
Az igekötők gépi annotálásának problémái

Kalivoda Ágnes

Budapest, 2017. február 10.

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 - Korpuszok

Korpusz neve	MNSZ v2.0.4 (Oravecz et al. 2014)	Pázmány Korpusz (Endrédy 2016)
Korpusz mérete	1,04 milliárd token	1,2 milliárd token
Szófaji egyértelműsítés	TnT, módosított (Oravecz et al. 2002)	PurePOS (Orosz et al. 2013)
Igekötő állomány	79 + <i>alább, benn, közé</i>	76
Ismeretlen szavak	SKIP címke	mindent elemez



all dogs must/must/FN+ACC **be**/be/IK muzzled

Módszertan - Munkafázisok

1. minden igekötőként elemzett szó lekérése a Pázmány Korpuszból
2. ugyanezen szavak lekérése nem-igekötőként
a lehetséges elemzések sora néha nagyon hosszú, pl. **mellé:**

Elemzés	NU	IK	HA hová?	FN+POS+NOM mié?	FN+FAC mivé?
Darab	89212	13014	128	57	1

3. hibatípusok megállapítása (Unix pipeline commands)
4. szabályok megfogalmazása (Python3.4 script)
5. tesztelés: MNSZ2, pseudo-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?

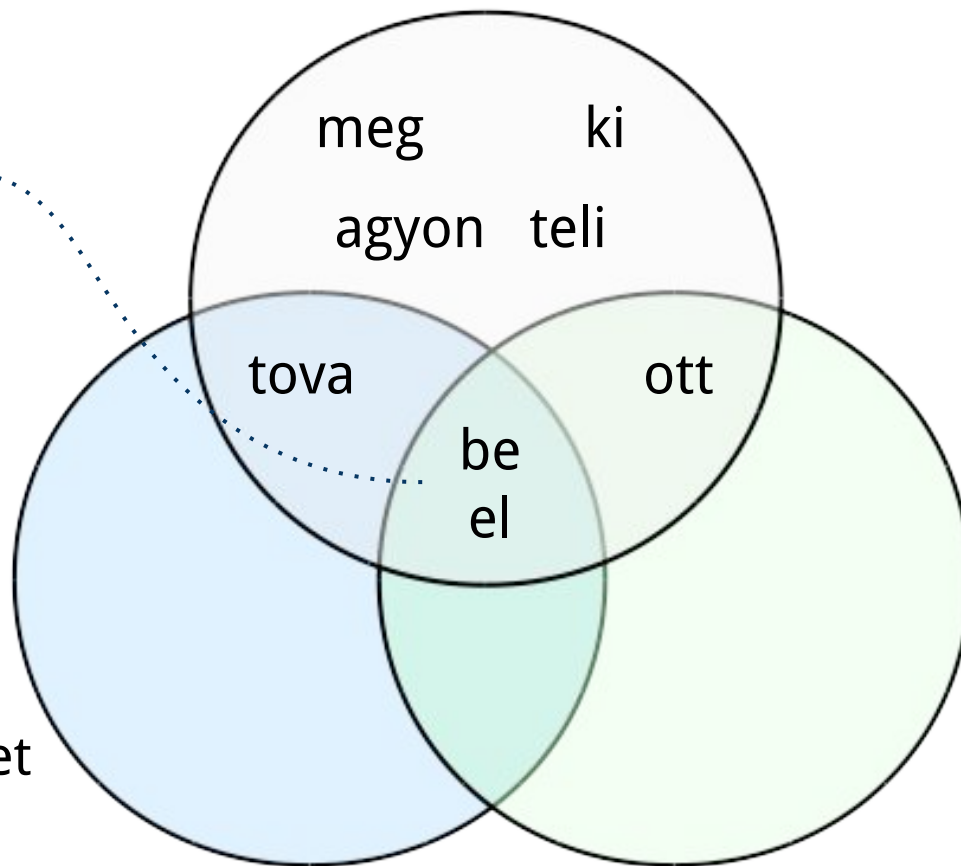
Homográfia – halmazokkal

Más magyar szófaj (1)

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

be (2):
It must be late.

be (3):
BE = Bowling
Egyesület



Elírások
utol só
össze gért
...

