Seminar on Speech and Language Technology Tools

Szeged, 19 October 2018

Bevezetés az e-magyar programcsomag használatába

Mittelholcz Iván
MTA Nyelvtudományi Intézet

Kik csinálták?

- MTA Nyelvtudományi Intézet
- MTA SZTAKI
- Szegedi Tudományegyetem
- Pázmány Péter Katolikus Egyetem
- Morphologic Kft.
- AITIA International Zrt.

Miért csinálták?

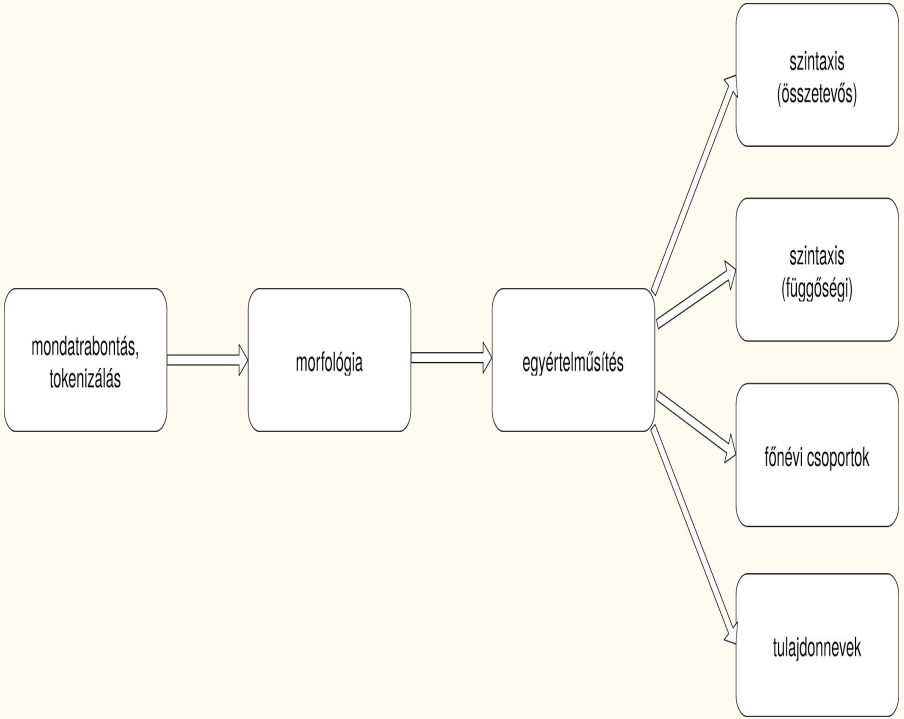
- különféle szövegelemzési feladatokhoz
- egy rendszerbe összeszedni a különböző műhelyek programjait
- háttér: modulok egymásra utaltsága
- nagyközönség számára elemzési lehetőséget biztosítani

Kiknek készült?

- érdeklődő laikusoknak: e-magyar.hu weboldal
- középhaladó szint: GATE integráció (gate.ac.uk)
- nyelvtechnológusoknak, fejlesztőknek

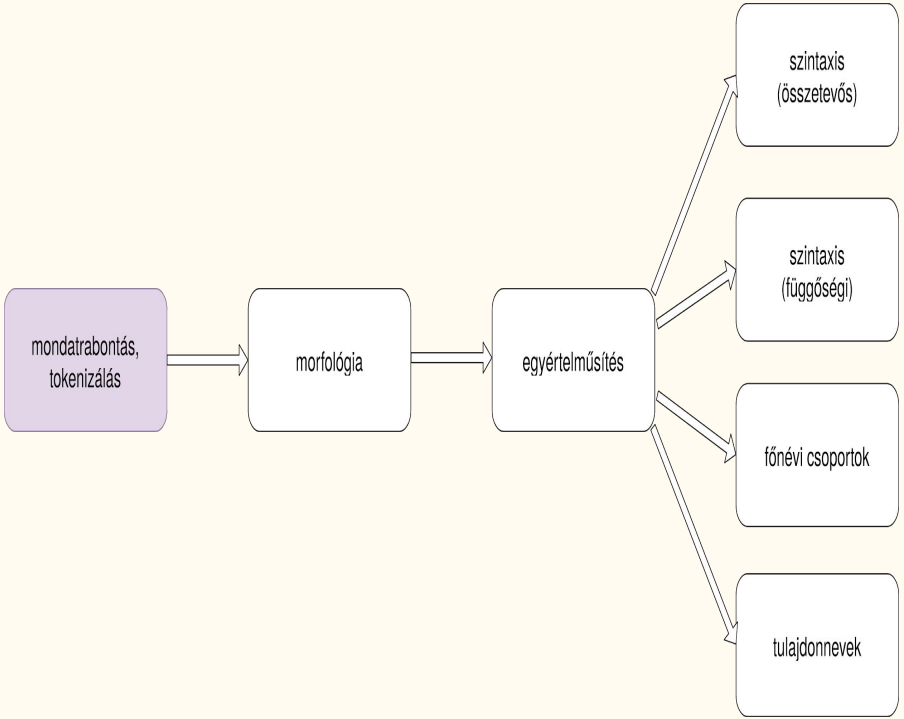
Hogyan működik?

