Korpuszlekérdezés

Nyelv és informatika – Pécs, 2022/23 tavasz
4. óra

Simon Eszter – Vadász Noémi

2023. február 18.

Tartalom

Korpuszlekérdezés

Terminológia

- nyers (*raw*) korpusz: szóalakok
- annotált korpusz: valamiféle annotáció kapcsolódik a tokenekhez vagy tokenszekvenciákhoz
- konkordancia: a lekrdezés eredménye
- találat (*hit*): kwic + kontextus
- kwic: *keyword in context*: a lekérdezett kifejezés
- ablak (*window*): a tokensorozat egy szakasza, amit kezdő- és végpozícióval definiálunk a kwic-hez képest (pl. -1...2: 4 szó, a kwic, egy szó balra és két szó jobba tőle)
- lekérdezőrendszer (*corpus query system, CQS*): a rendszer, amin keresztül a júzer konkordanciákat nyerhet ki lekérdezésekkel (pl. NoSketchEngine (NoSke), Emdros)
- formális korpuszlekérdező nyelv (*formal query language*) (pl. CQL, MQL)

Reguláris kifejezések

joker-karakter: `.`

mennyiségjelzők: `? + * { }`

választás: `| []`

csoportosítás: `( )`

Példák:

- `a|b*`
- `gr(a|e)y, gr[ae]y`
- `b[aeiou]bble`
- `colou?r`
- `go+gle`
- `z{3}, z{3,6}, z{3,}`

CQL

```
[attribútum="regkif"]
```

Néhány hazai korpusz lekérdezőfelülettel

Magyar Nemzeti Szövegtár (MNSZ)

- 1998-ban kezdték építeni
- összetétele: http://clara.nytud.hu/mnsz2-dev/stat.html
- MNSZ2 1,04 milliárd szövegszó
- hat stílusréteg, öt regionális nyelvváltozat
- 76 millió szavas beszéltnyelvi (rádiós) alkorpusz felolvasott szöveges tartalommal és spontán beszéddel
- személyes alkorpusz: fórum (57,9 millió szó), közösségi (243,2 millió szó)

MNSZ2

- 1,04 milliárd szövegszó (1,348 milliárd token)
- használatához hozzáférést kell igényelni
- morfoszintaktikai kódok:
  http://clara.nytud.hu/mnsz2-dev/msd.html

https://www.youtube.com/@magyarnemzetiszovegtar

Mazsola

- igék és argumentumaik közvetlen vizsgálata
- kollokációk a vonzatkeretek feltárásához
- 2006-2009 között fejlesztették az MNSZ anyagát felhasználva
- gyakorisági vonzatkeret-szótár alapjául szolgált (Magyar igei szerkezetek: http://www.tintakiado.hu/book_view.php?id=286)
- anyanyelvi nevelés, magyar mint idegen nyelv
- lexikográfia
- nyelvészeti kutatások (gyakoriság, szemantikai osztályozás bővítménykeret alapján, szemantikai szelekció stb.)
- http://corpus.nytud.hu/mazsola/s/mazsola_hun.html

Magyar Történeti Szövegtár

- jelenleg 30 millió szövegszót tartalmaz
- eredetileg: 1772 és 2000 között keletkezett különböző műfajú és stílusú szövegek gyűjteménye
- 2015-ben 3 millió szövegszónyi 2001 és 2010 közötti szemelvénnyel egészült ki, megtartva a regiszterek arányait
- részletes bibliográfiai adatokat is megjelenít az egyes találatok mellett
- gyakorisági listák, a lekérdezések szűrése, kollokációk keresése
- a Unicode szerinti karakterkódolás miatt a 18. században még gyakori régi grafémák is eredeti formájukban jelennek meg
- http://clara.nytud.hu/mtsz/

Ómagyar Korpusz (MGTSZ)

- 2009-2013, 2015-2018
- annotált korpusz, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) és néhány középmagyar kori (1526–1772) szövegemléket, és amely nyelvészetileg releváns kérdésekre tud választ szolgáltatni
- 3,2 millió szövegszó: 47 ómagyar kódex, 24 rövidebb ómagyar szövegemlék, 244 misszilis és 5 középmagyar kori bibliafordítás
- http://omagyarkorpusz.nytud.hu/hu-descr.html

Párhuzamos Bibliaolvasó

- Párhuzamos Bibliakorpusz:
  - bibliafordítások a magyar nyelv különböző korszakaiból
  - egyéb uráli nyelvek bibliafordításai
  - King James Bible az angol nyelvű glosszázáshoz
- https://parallelbible.nytud.hu/

Történeti magánéleti korpusz (TMK)

- az ó- és középmagyar kor magánéleti nyelvi regiszteréhez legközelebb álló műfajokat tartalmazza
- 1772 előtti magánlevelek és peres eljárások jegyzőkönyvei
- történeti morfológiai és szociolingvisztikai, történeti mondattani, pragmatikai és lexikológiai kutatásokhoz
- 8.6 millió karakter (magyar nyelvű rész: 7,68 millió karakter, 1 millió 112 ezer elemzett szövegszó) (2020 szeptemberi adat)
- https://tmk.nytud.hu/3/
- Emdros, MQL
- útmutató a kereséshez: https://tmk.nytud.hu/utmutato.php

Csángó korpusz

- Tánczos Vilmos kolozsvári néprajzkutató moldvai gyűjtése
- dialektológiai kutatásokra alkalmas
- hangfelvételek
- kutatópontok térképen
- magyar egyezményes hangjelölési rendszer
- folyamatban: normalizálás, morfológiai elemzés, egyértelműsítés
- https://nlp.nytud.hu/csango/index.html

Köszönjük a figyelmet!

https://github.com/esztersimon/nlp_at_pecs

simon.eszterke@gmail.com
vadasz.noemi@nytud.hu