Az igekötők gépi annotálásának problémái

Kalivoda Ágnes

Budapest, 2017. február 3.

PPKE BTK

Bevezetés

Mi a probléma? Homográf szóalakok hibás szófaji címkét kaphatnak

Mi a megoldás? Szabály alapú javítás (kontextus mintázatai)

Mi a haszna? Jobb annotáció → jobb alapanyag

Az előadás felépítése

1. Módszertan, a felhasznált korpuszok
2. Homográfia példákkal, halmazokkal
3. A javítás lépései (script)
4. Kiértékelés: pontosság, fedés, f-mérték
5. Összegzés, további teendők

Módszertan

- Felhasznált korpuszok:
  - **Pázmány Korpusz** (Endrédy 2016)
  - **Magyar Nemzeti Szövegtár v2.0.4** (Oravecz–Váradi–Sass 2014)
- Igekötő állomány (egytagú, elvált igekötők):
  Pázmány: **76**, MNSZ2: **79** (+ *alább, benn, közé*)
- Munkafázisok:
  - minden IK-nak elemzett szó lekérése a Pázmány Korpuszból
  - ugyanezen szavak lekérése nem-IK-ként
  - hibatípusok megállapítása (Unix pipeline commands)
  - szabályok megfogalmazása (Python3.4 script)
  - tesztelés: MNSZ2, pszeudo-random 5000 mondat

Homográfia – példákkal

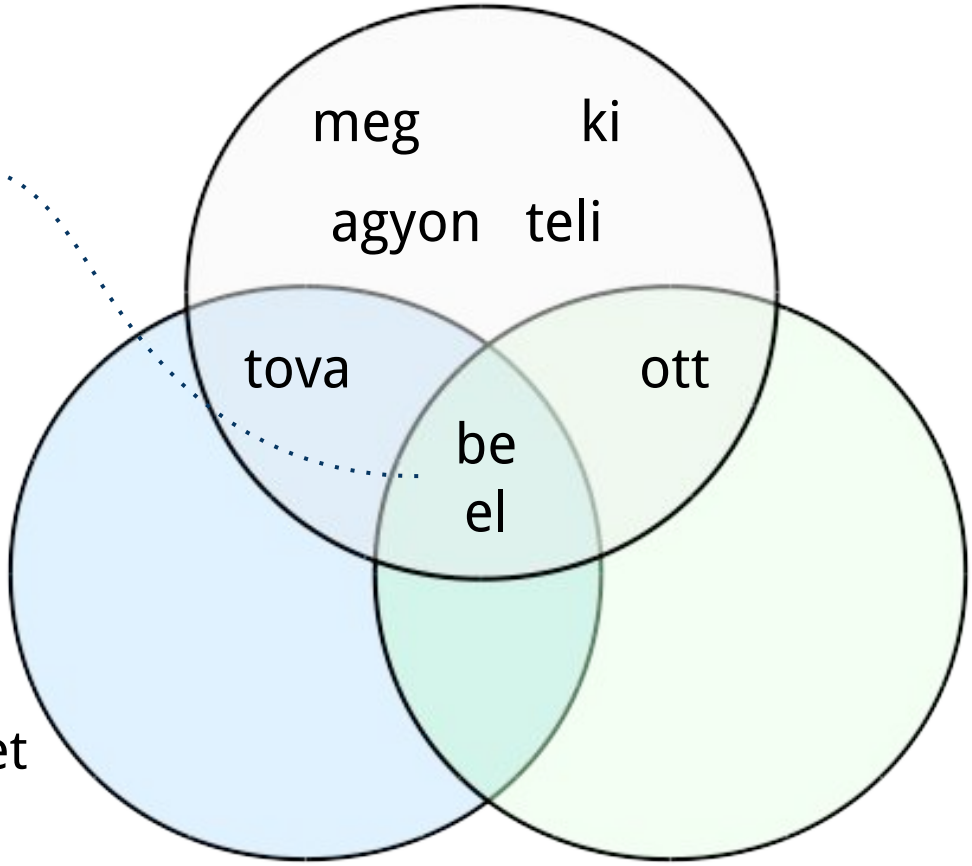
Más magyar szófaj	Ha a karbantartó sokáig vacakol, akkor az a konklúzió, hogy nem **ért hozzá**.	ért valamihez? hozzáér?
	4 telót rendeltem online tudom mit **ír ki**.	kiír valamit? (vala)ki ír?
Más nyelvű szóalak	[…] és a szingapúri Clara Vong **Van Ki**.	én kivagyok, ő kivan…
	[…] **vitte** a Sharm **el** Sheiki kórházból a per helyszínére.	elvitte?
Rövidítés, mozaikszó	LE (lóerő), OTT (oxytocin terheléses teszt)	
+ Elírás	[…] mert amit **meg ad** számot azon nem lehet!	megad? és amit ad?