- csővezeték architektúra (*pipeline*)
- az első modul bemenete sima szöveg
  - plain text, utf-8 karakterkódolás
- a többi modul bemenete az előtte lévő modul kimenete



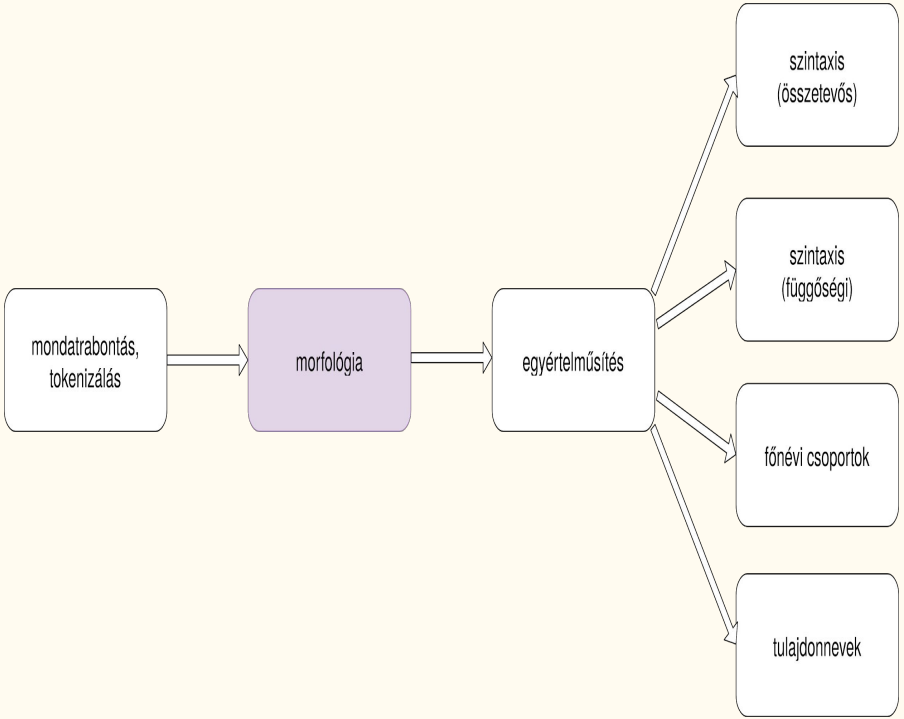
Mondatra bontás, tokenizálás

- bemenet: sima szöveg
- mondatra bontás
  - minden mondat!
- mondatokon belül szavak, szóközök és írásjelek azonosítása



