



# Bevezetés a korpuszok és nyelvészeti adatbázisok világába

Vincze Veronika

[vinczev@inf.u-szeged.hu](mailto:vinczev@inf.u-szeged.hu)



# Az előadás vázlatja

- **Bevezetés, alapfogalmak**
- **Korpusztípusok**
- **Annotáció**
- **Korpuszok**
- **Adatgyűjtés korpuszból**
- **Az adatok felhasználása**



# Bevezetés

- **Nyelvészeti kutatósmódszertan**
- **Adatorientált / elméletorientált módszerek**
- **Kompetencia / performancia**
- **Honnan származnak a nyelvi adatok?**

# Adatgyűjtés

- Intuíció alapján
- Adatközlőktől gyűjtött adatok
- Korpuszok
  
- A mai nap fókuszában: korpuszból történő adatgyűjtés





# Alapfogalmak

- **Korpusz: speciális célokra létrehozott, (gyakran tematikus) adatbázis – „szöveggyűjtemény”**
- **Annotáció: a szövegek nyelvi információval történő jelölése (és/vagy kézi ellenőrzése)**
- **Korpusznyelvészet: korpuszban található nyelvi adatok elemzése**



# Korpusztípusok

- **Hogyan gyűjtjük és csoportosítjuk a szövegeket?**
- **Nyelv szerint:**
  - Egynyelvű (pl. magyar nyelv korpuszai)
  - Többnyelvű – speciális esete a párhuzamos korpusz: ugyanazok az adatok egynél több nyelven
- **Modalitás szerint:**
  - Beszédkorpusz: hanganyagok (+ átiratok)
  - Írott nyelvi korpusz: szövegek
  - Multimodális: hang + kép + szöveg



# Korpusztípusok

- **Beszélők/szerzők szerint (pl.):**
  - Anyanyelvi beszélők
  - Adott tájegység beszélői
  - Gyermeknyelvi korpuszok
  - Nyelvtanulói korpuszok
  - Író/költő összes művei
- **Stílus / regiszter szerint:**
  - Jogi szaknyelvi korpusz
  - Orvosi szaknyelvi korpusz
- **Szövegek keletkezési ideje szerint:**
  - Pl. ómagyar nyelvemlékek korpusza



# Általános célú korpuszok

- **Általában több szövegtípus, stílusréteg, műfaj, beszélő stb. (heterogén adatok)**
- **Az adott nyelv vagy nyelvi réteg minél teljesebb reprezentációja a cél**
- **Sokszor nagyobb méretűek**





# Speciális korpuszok

- Egy adott nyelvi réteg, stílus, időszak stb. jellemző szövegei (homogén adatok)
- Lehet akár teljes körű is vagy erre törekvő (pl. nyelvemlékek korpuszai)
- Általában kisebbek, mint az általános célú korpuszok

# Annotáció

- **Nyelvi információ jelölése a szövegekben**
- **Automatikus annotáció: gépi algoritmus jelöli be**
- **Félig automatikus annotáció: gép jelöli, ember javítja**
- **Kézi annotáció: emberi szakértő jelöli**





# Annotáció típusai

- **Bármit annotálhatunk, ami számunkra érdekes nyelvészeti információ**
- **Alapszintű annotációk:**
  - Mondat- és szóhatárok
  - Szófajok és morfológiai jegyek az egyes szavakhoz
  - Szintaxis
- **Magasabb szintű annotációk (pl. szemantikai, pragmatikai jelenségek)**

# Magyar nyelvű korpuszok

- **Általános célú korpuszok**
  - Magyar Nemzeti Szövegtár (automatikus szótövezés és szófaji elemzés)
  - Webkorpusz (elemzetlen)
  - Szeged Korpusz (kézzel ellenőrzött morfológiai és szintaktikai (konstituens és függőségi) elemzés, névelemek, többszavas kifejezések, koreferencia)





# Magyar nyelvű korpuszok

- **Speciális korpuszok (lásd későbbi előadások)**
  - **Gyermeknyelvi korpuszok**
  - **Történeti korpuszok**
  - **Nyelvtanulói korpuszok**
  - **Tulajdonnévkorpuszok**
  - **Többszavas kifejezések korpuszai**
  - **Jelentés-egyértelműsített korpusz**
  - **Nyelvi bizonytalanságra annotált korpuszok**
  - **Stb.**



# A korpuszok felhasználhatósága

- Referenciaadatok
- (Gépi tanuló) algoritmusok tanítása
- Algoritmusok tesztelése
- Nyelvészeti adatok gyűjtése



# Milyen korpuszt használjunk?

- **Kutatási témához illeszkedő korpusz elérhető-e már?**
- **Ha igen, akkor a meglevő korpuszból érdemes adatokat gyűjteni (I. „Korpuszban keresés” előadás)**
- **Ha nem, akkor érdemes külön korpuszt építeni**



# Saját korpusz építése

- **Szövegek gyűjtése**
  - Tematika (jog, irodalom...)
  - Nyelvi regiszterek (hivatalos, köznyelv, internetes nyelvhasználat...)
  - Homogén/heterogén
  - Milyen egyéb (meta)adatok? (idő, szerző...)
  - Méret
  - Nyelv
  - Hozzáférhetőség (szerzői jogok, anonimizálás)
- **Annotáció**