Homográfia – halmazokkal

be (1):
Nem jött be.
Se ki, se be.
Be rút teremtés!

be (2):
It must be late.

be (3):
BE = Bowling
Egyesület



A javítás lépései (script)

1. Beolvassuk a mondatot
   ⇒ tejet/tej/FN.ACC meg/meg/IK ilyeneket/ilyen/MN_NM.PL.ACC

2. Keressük az igekötőnek jelölt elemet, mik a hibatípusai?
   meg = [1]

3. Adott hibatípusban több szabály, van illeszkedő?

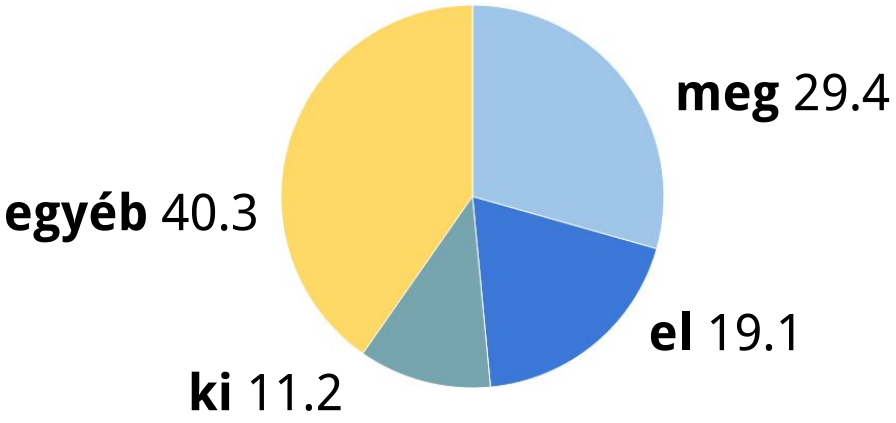
   [névszó+esetrag] [0-3 nem .?!] **[meg]** [0-3 nem .?!] **[névszó+esetrag]**

4. Ha valamelyik szabály talált: nem igekötő, különben igekötő

- Egyelőre egy irányban működik (igekötő ⇒ nem igekötő)
- Igekötő+ige kombinációkat is keresgél
  (csak bizonyos igekötőknél, pl. *észre, létre*)

Kiértékelés: módszer

- 5000 igekötős mondat az MNSZ2-ből, “majdnem” véletlenszerűen
- fő szempont:
  az igekötők arányát tekintve az MNSZ2 kicsinyített mása (alább %-ok)



- **összesen: 5000 / 1033 hibás 20.6 %**
- meg: 1429 / 510 hibás 35.6 %
- ki: 563 / 100 hibás 17.7 %

Kiértékelés: eredmények

- **Pontosság:** a hibásnak jelölt annotációk közül mennyi valóban hibás?
- **Fedés:** az összes hibás annotáció közül hányat talált meg?
- **F-mérték:** a pontosság és fedés harmonikus közepe

Teljesítmény a tesztkorpuszon (%):

Pontosság:	88.2
Fedés:	57.5
F-mérték:	**69.6**

------►

Miért ennyi?

1. Nehéz kontextusra támaszkodni, ha rossz a kontextus:
   ír/ír/MN.NOM alá/alá/IK

2. Rengeteg az elírás, hibás szóköz, ezt egyáltalán nem kezeljük
   fél → fel, még → meg, el → le ...

Összegzés

A helyzet javult, de bőven van még mit csinálni…

További kérdések, teendők

- a fedés javítása … de mit lehet kezdeni az elírásokkal?
- javítás a másik irányból (nem-IK ⇒ IK)
- az igekötő állomány vizsgálata kicsit elméletibben
  - nem konzisztens: ha a *zokon* igekötő, a *cserben* miért nem?
  - határozószó vs. igekötő, névmás vs. igekötő:
    nagyon sokszor nem dönthető el, pl. *oda* [IK|HA]?
- rövid válaszok kezelése
  Megtartottad az előadást? Én meg. (igekötő)
  Te megint ülsz le játszani… Én meg? (kötőszó)

Köszönöm a figyelmet!

Hivatkozások

- Endrédy István (2016):
  Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz. PhD. disszertáció.

- Kalivoda Ágnes (2016):
  A magyar igei komplexumok vizsgálata. Mesterszakos szakdolgozat. Budapest, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar.

- Makrai Márton (2007):
  Többértelműségek magyar mondatok számítógépes elemzésében – a „meg” szó szófajának vizsgálata gyakoriságokkal. Témalabor dolgozat.

- Oravecz Csaba – Váradi Tamás – Sass Bálint (2014):
  The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014.* Reykjavík. 1719–1723.