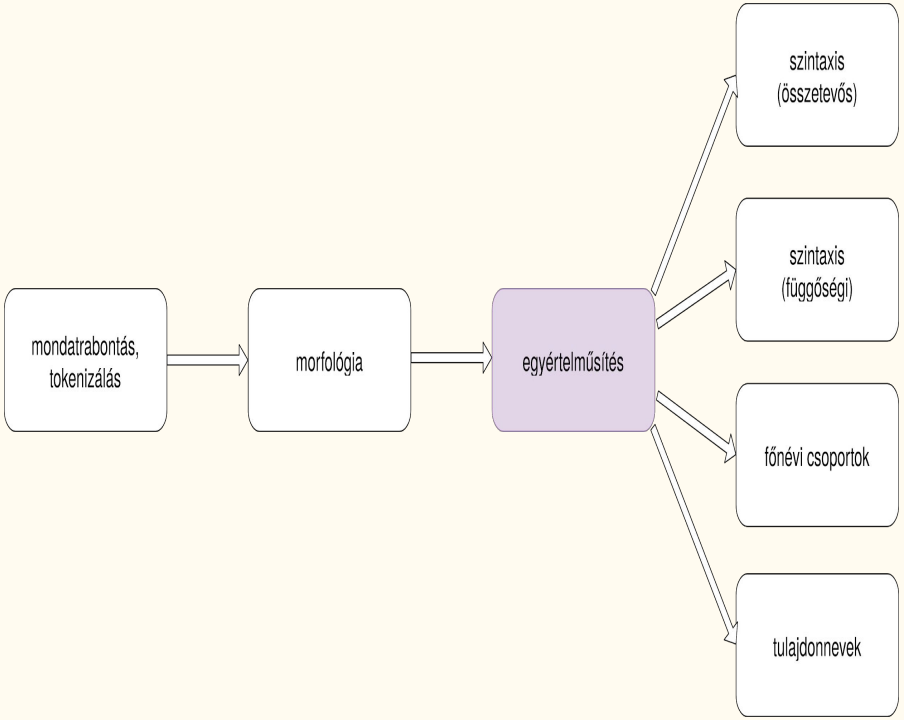
Morfológiai elemzés

- bemenet: szavak
- a szavak lehetséges morfológiai elemzését állítja elő
  - sok rejtett többértelműség
  - nincs kontextus
- lehetséges elemzéshez tartozó szótő és szófaj megállapítása



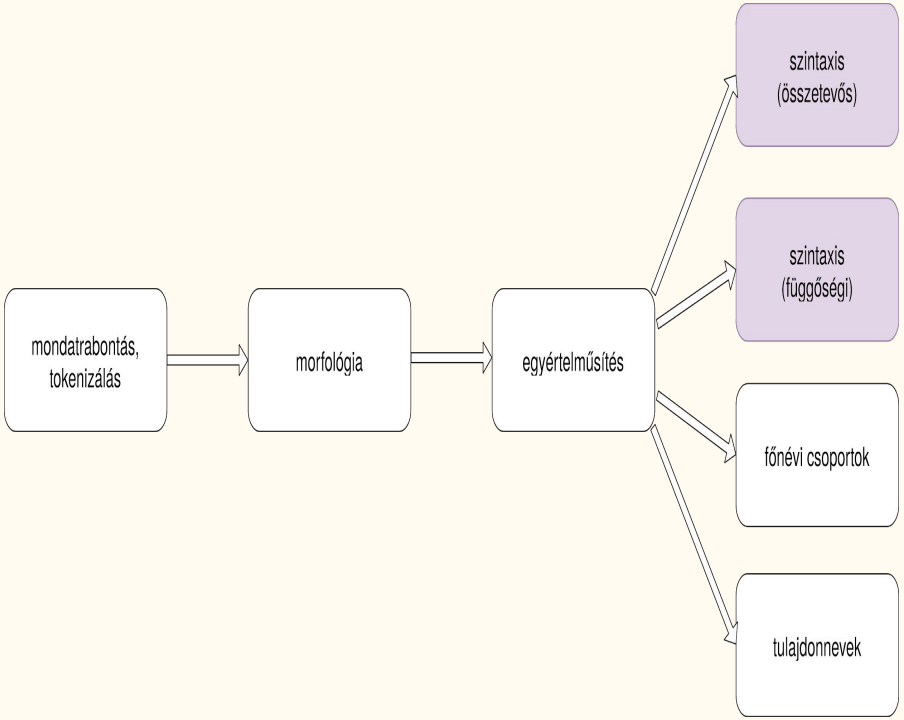
Egyértelműsítés

- bemenet: szavak listája a lehetséges elemzéseikkel együtt
- nézi a szavak kontextusát, és kiválasztja a legvalószínűbb elemzést (a hozzá tartozó szótővel és szófajjal együtt)



