

Andmebaasi koostamine Google'ga. Glossimine

09.12.2008

Andmebaasi koostamine Google abil

- Interneti eelised ja miinused.
- Päring Google's: mõned otsinguoperaatorid.
- Harjutamine
- Eksamitööst

Interneti eelised ja miinused

- Kolm eelist:

1) **SUURUS**: kõige suurem kirjutatud tekstide kogu maailmas

2) **KAASAEGSUS**: kõige kaasaegsem tekstide kogu maailmas

3) **ZHANRIROHKUS**: sisaldab igasuguseid tekstiliike ja zhanre; sisaldab ka palju sellist keelekasutust, mis on suulisele keelele väga lähedal.

- Neli miinust:

1) **EI LUBA OTSIDA LEMMASID**, vaid ainult spetsiifilisi sõnavorme; ei saa näiteks otsida kõiki *saama*-verbi esinemisi internetis, küll aga saab otsida sõnavormide *saab, sain, saadud, saades, saanuks* esinemisi.

2) **EBAMUGAV**: Google'i päringu tulemused on ebamugavas formaadis ja neid tuleb käsitsi läbi töötada.

3) **PUUDUB INFO** autori päritolu, haridustaseme ja muude isiklike omaduste kohta.

4) **Ei SOODUSTA STATISTILIST ANALÜÜSI.** Google annab ainult leidude umbkaudse arvu. Suurem probleem seisneb aga selles, et Google väljastab tihti ühe ja sama lehekülje mitu korda ja seega loeb ühte otsitava vormi mitu korda. Selleks, et koostada usaldusväärne statistika tuleb kõik Google'i leiud läbi vaadata ja kokku lugeda.

- Seetõttu on Google't parem kasutada siis kui

1) otsime **haruldasi** sõnu või grammatilisi konstruktsioone, mida on kergem kokku lugeda ja mida teistest korpustes ei leia. Näiteks kui tahame uurida, kas kirjanike Andrus Kivirähu, Kaur Kenderi ja Mehis Heinsaare stiil ja teemakäsitus on eesti publiku jaoks kergesti äratuntavad, võime otsida võrgust omadussõnu *kivirähulik*, *kenderlik* ja *heinsaarelik*.

Või näiteks kui me tahame teada, kas Aaviku uudissõnad *sulnis* ja *älnis* on ka tänapäeval kasutusel.

2) me ei ole kindlad teatud lause või konstruktsiooni **grammatilisuses**. Näiteks lause *armastan õppida prantsuse keelt* ei kõla väga hästi. Kuid äkki on võimalik armastada õppida teisi asju. Seetõttu tasub päringu “armastan õppida” tulemuste hulgas otsida selliseid, milles *õppida*-verbil on sihitis.

Proovige ise!

Päring Google's

Oluline teada enne alustamist

- Kui me otsime mõnda eesti sõna, siis tuleks eelistuste menüüs määrata otsingu keeleks eesti keele. Muidu on võimalik, et sama sõnakuju esineb ka mõnes teises keeles, ja sellisel juhul Google väljastaks tunduvalt rohkem esinemisi kui vaja.
- Kõige kergem on tulemusi lapata siis kui eelistuste menüüs on valitud “kuva 100 tulemust lehe kohta”.

Mõned kasulikud otsinguoperaatorid

- **Ülekomadega** saab otsida konkreetset sõnaühendit, mille elemendid on teatud järjekorras ja mille vahel ei ole muid sõnu. Võrrelge päringud *alles olnud*, "*alles olnud*" ja "*olnud alles*".
- **Tärn** markerib otsitud sõna(ühendile) eelnevat või järgnevat elementi ja niiviisi kergendab väljastatud tulemuste silmitsemist. Kui me tahame näiteks teada, keda on nimetatud silmakirjalikuks ja valelikuks, kirjutame päringu "**on silmakirjalik ja valelik*"; kui me tahame teada, milline sõna järgneb kollokatsioonile *sa oled mul* kirjutame "*sa oled mul**"

- **OR** operaatori abil saab otsida emba-kumba kahest sõnast. Näiteks kui me otsime väljendverbide *ennast välja puhkama* ja *ennast välja elama* põhivormide esinemisi, võime päringulahtrisse kirjutada "ennast välja puhkama OR elama"
- **Miinusmärgiga** saab välistada sõna. Kui me tahame teada kui tihti esineb väljend *untsu minema* muudes ajavormides kui lihtminevikus kirjutame päringulahtrisse *-läks untsu*.

Otsige inimest, kelle nimi on *Hansen Tammsaare*, kuid ta ei ole *Anton Hansen Tammsaare*!

- **Kaks punkti (..)** kasutatakse sellest, et piirata numbrivahemiku. Näiteks päring 5..10 annab numbrid 5, 6, 7, 8, 9 ja 10. Päring "5..100 krooni võrra kallim" annab Google's 25 vastet, päring "1000..15000 krooni võrra kallim" aga 4 vastet.