Más nyelvű szóalak (2)

Rövidítés, mozaikszó (3)

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)

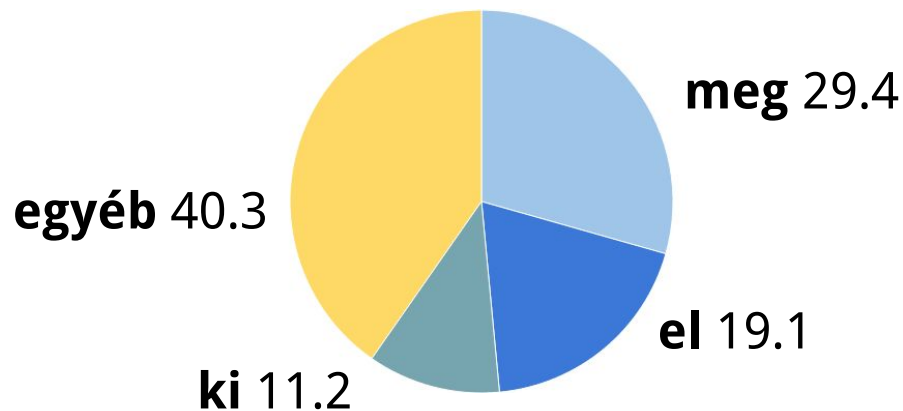
A javítás lépései (script)

Bővebben a szabályokról:

- Általános, minta alapú:
Pl. **át**: ha superessivus ragos névszót követ, akkor nem igekötő
ezt hallgattuk órákon át ✓ ↔ *mentünk a hídon át* ?
- Speciális, lexikon alapú:
Pl. **el**: kivételek felsorolása (pl. *El País, El Greco, Sharm el Sheiki ...*)
Probléma: folyamatosan jöhetnek új kivételek
- Igekötő+ige kombinációk:
Pl. **észre**: *szépség mellett észre is szükség van* → észre+van ✗
ezt a hibát korábban észre sem vettem → észre+vesz ✓
probléma: félreelemzett igék, pl. *észre is vétette* → észre+vét ?

Kiértékelés: módszer

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



- **összesen:** 5000-ből 1033 hibás $\Rightarrow$ **20.6 %**

Kiértékelés: hibák aránya

Igekötő	Darab	Helyes	Hibás	Hibásak %-a
meg	1470	960	510	34.7
ki	563	463	100	17.8
át	118	21	97	82.2
el	956	901	55	5.8
rá	70	22	48	68.6
hozzá	77	37	40	51.9
be	368	341	27	7.3
vissza*	123	96	27	21.9
fel	355	332	23	6.5

Kiértékelés: eredmények

- **Pontosság:** a hibásnak jelölt annotációk közül mennyi valóban hibás?

$$\frac{\text{valóban hibásak} \times 100}{\text{hibásnak jelöltek}}$$

- **Fedés:** az összes hibás annotáció közül hányat talált meg?

$$\frac{\text{megtalált hibásak} \times 100}{\text{összes hibás}}$$

- **F-mérték:** a pontosság és fedés harmonikus közepe

$$2 \times \frac{\text{pontosság} \times \text{fedés}}{\text{pontosság} + \text{fedés}}$$

Kiértékelés: eredmények

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 $\Rightarrow$ IK)
- az igekötő állomány vizsgálata kicsit elméletiben
 - 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.
- Makrai Márton (2007): [Többértelműségek magyar mondatok számítógépes elemzésében – a „meg” szó szófajának vizsgálata gyakoriságokkal](#). Témalabor dolgozat.
- Oravecz Csaba – Dienes Péter (2002): [Efficient Stochastic Part-of-Speech Tagging for Hungarian](#). In: Proceedings of LREC 2002. Las Palmas. 710–717.
- Oravecz Csaba – Váradi Tamás – Sass Bálint (2014): [The Hungarian Gigaword Corpus](#). In: *Proceedings of LREC 2014*. Reykjavík. 1719–1723.
- Orosz György – Novák Attila (2013): [PurePos 2.0: a hybrid tool for morphological disambiguation](#). In: Proceedings of Recent Advances in Natural Language Processing 2013. Hissar. 539–545.