

Eesti keele koondkorpuse esituse ja kasutusvõimaluste arendamine

Kadri Muischnek

TÜ

Taust

Jätkab EKKTT projekti “**Eesti keele koondkorpus**” 2006-2009

- ✓ kirjaliku eesti keele kogu suurusega ca 245 miljonit sõna
- ✓ koostis: 75% ajalehetekste, 2% ilukirjandust, 9% nn uue meedia keelt, 2% teadustekste jms
- ✓ märgendati: TEI P3 standardi järgi
- ✓ märgendatud on
 - teksti jaotumine osadeks, nt ajalehes: number, rubriik, artikkel, lõik, lause
 - teksti autor, väljajätted (“siin oli tabel”), šrifti muutused
- ✓ päis e *header* dokumenteerib teksti päritolu, töötluse jms

Vabalt kasutatav, vt cl.ut.ee/korpused

Käimasoleva projekti eesmärk ja ülesanded

Eesmärk:

Koondkorpuse täiustamine ja tema kasutusvõimaluste laiendamine

Ülesanded:

1. korpuse teisendamine UTF-8-sse ja XML-keelde, üleminek TEI P5 standardile
2. luua kollokatsioonide leidja esialgne versioon
3. morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks

Korpuse teisendamine UTF-8-sse ja XML-keelde

ASCII -> UTF-8 – valmis

SGML-> XML - valmis

TEI P3 -> TEI P4 -> TEI P5 – valmis

Päiste (*headerite*) süsteemi muutmine - lõpusirgel

TEI P3 standardi järgi märgendatud korpuses oli 1 päis igal korpusefailil, sisaldas kogu infot nii tervikkorpuse kui ka selle üksiku korpusefaili kohta

TEI P5 standardi järgi märgendatud korpuses on 3 tasandi päised: korpusel kui tervikul, allkorpusel (nt Eesti Päevalehe allkorpus) ja üksikul korpusefailil

Kollokatsioonide leidja esialgne versioon

Kollokatsioonid: sõnad, mis esinevad üksteise naabruses sagedamini, kui võiks oletada nende korpuses esinemise sageduste põhjal, nt *saama aru, teada, hakkama, kätte, alguse*

St sõnapaarid (või –mitmikud), mille tegelik koosesinemine O on suurem kui nende oodatav koosesinemine E

Kollokatsioonide tuvastamiseks korpusest kasutatakse seose tugevuse statistikuid (*association measures*), mis peaksid näitama, kui olulisel määral on $O > E$

Selle kohta täpsemalt: Uiboaed (2010)

http://www.rakenduslingvistika.ee/ul/files/ERYa6.19_Uiboaed.pdf

Kollokatsioonide leidja esialgne versioon

Kollokatsioonide leidja praegu:

- ✓ leiab kollokatsioonid Tasakaalus korpusest (15 milj sõna)
- ✓ arvestab ainult kõrvuesinevaid sõnu
- ✓ sisestatava sõna saab valida: lemma või sõnavorm
- ✓ kollokaadid sõnavormidena

Seose tugevuse statistikud (praegu): *log-likelihood* (log-tõepära), *local-MI* (vastastikuse info väärtus), koosinemise sagedus

Väljund:

küsitud sõna (lemmana või sõnavormina) 150 olulisemat kollokatsiooni, järjestatud seose tugevuse statistiku väärtuse või sageduse järgi

Kollokatsioonide leidja esialgne versioon

Näidatakse demosesioonil; aadress praegu: www.rabauti.ee/clc

Programmi autor Katrin Tsepelina

Edasiarendamisvõimalused:

- ✓ suuremad korpused
- ✓ kontekstiks terve osalause
- ✓ kasutaja saab valida

kollokaadi sõnaliiki

kas kollokaatide puhul arvestatakse lemmat või sõnavormi

Morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks

Koondkorpuses on **22 milj sõna uue meedia keelekasutust**
(foorumid, kommentaarid, uudisgrupid, jututoad)

muu Koondkorpus on juba varem morfoloogiliselt märgendatud,
uus meedia mitte, sest

uue meedia keelekasutus **erineb normeeritud kirjakeelest** nii
oma **leksika** kui ka **ortograafia** poolest

allkeeled (foorumid, uudisgrupid, kommentaarid, jututoad)
erinevad

Morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks

Leksika

- ✓ **partiklid:** nt *tre, irw, ok, kle, kule, we, auts, icc, krt, oih, wau* jpt
- ✓ **emotikonid:** :) :(jpt
- ✓ **uudissõnad ja toorlaenud:** nt *loogish* (ka *logish*), *friik, tydo, privama, ruulima, diskilt buutima*
- ✓ **lühendid** (sõnakatked): *suht, tegelt, norm, aint* jms
- ✓ **kõnekeelsused:** nud-partitsiip: *läind* jpt, *midagist, prääga* jpt
- ✓ **kokkukirjutised:** *eiole, midaiganes, niiet, eksole*
- ✓ **võõrkeelsed sõnavormid**

Morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks

Ortograafia:

nn **ortograafiamängud**, nt

- ✓ ühe **tähe** või tähejärjendi **asendamine** teise või teistega:
raffas, näitex, täica, h6be, t88 jpt
- ✓ **tähe või silbi mitmekordistamine** emotsiooni väljendamiseks:
laheee, hahahahaha
- ✓ **sõnaalgulise h** ärajätmine: *ommik, uvitav, ullem* jpt
- ✓ **suur- ja väiketähtede** kasutamine mitte ortograafiareeglitest lähtuvalt, vaid emotsioonide väljendamiseks: APPII

kirjavead, trükivead: *tegematta, siiksi*

Morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks

Lahendused: morfanalüsaatorile *t3mesta* saab lisada

kasutajasõnastiku (tekstisõna -> lemma, gram kategooriad)

Uued sõnaliigid: partiklid, emotikonid

Lemmad kasutajasõnastikus:

partikli lemma on sama, mis tekstisõna, st *kle* ja *kule* on 2 eri partiklit

muude sõnaliikide lemmaks on (võimalusel) kirjakeelne sõna, nt *raffas* lemma on *rahvas*

Sagedased tundmatuks jäävad sõnavormid -> kasutajasõnastik

Vähemsagedased -> sellest järgmises ettekandes

Täitjad ja finantsid

Kaarel Veskis

Kristel Uiboaed

Katrin Tsepelina

Raul Sirel

Eriline tänu: Tarmo Vaino

Finantseering 2010: 550 000