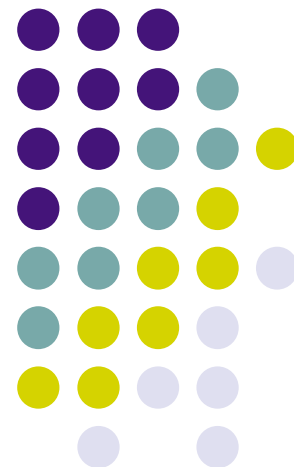


Eesti kirjakeele koondkorpus

Kadri Muischnek, TÜ





Milleks?

Programmi eesmärgid:

Suur kirjaliku keele korpus on vajalik kui:

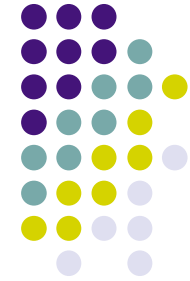
- formaalsete keelekirjelduste alusmaterjal
- elektrooniliste sõnastike ja andmebaaside koostamise abivahend
- keeleteadusliku uurimistöö materjal

Eesmärgid



- koondkorpuse arendamine 200 miljoni sõnani
- ajakirjandustekstide kõrval tuleb tähelepanu pöörata ka kirjaliku keele teistele allkeeltele (ilukirjanduskeel, teaduskeel) ning uut tüüpi kirjalikule keelekasutusele (jututoad, uudisgrupid)
- koondkorpus morfoloogiliselt märgendada

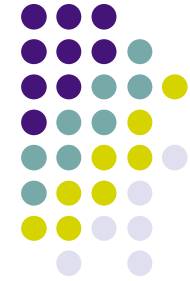
Veidi teooriat: suletud vs avatud korpus



Suletud korpus on representatiivne korpus ja sinna ei saa tekste lisada, ilma et tema representatiivsus kaoks.

Avatud korpus e monitorikorpus on selline, millesse pidevalt tekste juurde lisatakse. Representatiivsust ei taotleta. Kasutaja saab/peab ise koostama representatiivse valiku allkorpustest vastavalt oma vajadustele.

Mis tüüpi korpus on koondkorpus?



Seega: suletud vs avatud korpus =
koostaja vastutus vs kasutaja vastutus

Koondkorpus on avatud korpus, st kui kasutaja soovib mingi perioodi/allkeele suhtes representatiivset korpust, peab ta selle pakutavatest allkorpustest ise kokku panema.

AGA: Koondkorpuse alamhulk: Tasakaalus korpus



5 milj sõna ajalehekeelt

5 milj sõna ilukirjanduskeelt

5 milj sõna teaduskeelt

Võimaldab võrrelda kirjaliku keelekasutuse 3
tähtsamat tekstiklassi

Iga tekstiklass omaette on representatiivne teatud
perioodi suhtes

Kuid tervik ei ole representatiivne (sest pole
proportsioonis)

Koondkorpus: mis seal on 1



Eesti Ekspress 7 500 000 sõna

Postimees 33 000 000

Maaleht 4 300 000

Päevaleht töötluses (hinnanguliselt 50 miljonit)

Akadeemia (piiratud ligipääs) 7 200 000

Eesti Arst 700 000

Arvutustehnika ja andmetöötlus 625 000

Agraarteadus 300 000

Koondkorpus: mis seal on 2



Eesti Loodus 1 200 000

Horisont 260 000

Kroonika 600 000

Ilukirjandus 1995-... 6 000 000 (osa sellest
töötluses)



Koondkorpus: mis seal on?

Teadustekstid (sh dr-tööd) 3 900 000

Riigikogu stenogrammid 13 000 000

Jututoad 7 000 000 (töötluses)

Eesti seadused 1 800 000

Euroopa õigusaktide tõlked 9 600 000

Kuidas seda kasutada saab?



Kasutajaliides: päringule vastuseks saab terviklause ja soovi korral kuni 5-lauselise konteksti ette ja taha

Tekstid: Allalaaditavad www.cl.ut.ee/korpused

2006 ja 2007 plaanid ja tulemused



- Maaleht lausestada ja kasutajaliidese kaudu kättesaadavaks teha (korras)
- Kroonika kasutajaliidese kaudu kättesaadavaks (korras)
- Teadustekstide mahu suurendamine 5 miljoni sõnani (6,94 miljonit)
- Ilukirjandustekstide mahu suurendamine 5 miljoni sõnani (6 miljonit)

2006 ja 2007 plaanid ja tulemused 2



- Tasakaalus korpus (tekstid koos; kasutajaliidese alla panna)
- Jutukakorpus kasutajaliidese alla (töö käib)
- Päevalehe arhiiv korpuse kujule (töö käib)

Projekti eesmärgid



- koondkorpuse arendamine 200 miljoni sõnani
- ajakirjandustekstide kõrval tuleb tähelepanu pöörata ka kirjaliku keele teistele allkeeltele (ilukirjanduskeel, teaduskeel) ning uut tüüpi kirjalikule keelekasutusele (jututoad, uudisgrupid)
- koondkorpus morfoloogiliselt märgendada

Kes teevad?



Põhitäitjad

Kadri Muischnek

Heiki-Jaan Kaalep

Kaarel Veskis

+ TÜ üliõpilased ja magistrandid:

Liisi Pool

Siiri Pärkson

Kristel Uiboaed

Katrin Tsepelina