Mitmesõnalised verbid

Heiki-Jaan Kaalep
TÜ

Mitmesõnalised leksikaalsed üksused

- Mitu sõna, üks tähendus
  - millised need üksused on?
  - kuidas neid tekstis ära tunda?
  - nt.”välja tulema”
    - valmis saama, korda minema
    - väljuma

Täitjad:

- Heiki-Jaan Kaalep, Kadri Muischnek, Liisi Pool, Kaarel Veskis, Martin Mets
- TÜ tudengid

Andmebaas

- 12 500 kirjet
- kirje kuju:
  - väljend sõnastiku-kujul
  - liik (ühendverb, ahelverb, tugiverb, noomenverb)
  - millistest sõnastikest pärit (7 välja)
  - sagedus korpuses
  - morf. analüüsi kujuline (üldistatud) esitus

nt. aega (obj) maha võtma; tuulde (adit) lendama

Korpus

- Iga osa 100 000 sõna
  - ilukirjandus
  - ajakirjandus
  - “Horisont”
  - seadused – liiga erinevad
- Morfologiliselt ühestatud (käsitsi)
- 8600 verbikeskset püsiühendit (käsitsi)

Püsiühendite sagedus tekstides

	sõnesid	lauseid	püsi	üksikuid põhiverbe
Ilu	104200	9000	4000	16800
Aja	111100	9500	2600	14500
Hor	99000	7300	2000	12600
Kokku	314300	25800	8600	42900

Püsiühendite märgendaja (programm)

- Sisendiks on morf. ühestatud tekst
- Kasutab andmebaasi
- Märgendab püsiühendid
  1. Märgib sõnad, mis võiksid verbiga koos ühendi moodustada
  2. Vaatab, kas ja millised on antud kontekstis ühendi koosseisus (osa filtreeritakse välja)

Näited

- Üle tuleks vaadata #<-üle vaatama# aknatihendid.
- Kui tal 1992. aastal õnnestus presidendiks saada, siis Arnold lihtsalt visati #->välja viskama# Kadrioru lossist välja.
- Savisaar lõi #->kaarte segi lööma# kaardid segi, lubades avalikult toetada Siim Kallast, kui too soostub presidendiks kandideerima.

Valiku algoritm

- Püsiühend ei saa ulatuda üle osalause piiri (osalause piirid leitakse automaatselt)
- Kuidas valida, kui lauses on mitme väljendi komponente?
  - eelistatakse püsiühendit, mille komponendid on üksteisele lähemal
  - eelistatakse pikemat püsiühendit lühemale
- S.t. filtreerimisel kasutatakse väljendi-kandidaate endid

Saak ja täpsus

- Saak: õigesti ära tuntud jagada tegelikult olemas olnute arvuga

  7958 / 8663 ≈ 92%

- Täpsus: õigesti ära tuntud jagada programmi poolt leitute arvuga

  7958 / 8815 ≈ 90%

Areng 2007-2008

- Nii saak kui täpsus suurenenud 80% pealt 90% peale
- Kuidas?
  - AB ühtlustamine
  - Korpuse ühtlustamine
  - Lisatud verbi + verbi ühendite (nt. *minema jooksma*) äratundmine => parem filter valede eemaldamiseks

vead

- Väljendikandidaatide äratundmine arvestab morf. analüüsi, et väljendite eri variante ära tunda:

  saba jalge (e jalgade) vahele tõmbama –>

  saba (obj) jalge (pl gen) vahele (K) tõmbama

- Sõnavorm on AB-s ja korpuses erinevalt analüüsitud:

  tõeks saama –>

  tõde (sg tr) saama või

  tõsi (sg tr) saama ???

⇒ s.t. liigne lingvistiline töötlus (algvormil polegi tähtsust!)

⇒ Lahendus: genereerida sõnavormid AB põhjal ja otsidagi korpusest sõnavorme

Aitäh!