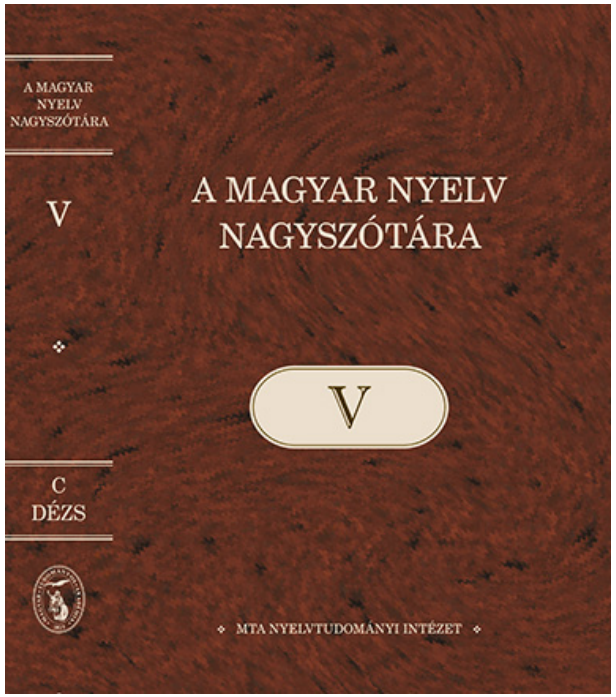
A kibővített *Magyar történeti szövegtár* új keresőfelülete

Sass Bálint
MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

A nyelvtörténeti kutatások újabb eredményei IX.
2016. április 27., Szeged

Nszt és Mtsz



`http://nszt.nytud.hu/nszt.html`

Magyar történeti szövegtár

- 30 millió szövegszó
- 1772-2010-ig $\sim$ 240 év
- 2015-ben készült el a bővítés

1.
Az Mtsz használata

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
– *funkciók:* alkorpuszok, mentés, szűrés, gyaklisták, kollokációk
– CQL = Corpus Query Language – formális lekérdezőnyelv

elérhetőség: `http://clara.nytud.hu/mtsz`

CQL – reguláris kifejezések

Bizonyos tulajdonságú karaktersorozatok megadására.

.	tetszőleges karakter
*	a megelőző karakterből bármennyi (0 vagy több)
[ab]	’a’ vagy ’b’ karakter

Példák:

(1) tejf.l

(2) .*

(3) .*bb

(4) **nélk[üűúű]l**

CQL

Corpus Query Language

[..]	egy tokenre vonatkozó megkötések
x="y"	*x* attribútum értéke legyen *y* – Mtsz: csak *word* attribútum van
x!="y"	*x* attribútum értéke *ne* legyen *y*
&	*és* kapcsolat megkötések között

Példák:

1. [] []
2. [word="majd"]
3. "majd"
4. [word!="a.*"]

Mtsz példalekérdezés

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

Mtsz példalekérdezés

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

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

Mtsz példalekérdezés

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

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

most itt – ???

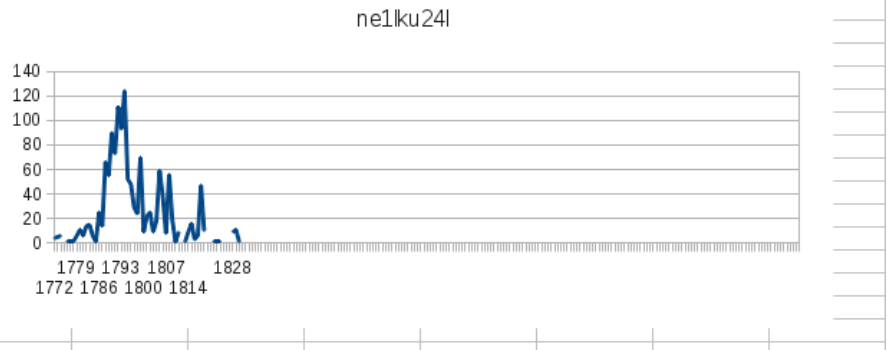
Mtsz példalekérdezés – a vizsgálat lépései

