Keelekogude osast keeleteaduses

Liina Lindström
Tartu Ülikool

Sisu

1. Kust saab keeleteadlane materjali?
2. Mis on korpus?
3. Keelekogust korpuseks murdekorpuse näitel
4. Eesti keele korpused

Kust saab keeleteadlane materjali?

- mõtleb ise välja (st toetub enda keelepädevusele);
- kogub küsitluste/testide teel (toetub teiste keelepädevusele);
- fikseerib tegelikku keelekasutust erinevates situatsioonides:
  - kogub materjali raamatutest, internetist; lindistab ise ja litereerib;
  - otsib olemasolevatest keelekogudest;
  - otsib korpustest.

Korpuste kasutamine

- Umbes pooled magistri- ja bakalaureusetöödest põhinevad korpustest pärit materjalidel.
- Korpuste materjale on väga palju kasutatud ka doktoritöödes, nt
  - Renate Pajusalu 1999. Deiktikud eesti keeles (suul. kõne korpus)
  - Ilona Tragel 2003. Eesti keele tuumverbid (kirjak. korpus, suul. kõne korpus)
  - Külli Habicht 2001. Eesti vanema kirjakeele leksikaalsest ja morfosüntaktilisest arengust ning Heinrich Stahli keele eripärast selle taustal (vana kirjakeele korpus)
  - Liina Lindström 2005. Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles (suul. kõne korpus)
  - Larissa Degel 2007. Intellektuaalsfäär intellektuaalseid võimeid tähistavate sõnade kasutuse põhjal eesti ja vene keeles (kirjak. korpus)
  - Kadri Muischnek 2006. Verbi ja noomeni püsiühendid eesti keeles (kirjak. korpus)
  - Merike Parve 2003. Välted lõunaeesti murretes. (murdekorpus)
  - Heiki-Jaan Kaalep 1999. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös (kirjak. korpus)
  - Pire Teras 2003. Lõunaeesti vokaalisüsteem: Võru pikkade vokaalide kvaliteedi muutumine. Dissertationes philologiae estonicae universitatis tartuensis 11.

Mis on korpus?

Arvutiajastul on **korpus** polüfunktsionaalne elektroonilisel kujul olev tekstikogu, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest (selle hetkeseisust või muutumisest).

Keelekogu on igasugune kollektsioon, mis sisaldab keelematerjali.

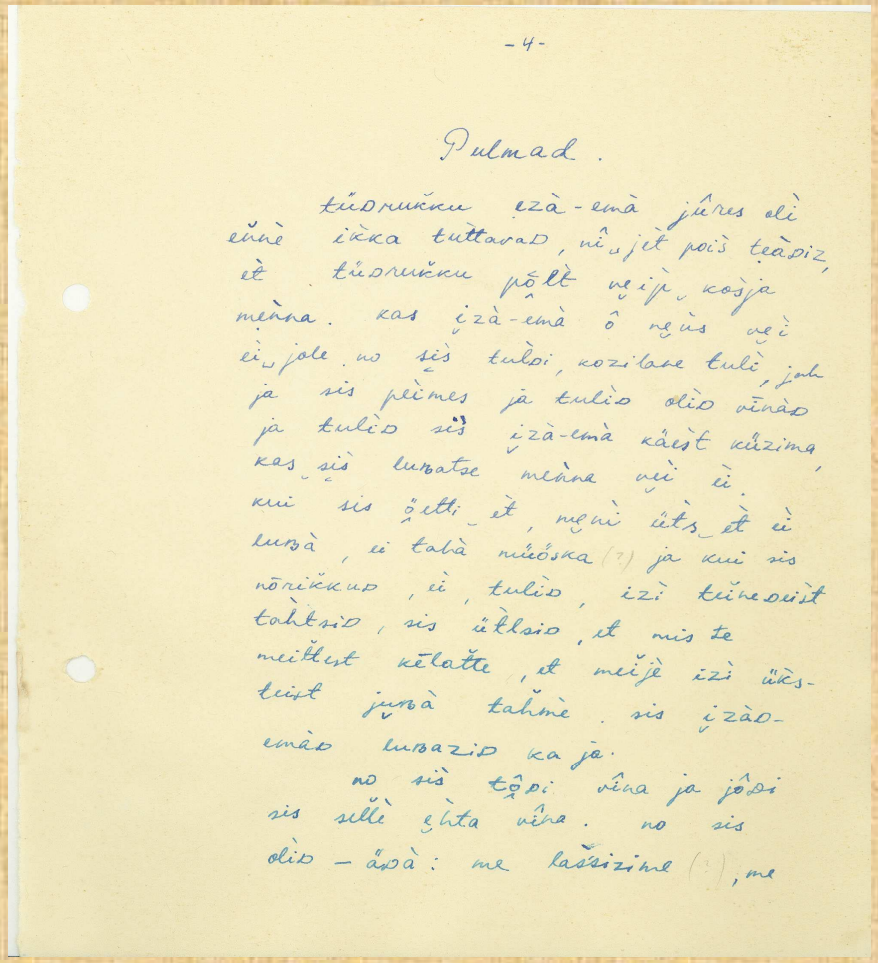
Mis on korpus?