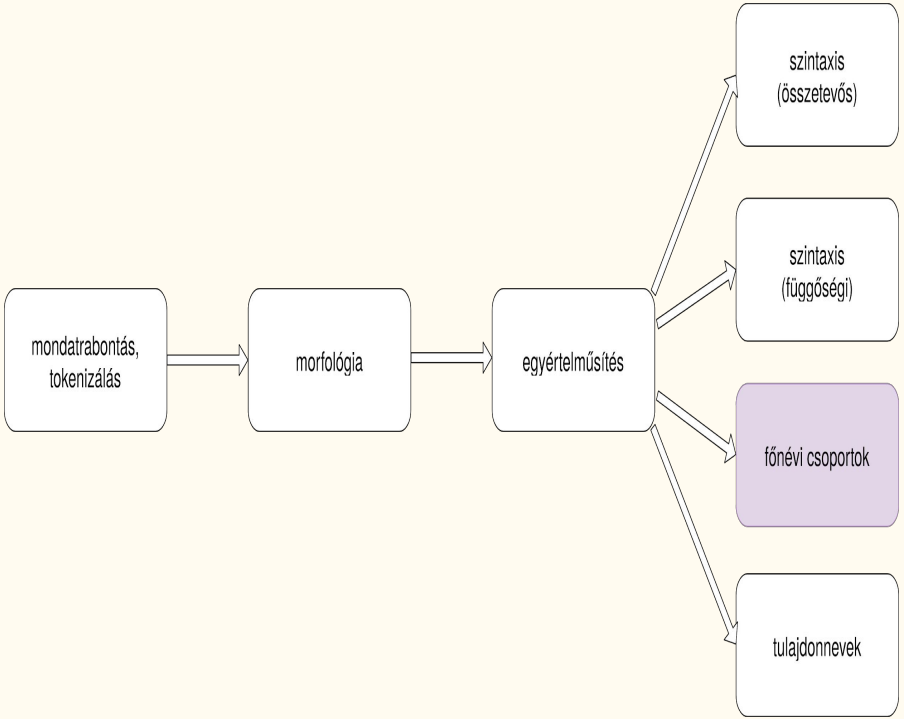
Szintaktikai elemzés

- bemenet: mondat
- összetevős elemzés
  - kifejezések, ágrajz
- függőségi elemzés
  - szavak közti függőségi viszonyok



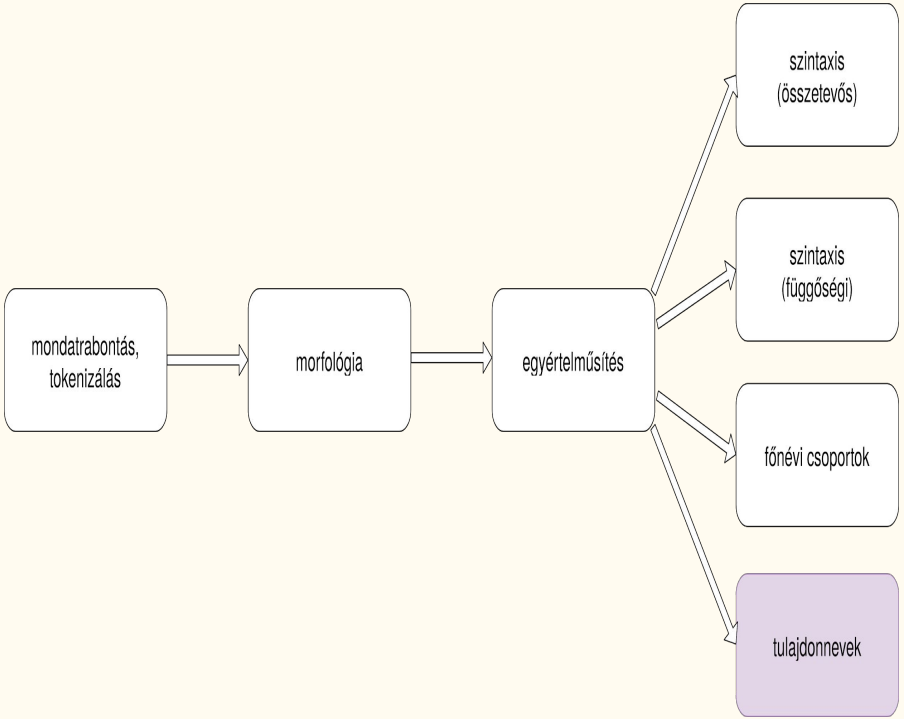
Főnévi csoportok

- bemenet: mondat
- főnévi kifejezések azonosítása
  - maximális
  - sekély



Tulajdonnév-felismerés

- tulajdonnevek keresése
  - személynevek
  - helynevek
  - intézménynevek
  - *egyéb*



Korlátok

- pontosság és fedés
- helyesírás
- szleng, közösségi média

Hasznos linkek

- honlap: http://e-magyar.hu/
  - szövegelemző: http://e-magyar.hu/hu/parser
- github: https://github.com/dlt-rilmta/hunlp-GATE
- cikkek: MSZNY 2017, 49-112. o.
  - http://rgai.inf.u-szeged.hu/project/mszny2017/files/kotet.pdf

Köszönöm a figyelmet!

mittelholcz.ivan@nytud.mta.hu