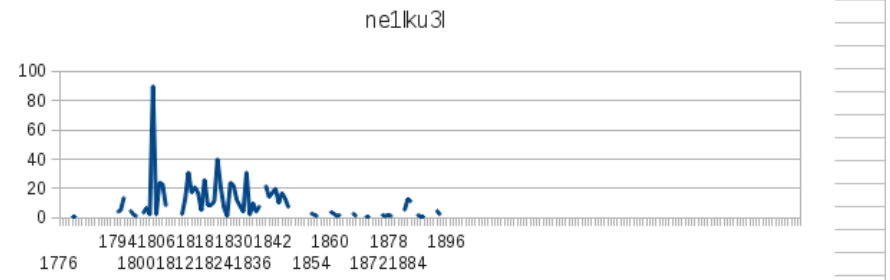
1. CQL: `".*t" ".*tt"`
2. Gyakoriságok / szóalakok
3. $p \rightarrow$ erőt vett
4. Milyen szó jön utána? $\rightarrow$ Gyakoriságok: 1R
5. $p \rightarrow$ rajta
6. Rendezés / jobb: hogy *mi* vesz erőt rajta

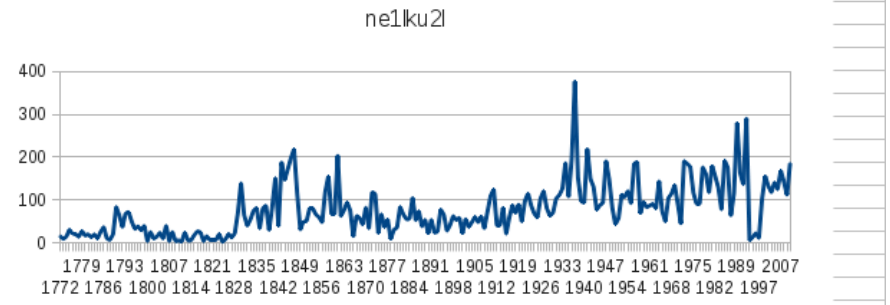
$\rightarrow$ félelem, féltékenység, habozás, kacagás, kishitűség, kíváncsiság …

Diakrón vizsgálat: *nélkül* helyesírása

	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.
Az Mtsz építése

Az Mtsz építése

Mi mindent kell csinálni ahhoz,
hogy sima szövegből ilyen korpuszlekérdezőfelület legyen?

1. Unicode (UTF-8) karakterkódolás é ✓ ő ✓ f ✓ ő ✓
2. XML – saját / TEI

korpuszkezelő rendszer: NoSketchEngine (NoSkE)
`https://nlp.fi.muni.cz/trac/noske`

3. NoSkE XML ← XSLT

NoSkE XML

```
<doc mtsz_id="7021030" author="Baróti Szabó Dávid" wdate="1808" ...>
  <oldal oldalszam="18">
    <par>
      Ám
      de
      viſzont
      hallá
      <g/>
      ,
      hogy
      ...
      <br/>
      ...  további tokenek
    </par>
    ...  további bekezdések
  </oldal>
  ...  további oldalak
</doc>
```

NoSkE XML

```
<doc mtsz_id="7021030" author="Baróti Szabó Dávid" wdate="1808" ...>
  <oldal oldalszam="18">
    <par>
      Ám
      de
      viſzont
      hallá
      <g/>
      ,
      hogy
      ...
      <br/>
      ...  további tokenek
    </par>
    ...  további bekezdések
  </oldal>
  ...  további oldalak
</doc>
```

Az Mtsz építése

Mi mindent kell csinálni ahhoz,
hogy sima szövegből ilyen korpuszlekérdezőfelület legyen?

1. Unicode (UTF-8) karakterkódolás é ✓ ő ✓ ſ ✓ ő̈ ✓
2. XML – saját / TEI

korpuszkezelő rendszer: NoSketchEngine (NoSkE)
`https://nlp.fi.muni.cz/trac/noske`

3. NoSkE XML ← XSLT
4. tokenizálás (…és egyéb elemző lépések)
5. „indexálás” = a korpusz betöltése a korpuszkezelőbe

Nemzeti Korpuszportál (NKP)

Együtt, egy helyen minden 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: automatizálás

A kibővített *Magyar történeti szövegtár* új keresőfelülete

Sass Bálint
MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

A nyelvtörténeti kutatások újabb eredményei IX.
2016. április 27., Szeged