- Korpus on ideaalis representatiivne keele / allkeele suhtes
- Korpus on ideaalis multifunktsionaalne
- Korpus on tehniliselt ühtlustatud ja märgendatud

- Praktilises elus nimetatakse korpuseks igasuguseid elektroonilisi keelekogusid, mis on teatud viisil korrastatud (ühtlustatud)

Keelekogust korpuseks: murdekorpus

- Lindistused EKI ja TÜ eesti ja sugulaskeelte arhiivist
- Lindistusele lisaks olemas ka transkriptsioon
- Sihtrühm: keeleuurijad



-4-

Pulmad.

tüdrukku ezà-emà jûres oli
eŭnè ikka tuttavad, nii jèt pois teadiz,
èt tüdrukku põlt wejè kosja
mènna. kas izà-emà ô nęüs wei
ei jole. no sis tulsi, kozilane tuli, joh
ja sis pèimes ja tulid olid vînàd
ja tulid sis izà-emà kàest küzima,
kas sis lubatse mènna wei ei.
kui sis õeldi èt, męni ütš èt ei
lubà, ei tahà nüóska (?) ja kui sis
nõrikkud, ei, tulid, izi türedest
tahtsid, sis üldsid, et mis te
meildest kõlatte, et meije izi üks-
teist jurà tahme. sis izàd-
emàd lubazid ka ja.
no sis tõdi vîna ja jõdi
sis sellè ęhta vîna. no sis
olid – õdà: me lassizime (?), me

Keelekogust korpuseks: murdekorpus

Tekstide valiku printsiibid:

- tekstid kõigist eesti murretest, igast murdest vähemalt kahest kihelkonnast;
- vanem kiht murdelindistusi (1960-1970ndad peamiselt), vanad kõnelejad
- klassikalised murdeintervjuud
- igast kihelkonnast mitu keelejuhti

Keelekogust korpuseks: murdekorpus

- Töö käik

1. Murdetranskriptsioonide ülekuulamine, täiendamine ja sisestamine (SU foneetiline transkriptsioon, Word, spetsiaalsed fondid)

rongu.mp3

rongu.pdf

Keelekogust korpuseks: murdekorpus

Lihtsustatud transkriptsiooni tegemine (vastab foneetilisele; multifunktsionaalne – saab kasutada kõikvõimalike programmidega jne; märgendatakse kõnelejate voorud, kommentaarid)

<com> Tartu murre, Rõngu, Pühaste küla, EMH 342, KJ = Juuli Antsik (82a, s 1879). Lind. 1961. HK = Hella Keem. </com>

<u who=KJ> inemine omm=jo `väega (.) näottu ku=ta (.) vanass lätt är </u>

<u who=HK> eij=ole </u>

<u who=KJ> (---) (...) nigu ärä kujunu (.) t'suug (...) </u>

<u who=HK> <com> naerab </com> jahh (.) ni=et (.) saa=nüt kas=sul=omm `meelen midägi oma=ni latsõ+põlvitsest värgist nii=kahh (...) </u>

<u who=KJ> mes=sa säält mõistat mälettädä viil (...) </u>

Keelekogust korpuseks: murdekorpus

3**. Morfoloogilise märgenduse lisamine**

Märgendamiseks nimetatakse interpretatiivse info lisamist suulist või kirjalikku keelt esindavasse keelekorpusesse