# Automatikus annotáció

- Alapszintű elemzésre számos (online) eszköz elérhető a magyar nyelvre is:
  - e-magyar (l. később)
  - magyarulanc
  - Universal dependencies
- Szöveg (txt) formátum



# magyarlanc

- <http://www.inf.u-szeged.hu/rgai/magyarlanc>
- Mondatra és szavakra bontás
- Morfológiai elemzés
- Kétféle szintaktikai elemzés
- Online demo (1 mondatra):
- <http://rgai.inf.u-szeged.hu/magyarlanc-service/>
- Szövegfájlok elemzése is lehetséges parancssorból a letölthető program segítségével





# Universal dependencies

- Nemzetközileg egységes morfológiai és szintaktikai annotációs séma kb. 50 nyelvre
- UDPipe:  
<http://lindat.mff.cuni.cz/services/udpipe/>
- Mondatra és szavakra bontás
- Morfológiai elemzés
- Szintaktikai elemzés
- 1-1 mondat és szövegfájl elemzése egyaránt lehetséges online

# Adatgyűjtés a magyarulanc és a UDPipe kimenetéből

- Ha nincs programozói ismeretünk...
- Excel-szűrésekkel is tudunk egyszerű statisztikai adatokat gyűjteni, pl.:
  - Adott szófajok aránya
  - Leggyakoribb főnevek, amik alanyként szerepelnek a szövegben
  - Tulajdonnevek a szövegben
- TANIT: magyarulancra épülő online szolgáltatás, alapvető statisztikai adatokat automatikusan összegyűjti (l. későbbi előadás)





# Mire lehet használni?

- **Saját nyelvészeti kutatás**
- **Gépi tanuló algoritmusok fejlesztése (pl. Szeged Korpusz annotációin tanult a magyarulanc)**

Ember

- Kézzel annotált adatok
- Korpuszok

Gép

- Gépi tanulórendszerek
- Statisztikai módszerek

Szöveg

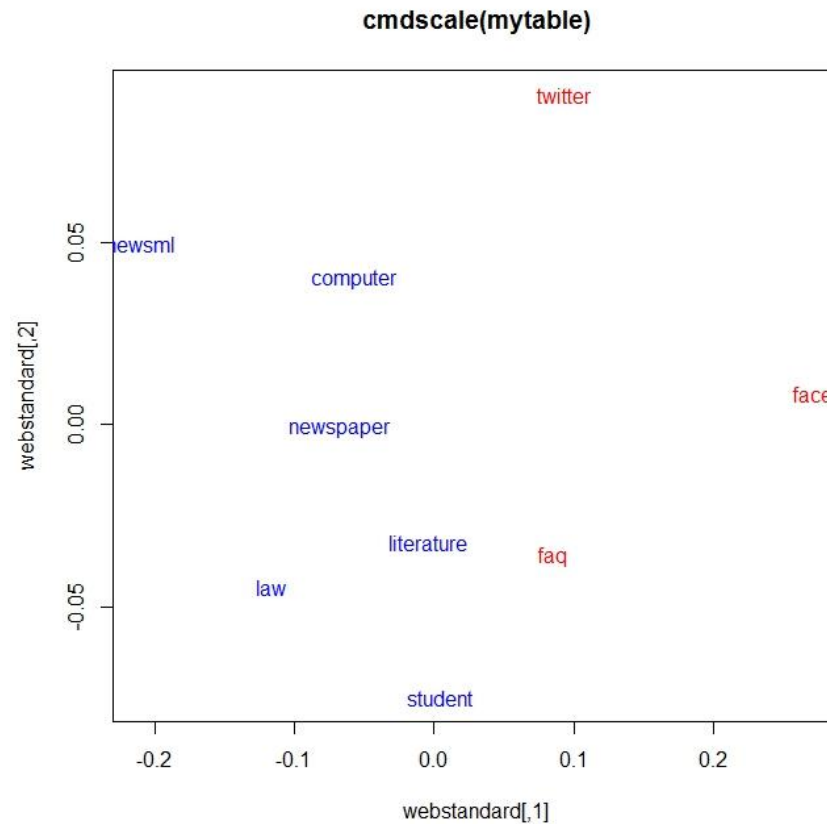
- Új szövegek automatikus feldolgozása





# Mire lehet használni?

- Szövegtípusok összevetése, hasonlósági metrikák





# Próbáljuk ki!

- Tetszőleges cikk a webről
- <https://boilerpipe-web.appspot.com/>
- A kinyert szöveget ide másoljuk a megfelelő beállításokkal
- <http://lindat.mff.cuni.cz/services/udpipeline/>
- Az eredményt elmentjük, majd Excelbe illesztjük
- Szűrések (pl. tulajdonnevek, főnevek...)
- Szófelhő készítése
- <http://www.wordle.net/>





# Összegzés

- **Korpusznyelvészeti bevezető**
- **Néhány korpusz bemutatása**
- **Saját korpusz létrehozása**
- **Adatok gyűjtése és felhasználása**