

Korpusepäring Keeleveebis

Heiki-Jaan Kaalep

Filosoft

Portaal keeleveeb

- <http://www.keeleveeb.ee> ,
<http://keeleveeb.edu.ee>
- Sõnastikud ühispäringus (50+)
- Tekstikorpused

Täitjad:

- Rene Prillop, Heiki-Jaan Kaalep, Tarmo Vaino,
Katriin Tsepelina

Korpuste liidesed

- <http://www.murre.ut.ee/otsing.html>
- <http://www.eki.ee/corpus/>
- <http://www.cl.ut.ee/korpused/kasutajaliides/>
- <http://www.murre.ut.ee/vakkur/Korpused/korpused/>

Korpuse morfoloogiline märgendus

- Morfoloogiline analüsaator
 - FiloSoft
 - sõnastikupõhine (=speller)
 - oletamine (aktiivne morfoloogia)
 - Grammatilised kategooriad VVS järgi
- Statistiline ühestamine (trigrammidega Markovi peitmudel); vt (Kaalep, Vaino 2001)
 - Korrektsus 95% (Veskis, Liba 2008)
 - Mitmesus
 - nud, tud, on
 - algvorm, nt *manner v mander*

Korpusepäring

- eesmärk: tasakaal kasutusmugavuse ja võimaluste vahel
- paremad võimalused kui tavalise sõnavormi otsingu puhul, kuid kehvemad kui UNIXi käskudega saaks teha
- päringus sõnavorm, lemma, grammatiline info
- võimalik neid omavahel kombineerida ja seda ka mitme sõna puhul

Korpuste allikad

www.cl.ut.ee

- TÜ tasakaalus korpus
 - Ilukirjandus, ajakirjandus, teadus (15)
- TÜ koondkorpuse osad
 - Eesti ilukirjandus 1990- (5,6), «Postimees» (32,9), «Eesti Ekspress» (7,5), «Eesti Päevaleht» (87,9), «Maaleht» (4,3), «Horisont» (0,3), «Luup» (1,9), «Kroonika» (0,6), «Eesti Arst» 2002 - 2004 (0,7), «Arvutitehnika ja Andmetöötlus» (0,6), ajakiri «Agraarteadus» (0,3), Mitmesugused teadusartiklid (1,3), Doktoritööd (0,5)
- Plaan: ülejäänud osad TÜ korpusest

Näited

- Mitmuse osastaval on õigekeelsuse normi kohaselt võimalikud paralleelvormid, nt *dokumente v dokumentisid*
 - ...ente vs ...entisid
- Umbisikulist tegumoodi iseloomustab tegija ebamäärastamine, kuid tegijat saab esitada *poolt* abil, kusjuures tegija ei tohiks olla päris konkreetne isik (EKK)
 - ..takse, ...ti, ...tud kellegi poolt

Piirangud

- Vastuseks kuni 200 lauset
- Lühemat kui 3-tähelist sõnaosa ei otsita

Aitäh!