<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE record SYSTEM "morfo.dtd">
<record khk="RÕN" kla="Pühaste">
<com> Tartu murre, Rõngu, Pühaste küla, EMH 342, KJ = Juuli Antsik (82a, s 1879). Lind. 1961. HK = Hella Keem. </com>
<u who="KJ"><mark><sne>inemine</sne><msn>inimene</msn><mrf slk="S">sg n</mrf></mark>
<mark><sne>omm</sne><msn>olema</msn><mrf slk="V">ps ind pr sg 3</mrf></mark>=
<mark><sne>jo</sne><msn>ju</msn><mrf slk="Par"/></mark>
<mark><sne>`väega</sne><msn>väga</msn><mrf slk="Adv"/></mark>(.)
<mark><sne>näottu</sne><msn>näotu</msn><tah>kole</tah><mrf slk="A">sg n</mrf></mark>

- morfoloogiline andmebaas internetis: http://www.murre.ut.ee/

Murdekorpuse seis

Murre	litereeritud sõnu	märgendatud sõnu
Idamurre	69240	25617
Keskmurre	122815	77355
Läänemurre	146605	64620
Saarte murre	195971	98831
Alutaguse	43852	34176
Rannamurre	55246	42257
Mulgi	65392	31115
Tartu	67682	38064
Võru	93304	30593
Setu	100711	21604
keelesaared	24183	0
kokku	985001	464232

Valik murdekorpuse põhjal tehtud uurimusi

Liina Lindström, Varje Lonn, Mari Mets, Karl Pajusalu, Pire Teras, Ann Veismann, Eva Velsker, Jüri Viikberg 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. - Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. TÜ eesti keele õppetooli toimetised 17. Toim. R. Kasik, Tartu, lk 186-211**.**

Liina Lindström 2001. Eesti murrete korpuse iseloomustus argivestlustega võrrelduna. - Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. TÜ eesti keele õppetooli toimetised 17. Toim. R. Kasik, Tartu, lk 212-221.

Mari Mets 2004. Võru kõnekeel: *nud*-partitsiibi tunnuse varieerumine Vastseliina murrakus. Magistritöö Tartu Ülikooli eesti keele ajaloo ja murrete õppetoolis.

Mari-Liis Kalvik 2004. Kvantiteedisuhted rannikumurdes. Magistritöö Tartu Ülikooli eesti ja üldkeeleteaduse instituudis.

Liina Lindström, Karl Pajusalu 2003. Corpus of Estonian Dialects and the Estonian Vowel System. - Linguistica Uralica nr 4, lk 241-257.

Rutt Läänemets 2007. Sõna *pool* grammatiseerumine muutumatuteks sõnadeks eesti murrete näitel. Bakalaureusetöö.

Liina Lindström, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann, Eva Velsker 2006. Sõnaliigituse küsimusi eesti murrete korpuse põhjal. – E. Niit (toim.) Keele ehe. Tartu Ülikoolieesti keele õppetooli toimetised 30. Tartu. 154-167.

Mervi Kalmus 2007. Isikuühildumine ja pronoomeni väljajätt Muhu murraku näitel.Bakalaureusetöö

Merike Parve 2003. Välted lõunaeesti murretes. Dissertationes philologiae estonicae universitatis tartuensis 12.

Pire Teras 2003. Lõunaeesti vokaalisüsteem: Võru pikkade vokaalide kvaliteedi muutumine. Dissertationes philologiae estonicae universitatis tartuensis 11.

Kristel Uiboaed 2008. Ühendverbid eesti murrete korpuses. Magistritöö Tartu Ülikoolis.

