28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet

Sass Bálint
MAGYAR TUDOMÁNYOS AKADÉMIA
NYELVTUDOMÁNYI INTÉZET
sass.balint@nytud.mta.hu

Két nagy méretű, magyar nyelvi erőforrást teszünk közzé. Elérhetők a **http://corpus.nytud.hu/isz** címen.

❶ a régi MNSZ tagmondatai sekély szintaktikai elemzéssel: Mazsola adatbázis [1]

Tartalom, formátum

igekötő! „engem” fni volt! alany lett!

engem meg sem hallgattak . *stem@@meghallgat ACC@@én*
A hasmenéstől szenvedő betegeknek sokat kell inniuk , *stem@@iszik ACC@@sok NOM@@beteg*
A Profi egyik támadójátékosa elhúzta mellettem a labdát ,
stem@@elhúz ACC@@labda mellett@@én NOM@@támadójátékosPOSS

„mellettem” birtokos személyjel

- alapvetően az igei szerkezetek kinyerése céljából jött létre
- erőforrásként önmagában közzétesszük a további felhasználás érdekében

Így készült

régi MNSZ (187 millió szó) + tagmondatra bontás + részleges szintaktikai elemzés
= mi az **ige**? (igekötő, főnévi igenév) **+** mik a felső szintű **NP/PP bővítmények**?

(részletek: [2] 2.2. fejezet)

Mire jó? – Kutatási ötletek

heterogén →
1. száll -rA 610
2. száll 463
3. száll vita-bA -vAl 359
4. száll -bA 292

vonzat ❹ →
5. száll -ért sík-rA 150
6. száll -ért harc-bA 142
7. száll -bAn 141
8. száll -vAl 134
9. száll ring-bA 103

dicsőség; vér; ital ❶ →
10. száll fej-A-bA 101

11. száll =mellett sík-rA 94
12. száll vonat-rA 80 ← villamos, busz, hajó . . . ❶
13. száll maga-A-bA 72
14. száll -n 71 ← szabad határozó ❸
15. száll sík-rA 69
16. száll -bA -vAl 67
17. száll -ért ring-bA 65
18. száll part-rA 64 ← komplex ige ❷
19. száll harc-bA 63
20. száll -rÓl -rA 61

❶ **szemantikailag koherens szóosztály** (több ilyen osztály is lehet!)
= literális jelentésű szavak adott vonzati helyen – pl.: *eszik* tárgyai: ételek

❷ **„kakukktojásként”**: komplex igék, idiómák, szólások
– *eltörik* alanyai: testrészek ↔ DE! *mécses*

❸ *-bAn*, *-n* a szabad határozók miatt jelenik meg; **eloszlásuk → vonzatság?** pl.: szerepel -bAn

❹ **vonzatok kötelezősége**:
felszólít, felkér, tanít: -t ≫ -t -rA → -rA nem kötelező?
bíz, kényszerít, alapoz: -t ≪ -t -rA → kötelező -rA ragos vonzat?

❺ bővítményi szó → jellegzetes igei szerkezetek:
vagyon → *szert tesz*, *felél*, *gyarapít*, *elkoboz*, *kiforgat*; *kenyér* → *eszik*, *süt*, *szel*, *keres*, *visszadob*.

← ❷ ebből az adatbázisból automatikusan származtatott igeiszerkezet-lista

Tartalom, formátum

becsap -t 1248
lecsap -rA 620
mér csapás-t -rA 360
átcsap -bA 345
megcsappan 217
lesz csapadék 205
csap -t hón-A=alá 80
becsap ajtó-t maga=mögött 28
átcsap =fölött 20

soronként: igei szerkezet + gyakorisági mérőszám

← **vonzatos komplex ige** – lexikailag szabad + kötött bővítmények

gyakorisági mérőszám:
„ennyi olyan mondat volt a korpuszban, ami megfelel az adott szerkezetnek, és nincs olyan specifikusabb szerkezet a listán, aminek megfelelne”
becsap + tárgy – hány ilyen mondat van? össze kell számolni!

- kompozicionális szerkezetek is! = szabad határozók ill. literális jelentésű gyakori szavak
- a Magyar Igei Szerkezetek [3] szótárral összevetve: tisztítatlan, nyers adat
- a szegedi FX listával [4] összevetve: sok félig kompozicionális szerkezet
- „teljes” szerkezetek: *zsebre vág* helyett vág -t zseb-rA
- a Mazsola adatbázis elemzése közel sem hibamentes
  előny: **nagy korpuszméret** → ritka szerkezetek azonosítása, gyakoriságuk becslése:
  visz prím-t -bAn, terjeszt rémhír-t, telik erő-A-bÓl -rA, tapos -t sár-bA
  → a kis plusz hozzáadott információt tartalmazó, de nagy méretű korpuszok hasznossága

Így készült

speciális igeiszerkezet-kinyerő algoritmus: mondatvázakat összesíti
- a ritka (fq≤5) mondatvázak → egy rövidebb, illeszkedő mondatvázhoz
- végül: mindent → a lehető legspecifikusabb meglévő mondatvázhoz

képes feltárni: **jellegzetes-e a bővítmény? + jellegzetes-e a konkrét szó?**
szemére vet vmit: -rA:fix -t:? ↔ *pillantást vet vmire*: -rA:? -t:fix

→ eredmény:
1. csak vonzat: *hisz vmiben*
2. kollokatív szerkezet: *süt (a) nap*, *döntés születik*
3. kombináció = vonzatos komplex ige: *szó van vmiről*, *igényt tart vmire*

(részletek: [2] 3.3. fejezet)

Hivatkozások

[1] Sass, B.: „Mazsola” – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest (2009) 117–129
[2] Sass, B.: Igei szerkezetek gyakorisági szótára - egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD thesis, PPKE ITK (2011)
[3] Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
[4] Vincze, V., Csirik, J.: Hungarian corpus of light verb constructions. In: Proceedings of COLING 2010, Beijing, China (2010) 1110–1118

XI. Magyar Számítógépes Nyelvészeti Konferencia, 2015. január 15-16., Szeged