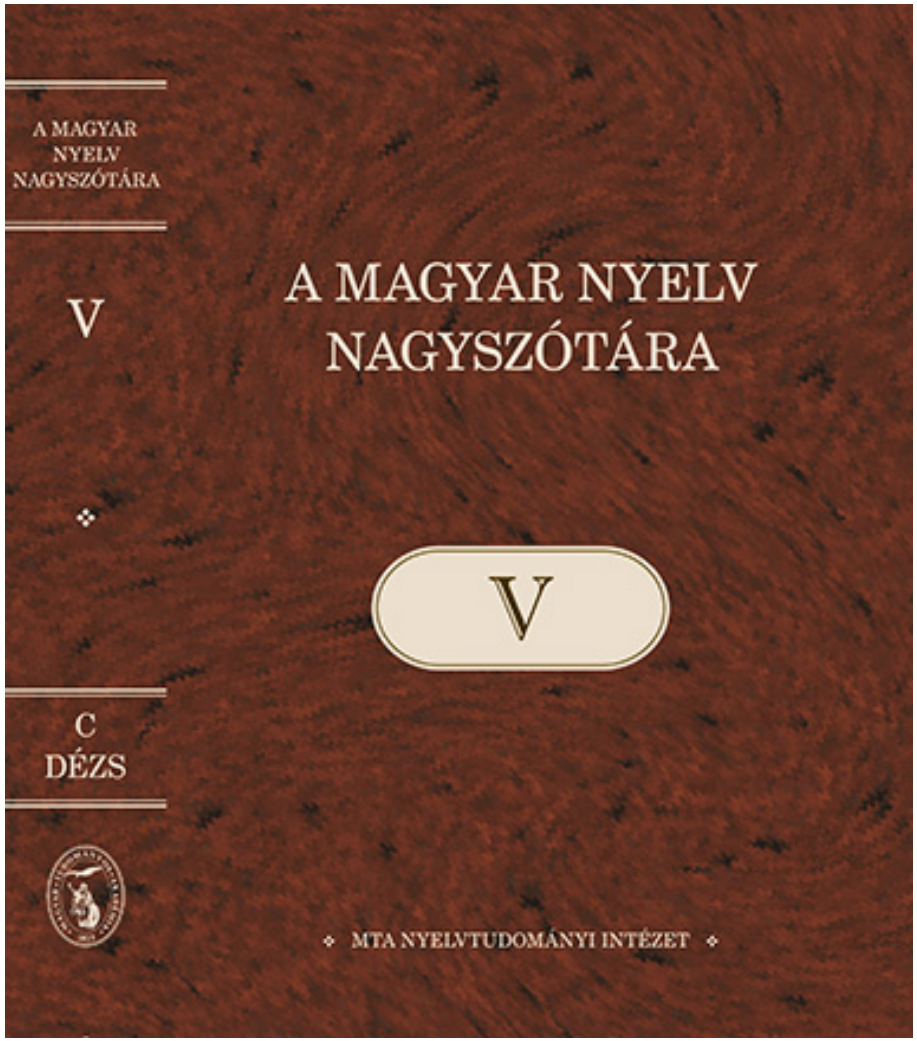
Korpusznyelvészet

2016. április 18., ELTE

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



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

Mi mindent kell csinálni ahhoz,
hogy sima szövegből ilyen korpuszlekérdezőfelület legyen?
Ezt fogjuk most megnézni lépésről lépésre.

Példa:

/ Ám de viſzont hallá, hogy majd a’ Trójai vérből /

1.
Az Mtsz építése

Karakterkódolás

1 karakter = 1 byte

kitûnõ és idõszerû

1 karakter > 1 byte: Unicode

helytakarékos kódolás: UTF-8

HÃ©tfÅ n talÃ¡lkozunk.

Megvan az egységes UTF-8 kódolású szövegünk.

é ✓ ő ✓ f ✓ ő ✓

XML

Kenjük a vajat

a késsel

a kenyérre.

XML – tagek

<recept>

Kenjük a **<hozzavalo>**vajat**</hozzavalo>**

a **<eszkoz>**késsel**</eszkoz>**

a **<hozzavalo>**kenyérre**</hozzavalo>.**

</recept>

XML – attribútumok

```
<recept nev="vajas kenyér">

Kenjük a <hozzavalo id="41">vajat</hozzavalo>

a <eszkoz id="5">késsel</eszkoz>

a <hozzavalo id="12">kenyérre</hozzavalo>.

</recept>
```

Mtsz XML

```
<section>
  <head>
    <id>7021030</id>
    <author>Baróti Szabó Dávid</author>
    <wdate>1808</wdate>
    ...  egyéb adatok
  </head>
  <text>
    <page>
      <p>18</p>
      <par>
        Ám de viſzont hallá, hogy majd a' Trójai vérből<br/>
        Nemzet ered, melly e' várat valahára le-dönti;<br/>
        ...  további sorok
      </par>
      ...  további bekezdések
    </page>
    ...  további oldalak
  </text>
</section>
```

TEI

„The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.”

