

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
- TÜ tudengid

Andmebaas

- 13 000 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
- Morfoloogiliselt ühestatud (käsitsi)
- 8200 verbikeskset püsiühendit (pool-käsitsi)

Püsiühendite sagedus tekstides

	sõnesid	lauseid	püsi	lihtv
Ilu	104200	9000	3800	21200
Aja	111100	9500	2400	18000
Hor	99000	7300	1900	15500
Kokku	314300	25800	8100	54700

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# aknatiendid.
- 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

Hindamine

	Valesti	Õigesti	Kokku
ÜV	1419 (29%)	3487 (71%)	4906 (100%)
VNÜ	1165 (26%)	3296 (74%)	4455 (100%)
	2584 (28%)	6777 (72%)	9361(100%)

Saak ja täpsus

- Saagis: õigesti ära tuntud jagada tegelikult olemas olnute arvuga

$$\text{Kokku saagis} = 6800 / 8200 \approx 82\%$$

$$\text{ÜV saagis} = 3500 / 4000 \approx 87\%[1\]](#)$$

$$\text{VNÜ saagis} = 3300 / 4200 \approx 78\%$$

- Täpsus: õigesti ära tuntud jagada programmi poolt leitud arvuga

$$\text{Kokku täpsus} = 6800 / 7900 \approx 86\%$$

$$\text{ÜV täpsus} = 3500 / 4400 \approx 80\%[2\]](#)$$

$$\text{VNÜ täpsus} = 3290 / 3535 \approx 93\%[3\]](#)$$

Tüüpilised vead

- Eksimus osalause piiri tuvastamisel – 300
- Kaassõna on peetud mäarsõnaks – 500

Oletasime, et võime antud etapil ignoreerida muutumatute sõnade tüüpe, sest nad on automaatselt raskesti eristatavad ja just püsiühendite äratundmine võiks nende eristamisel abiks olla

- Tehnilised vead
 - “ei” ja “olema” on arvestatud kui põhiverbe – 300
 - muud tehnilised vead – 200

Aitäh!