálnak,
hogy segítségükkel nyelvészeti kérdésfelvetéseket, hipotéziseket
alátámasztani vagy cáfolni lehessen.

Ha szembetalálkozunk egy nyelvészeti állítással, akkor
ha rendelkezésre áll a megfelelő korpusz,
azonnal ellenőrizhetjük az állítás igazságtartalmát, megfelelőségét.

Kialakítható egy olyan hozzáállás, gondolkodásmód,
hogy amikor felmerül egy ilyen állítás vagy kérdés, akkor
készségszinten, természetes módon nyúljunk a korpuszhoz,
és ott keressünk választ.

Korpuszok együttműködése: cigány eredetű szavak (1/3)

szavak:

csaj, csávó, csór, gádzsó, gizda,
góré, kaja, kéró, lóvé, nyikhaj,
pia, pimasz, séró, verda

→ Melyik a kakukktojás?

Korpuszok együttműködése: cigány eredetű szavak (1/3)

szavak:

csaj, csávó, csór, gádzsó, gizda,
góré, kaja, kéró, lóvé, nyikhaj,
pia, pimasz, séró, verda

→ Melyik a kakukktojás?

első előfordulás az **Mtsz**-*ben:*

csaj – 1963, csávó – 1971, csór – 1913, gádzsó – ∅, gizda – ∅,
góré – 1965, kaja – 1948, kéró – ∅, lóvé – 1968, nyikhaj – 1978,
pia – 1954, pimasz – 1785, séró – 2003, verda – 2004

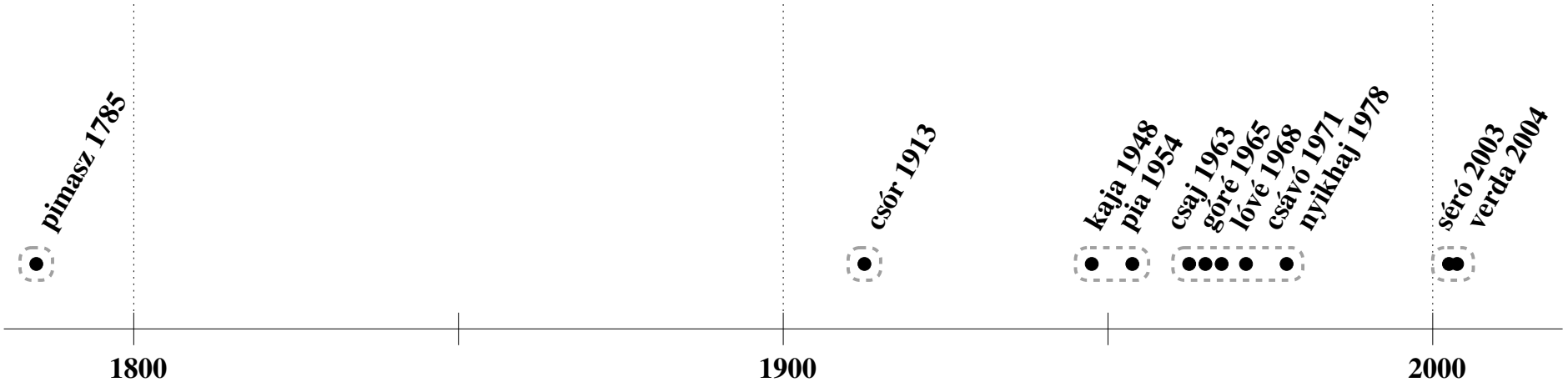
Korpuszok együttműködése: cigány eredetű szavak (1/3)

szavak:

csaj, csávó, csór, gádzsó, gizda,
góré, kaja, kéró, lóvé, nyikhaj,
pia, pimasz, séró, verda

→ Melyik a kakukktojás?

első előfordulás az **Mtsz**-*ben:*



→ a *pimasz* régi magyar szó!:)

Korpuszok együttműködése: cigány eredetű szavak (2/3)

Mennyire köznyelvi?

ötlet: gyakoriság közeli szinonimával összevetve: **MNSZ2**

lány	198000	csaj	10000	$\times$**20**
szemtelen	1976	pimasz	1825	**=**

$\rightarrow$ *pimasz* teljesen köznyelvi
$\rightarrow$ *csaj* kevésbé köznyelvi, van stílusértéke!

Korpuszok együttműködése: cigány eredetű szavak (3/3)

A *csaj* már *nagyon* magyar szó:
van pl. *csajos*, ami ráadásul nagyon $\neq$ *lányos*

MNSZ2/1R gyakorisági lista alapján az eltérés:

- *csajos:* mobil film este buli könyv zenekar program
- *lányos:* ház arcú/képű zavarában apák/anyák/szülők

Ami még kevésbé épült be: gádzsó, gizda, kéró, séró, verda

A korpuszkeresés elvei #2

6. A találatszám-minimalizálás elve.
7. A „mire vagyok kíváncsi” elve.
8. A kontextusalapú keresés elve.

A találatszám-minimalizálás elve 6.