Korpuszkezelő

NoSketchEngine (NoSkE)

`https://nlp.fi.muni.cz/trac/noske`

NoSkE XML

```
<doc mtsz_id="7021030" author="Baróti Szabó Dávid" wdate="1808" ...>
  <oldal oldalszam="18">
    <par>
      Ám de viſzont hallá, hogy majd a’ Trójai vérből<br/>
      Nemzet ered, melly e’ várat valahára le-dönti;<br/>
      ...  további sorok
    </par>
    ...  további bekezdések
  </oldal>
  ...  további oldalak
</doc>
```

NoSkE XML ← XSLT ← Mtsz XML

```
<doc mtsz_id="7021030" author="Baróti Szabó Dávid" wdate="1808" ...>
  <oldal oldalszam="18">
    <par>
      Ám de viſzont hallá, hogy majd a’ Trójai vérből<br/>
      Nemzet ered, melly e’ várat valahára le-dönti;<br/>
      ...  további sorok
    </par>
    ...  további bekezdések
  </oldal>
  ...  további oldalak
</doc>
```

Találatok időrendben

Hogy lehet a találatok idő szerinti rendezését megvalósítani?

1. Lekérdezés után az eredményt mindig rendezzük.
2. Előre rendezzük és azt kérdezzük le.

Megoldás (erre is): XSLT

„Tokenizálás”

```
<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>
```

„XML+TAB” formátum

```
<doc mtsz_id="7021030" author="Baróti Szabó Dávid" wdate="1808" ...>
  <oldal oldalszam="18">
    <par>
      Ám        ám       KOT
      de        de       KOT
      viſzont   viſzont  KOT
      hallá     hall     V.Ipf.S3.Def
      <g/>
      ,         ,        WPUNCT
      hogy      hogy     KOT
      ...
      <br/>
      ...  további tokenek
    </par>
    ...  további bekezdések
  </oldal>
  ...  további oldalak
</doc>
```

2.
Az Mtsz használata

Reguláris kifejezések

Bizonyos tulajdonságú karaktersorozatok megadására.

.	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

Példák:

1. alma
2. tejf.l
3. mentők?
4. **nélk[üűú]l**
5. .*
6. .*bb
7. alma|almá.*
8. mondjá(to)?k

CQL (Corpus Query Language)

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

Példák:

1. [] []
2. [word="majd"]
3. "majd"
4. [word!="a.*"]
5. []?
6. [word="nem"] [word="kellett"] [word="volna"]? [word=".*ni"]

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

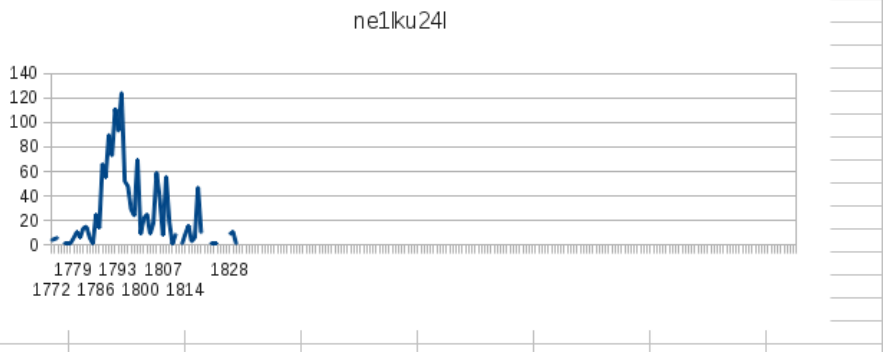
Feladat. Keressünk ilyet: *tárgyesetű szó + múltidejű E/3 ige!*