Uurige kui tihti esineb fraas *võttis ... kilo alla* siis kui puuduv kilode arv on vahemikus 10-50 ja siis kui ta on vahemikus 50-100!

- **Site:** väljastab tulemusi ainult teatud lehekülgedelt. See on väga kasulik operaator siis kui me otsime teatud sõna või väljendi esinemist teatud tüüpi tekstides. Näiteks kui me otsime vormi *ei midagit* esinemist ajakirjanduses, võime kasutada päringut "ei midagit" site:www.postimees.ee

Uurige sõna *minetea* esinemust Päevalehes (www.epl.ee)!

- **Aja piiramine.** Google võimaldab uurida sõna või väljendi esinemist erinevates ajavahemikes, mis kõik lõpevad tänapäevaga. Tuleb minna Google täpsema otsingu menüüsse ja sealt valida 'Tagasta leheküljed, mida nähti esmalt'. Variandid on 'millal iganes', 'viimase 24 tunni jooksul', 'viimase nädala jooksul', 'viimase kuu jooksul' ja 'viimase aasta jooksul'.

Proovige erinevaid variante sõnavormiga *tagandamisohus*. Mitu tulemust leiate iga ajavahemiku kohta?

Rohkem infot Google otsinguoperaatorite kohta leiate järgmistel lehekülgedel:

- <http://www.google.com/help/operators.html>
- <http://www.google.com/support/bin/static.py?page=searchguides.html&ctx=advanced>
- http://www.google.com/advanced_search

Glossimine

- Glossimine tähendab **tõlkimist sellisesse metakeelde**, mis sisaldab infot sõnade ja nende koostisosade tähenduse ja grammatiliste omaduste kohta. Glossimist võib nimetada ka **morfeemipõhiseks tõlkeks**; näiteks

Ma tule-n *Pärnu-st*

I come-PRS.1SG Pärnu-ELA

‘I come from Pärnu’

- **Milleks on seda vaja?**

Kui me kirjutame keeleteaduslikku uurimust ja keel mida me uurime erineb keelest, milles kirjutame, tuleb uuritava keele näitelaused varustada glossidega, et lugeja saaks aru nende grammatilisest struktuurist.

- **Glossimise juhend**

Mingit universaalset glossimise süsteemi ei ole; viimastel aastatel on hakatud ühe rohkem kasutama

Leipzig Glossing Rules:

<http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

- Glossimine hõlmab kahte tegevust:

1) **morfeemide tõlkimist** uuritavast keelest metakeelde, s.h grammatiliste morfeemide **asendamist** adekvaatsete grammatiliste lühenditega ehk **glossidega**; vt nt PRS, 3SG, INF ja PART järgnevas näitelause

Kaasa saa-b võtta 10 kroon-i.

with can-PRS.3SG take-INF 10 kroon-PART

‘One is allowed to take 10 kroons along.’

NB! Leipzig Glossing Rules’i grammatiliste lühendite nimekiri on kahjuks väga puudulik.

2) **morfeemide sidumist üksteisega** õigete märgendite abil; vt nt märgendid __, - ja . järgmises lauses:

Iga sell varja-b, nagu jumaltea-b
every guy hide-PRS.3SG as_if God_know-PRS.3SG
mihukes-t ametisaladus-t, aga
what_kind_of-PART professional_secret-PART but
paista-b, nagu ole-ksi-n mina üksi see
seem-PRS.3SG as_if be-COND-1SG I alone that
togu, kes midagi ei tea.
fool who nothing NEG know.CONNEG

‘Every guy is secretive as if hiding God knows what kind of a professional secret, but it seems to me that I am the only fool who doesn’t know anything.’

Mõned põhireeglid

- **Morfeeme eraldatakse sidekriipsuga** nii näitelauses kui ka glossides. Sidekriipsude arv näitelauses ja glossides peab olema sama; nt

Voodi-s on maga-tud.

bed-INE be.PRS.3SG sleep-PASS.PST.PTCP

‘The bed has been slept in.’

- Sõltuvalt uurimuse teemast ja muudest asjoludest võime olla **rohkem või vähem üksikasjalikud**; vrd:

Me läksi-me naabri-ga linna.

1PL go.ACT.PST-1PL neighbour-COM town.ILL

we went neighbour-COM town.ILL

- Kui **ühele sõnale** uuritavas keeles vastab metakeeles **mitu sõna**, kasutatakse allkriipsu:

Nägi-n jaoülema-t.

see.pst-1sg squad_commander-part

‘I saw the squad commander.’

- Mõnikord on erinevaid võimalusi; nad on kõik õiged; vrd:

Luge-si-n raamatu läbi.

read-PST-1SG **book.GEN** through

Luge-si-n raamatu-Ø läbi.

read-PST-1SG **book-GEN** through

- Proovige panna glosse järgmistele lausetele:

Kirjastus andis välja kaks raamatut Eesti luule kuldsest klassikast ja kolmanda maailmaluule klassikast.

Tugevad riigid ei uuri kunagi omaenda kuritegusid, seepärast ei tea me iial nende üksikasju.