Törekedjünk rá, hogy minél kevesebb találatot kapjunk. :)

Ha `DET + MN.NOM`-ra keresünk,
akkor először `MN.NOM`,
aztán szűrés: `DET`.

A „mire vagyok kíváncsi” elve 7.

mire vagyok kíváncsi = miből szeretnék gyakorisági listát készíteni

4. példa volt: Készítsünk gyakorisági listát
a *fog*-tól jobbra 1, 2 vagy 3 szó távolságban lévő FNI-kből.

megoldás volt: 1. FNI + 2. *fog*

A „mire vagyok kíváncsi” elve 7.

Többszavas lekérdezés vagy szűrés?

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.

A kontextusalapú keresés elve 8.

alap: a keresett szó (formai/elemzési) tulajdonságai alapján keresünk

ha ez nem megy: megpróbálhatunk a *kontextus* (formai/elemzési) tulajdonságai alapján keresni!

– **3. példa** volt: alanyesetű melléknév?
  ezt alkalmaztuk: főnév *előtti pozícióban* fogjuk megtalálni
– keressünk ilyet: *villany lekapcsol, ajtó becsuk, ...*
  tipp: ez felsorolásként jelenik meg
  („halmozódás elve”)
– keressünk pozitív értékelőket
  tipp: (*jó* utáni szavak – *rossz* utáni szavak) előtt mi van!
  => különleges, szülőkímélő, kihagyhatatlan
  („áttételes keresés elve”)

6.
Példák az MNSZ2 logból „korpuszlekérdezés-korpusz”

Példák az MNSZ2 logból

1. érdekes/értelmes lekérdezések
2. hibák
3. „ilyet ne”

a) `"tudatjuk" "mindazokkal"`

b) `[lemma="felkap"] [lemma="a"] [lemma="víz"]`

c) `[word="."]`

d) `[lemma = "k[K]andeláber"]`

e) `ama i*`

Példák az MNSZ2 logból

1. érdekes/értelmes lekérdezések
2. hibák
3. „ilyet ne”

f) `[word="\."] [word="[Mm]indig"] [word="\."]`

g) `[msd="Det.*"] [msd="FN.PSe2.*"]`
`[lemma="fog"] [word=".*ni" & pos="V.*"]`

h) `".*"`

i) `[][][][]*`

j) `[word = "elé"] [word = "a.?"] [word = ".+n(a|e)k"]`

7.
Feladatok

Feladatok

1. a melléknevek középfoka „mindig alsó nyelvállású kötőhangzóval jár: *-abb/-ebb*, ennek csak az amúgy is kivételes, mert nem nyitó *nagy* melléknév áll ellen: *nagyobb*.” (nyest.hu) → Ellenőrizzük!
2. Ikes feltételes ragozás (*aludnám, aludnék, aludna*) diakrón változása
3. *farmerben/farmerban* típusú szavak keresése
4. Mióta van meg a *köszönhetően* alak?
5. Fosztóképzős (*talan* ill. *tlan* morfémát tartalmazó) alakok?
6. Van-e az ómagyarban egyenes szórendű tagadás, azaz a mai *nem futott ki* helyett *ki nem futott*?
7. Keressünk olyan ómagyar nyelvi adatot, ahol nincs ott a névelő, pedig várnánk.

Feladatok

8. Mik a *munka* tipikus jelzői?

9. *kiküszöböli a csorbát* – Fura, nem?

10. Igekötős ige összes alakjának keresése az MNSZ2-ben

11. Hogy viszonyul egymáshoz az *össze* és a *-vAl*?
    összefügg, *összeköt* vs. *összehív*, *összeszed*

12. Mennyire jó a *szomszéd* fn/mn annotációja az MNSZ2-ben?

13. `prev` segítségével: adott/összes igekötő mennyire szeret elválni?

14. *a kutya és a macska és az egér…* típusú szerkezetek keresése

Összefoglalás

- az elemzett korpusz = egy **táblázat**
- NoSkE korpuszkezelő
- **szűrés 1..1** ***és*** **gyakorisági lista 1R**
- **regex + CQL =** `".+t" ".+tt"`
- Mtsz, MNSZ2, Ómagyar, BUSZI, Mazsola
- „Példák korpuszból”
- „Minden találat kell!”
- „Ne bízzunk vakon az annotációban!”
- **„Nézzük meg a korpuszban!” :** ***pimasz***
- „Kontextus alapú keresés”

Sass Bálint – `sass.balint@nytud.hu`