TÜ kirjakeele korpus

- Sihtrühm: keeleuurijad, sõnaraamatute koostajad, keeletehnoloogia
- Peamised osad:
  - Baaskorpus (1980ndate korpus)
  - Läbilõikekorpus (1890-1990)
  - Koondkorpus e segakorpus
- Korpuse põhjal on välja töötatud automaatse morf. analüüsi vahendid ja süntaksi analüüsi vahendid

Üldmaht *ca* 180 miljonit

http://www.cl.ut.ee/korpused/kasutajaliides/

TÜ vana kirjakeele korpus (VAKK)

- Eesmärgiks on teha vana kirjakeele tekstid uurijatele ja huvilistele elektrooniliselt kättesaadavaks ja töödeldavaks.
- Võimaldab uurida keele ajalugu.
- Sisaldab eestikeelseid tekste alates kõige vanematest tekstidest kuni 19. sajandi I pooleni.
- Tekstid morfoloogiliselt märgendatud ja märksõnastatud – võimalik teha päringuid tänapäevase keele põhjal. (Pole veel veebis)

TÜ vana kirjakeele korpus (VAKK)

- Maht *ca* 1,8 mln sõna, sh:
  - vanimate eestikeelsete tekstide lauskorpus (16. sajandist ja 17. sajandi esimesest kümnendist, k.a käsikirjad) kuni 1660. aastateni. Umbes 900 000 tekstisõna.
  - 18. sajandi tekstide valikkorpus. Umbes 600 000 tekstisõna
  - 19. sajandi esimese poole tekstide valikkorpus. Umbes 300 000 tekstisõna

http://www.murre.ut.ee/vakkur/Korpused/korpused.htm

Käsikirjade näited:

http://www.murre.ut.ee/vakkur/Gooti/pildid.htm

TÜ suulise kõne korpus

- Kogutud kokku umbes 356 tundi
- Litereeritud umbes 300 tundi (730 linti)
- 2011 litereeritud teksti, 1333279 tekstisõna
-
- 1. Silmast-silma vestlused: 562 teksti, 722654 sõna
  - 184 argivestlust
  - 342 institutsionaalset (aktus, polikliinik, loeng, intervjuu, koosolek jne)
  - 36 muud (teeküsimine, külapood, torulukksepp, naabrid)
- 2. Telefonivestlused: 1297 teksti, 361330 sõna
  - 176 argivestlus
  - 1113 institutsionaalne (arst-patsient, bussiinfo, infotelefon, reisibüroo jms)
  - 8 muu (toa üürimine, auto ost, valeühendus)
- 3. Raadio- ja TV-saated: 152 teksti, 249295 sõna

- Morfoloogiliselt märgendatud ja ühestatud: 100 000 sõna 2008. aastal. (Kirjakeele morfol. analüüsi vahendeid on kohandatud suulisele keelele)