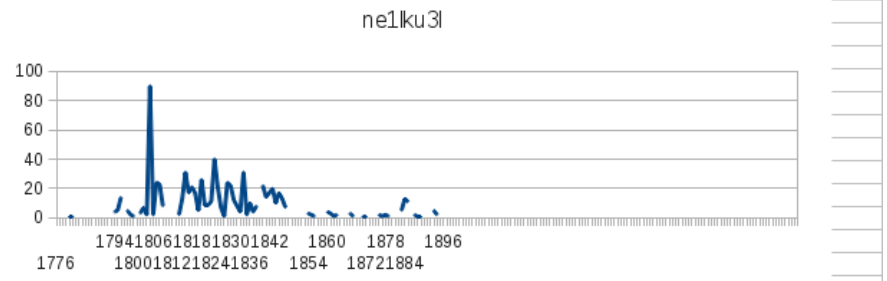
```
".*t" ".*..tt"
```

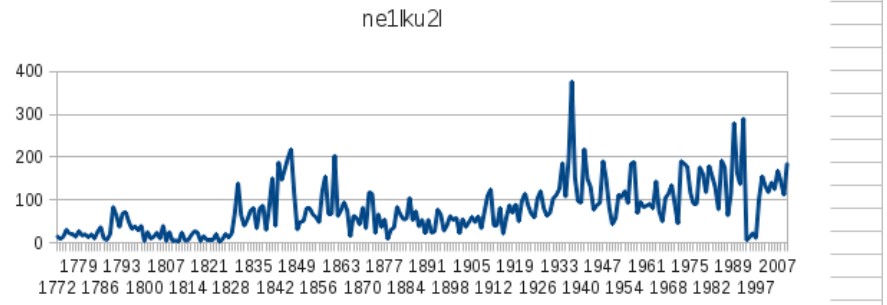
```
[word=".*t"] [word=".*..tt" & word!="alatt" & word!="azelőtt"]
```

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







3.
Korpuszok

Korpuszok

MNSZ2 – elemzett, nagy méretű

* körülültük, felszedegettük, elsimítottuk, végigcsináltuk

* forrásokban, fellegekben, falvakban, fejekben (bazi lassú)

* cél, csal, csaj, csel, dzsal

Mazsola – igék és bővítmények

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

BUSZI – beszélt de írott

... bizonyos dógokban □ mmm tát, hogy öööö lustább annál, mint amilyennek elképzeltem, ...

Ómagyar korpusz – normalizálás, ómagyar morfológia

NKP (Nemzeti Korpuszportál)

`http://corpus.nytud.hu/nkp`

4.
Korpuszvezérelt kutatás

Korpuszalapú és korpuszvezérelt

„A korpusz segédeszköz, ami empirikus adataival támogatja az intuíciót, mérhetővé teszi a nyelvi jelenségeket, meglévő elméleteket bizonyít/cáfol.”

„A korpusz maga szolgáltatja az «elméletet», a nyelvész minden feltevés és elvárás nélkül fordul az adatokhoz. Minden következtetést kizárólag korpuszmegfigyelésekből von le.”

serendipity principle:
lényeges jelenség véletlen felfedezése

pl.: Tognini-Bonelli: Corpus Linguistics at Work (2001)

Braille-rövidírás bővítése

ban/ben → b (⠃ → ⠃) hoG → h (⠓ → ⠓)

Alapötlet: a maximális rövidítési képességgel bíró ideális rövidírás a magyar nyelv korpuszgyakorisági adatai alapján **korpuszvezérelt** módon, automatikusan kialakítható.

Elv: a lehető leggyakoribb *betűkapcsolatokat* kell a lehető legrövidebbre rövidíteni.

Eredmény: 33 új szabály

meg → mg (⠍ ⠑ ⠛ → ⠍ ⠛) maGar → mG (⠍ ⠁ ⠛ ⠁ ⠗ → ⠍ ⠛)

Tanulság:
A korpuszvezérelt módon létrehozott rendszer még úgy is kétszeres teljesítményre képes az intuíció illetve hagyomány talaján álló rendszerrel szemben, hogy már eleve jelentősen rövidített szövegen kell dolgoznia.

Ha valamit meglévő (korpuszgyakorisági) adatokból automatikusan származtatni tudunk, akkor nem érdemes intuitív megközelítést alkalmazni.

Igei szerkezetek felfedezése

vet [15728]

vet -nAk VÉG-t [1463] *vessen véget az erőszaknak*
vet SZEM-A-rA -t [805] *hasonló diszkriminációkat vetnek az albán hatóságok szemére*
vet -rA PILLANTÁS-t [708] *vess egy pillantást a térképre*
vet -t [703] *vetem a magot*
vet -rA -t [380] *a humanista könyveket máglyára vetették*
vet PAPÍR-rA -t [377] *vesse papírra az új problémákat*
vet SZÁM-t -vAl [297] *ez rossz fényt vet az edzők nevelő munkájára*
vet -rA FÉNY-t [267] *vessünk számot eddigi politikánkkal*
vet -bA -t [252] *a tó vizébe vetette magát*

csóvál [1078]

csóvál FEJ-A-t [754] *csóválta a fejét*

Korpusznyelvészet

2016. április 18., ELTE

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