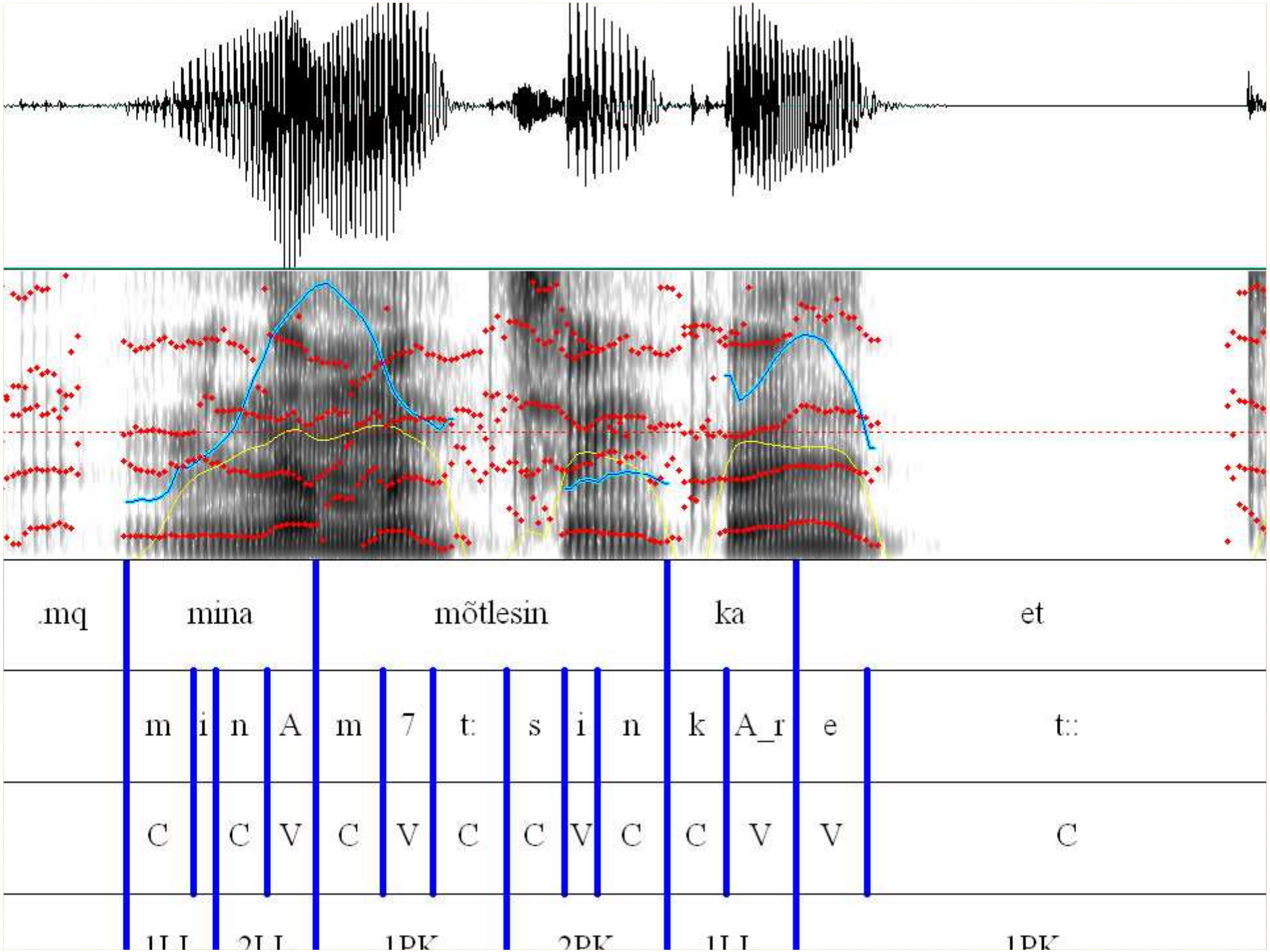
Dialoogikorpus

- Suulise kõne korpuse allkorpus
- Eesmärk: dialoogi modelleerimise programmide väljatöötamine
- märgendatakse suhtlusakte
- kogumaht 265071 sõna

TÜ spontaanse kõne foneetiline korpus

- Alustatud 2006
- sihtrühm: foneetikauurijad, keeletehnoloogia
- spontaanse ja poolspontaanse kõne lindistused
- programm Praat
- ca 10 tundi märgendatud helilindistust
- märgendus 5 tasandil:
  - sõnad (ortograafiline kirjaviis)
  - häälikud (SAMPA transkriptsioon)
  - häälikustruktuurid (CV)
  - silbid
  - kõnetaktid

helinäide


.mq
mina
mõtlesin
ka
et
m i n A m 7 t: s i n k A_r e t::
C C V C V C C V C C V V C
1LL 2LL 1PK 2PK 1LL 1PK

Muud korpused

EKI tekstikorpus

http://www.eki.ee/corpus/

- ca 10 mln tekstisõna, peam. aja- ja ilukirjandus (ajakirjandust ca 80%)
- juhuslik tekstivalik
- märgendamata korpus

EKI eesti emotsionaalse kõne korpus

http://urve.eki.ee:5000/

- Korpusel on kaks eesmärki:
- olla korpuspõhise emotsionaalse tekst-kõne sünteesi akustiline baas;
- olla usaldusväärne andmekogu kõnes avalduvate emotsioonide uurimiseks.
- erineva emotsiooniga sisse loetud laused: viha, rõõm, kurbus jne.

Õppijakorpused

- vajalik keeleõpetuses, vigade korpus

- **TLÜ eesti vahekeele korpus**

eesti keele võõrkeelena õppijate tekstid
vead on märgendatud
ca 550 000 sõna
http://evkk.tlu.ee/what_is_evk

- **TÜ emakeeleõppija korpus** – algusjärgus, ei ole veel internetis (Kadri Sõrmus)

emakeeleõppijate tekstid, peamiselt kirjandid
tekstid elektrooniliselt, märgenduspõhimõtted välja töötatud, osaliselt märgendatud
gümnaasiumiaste: ca 320 000 sõna
põhikooliaste: ca 12 000 sõna

TÜ eesti keele õppijate vigade korpus – algusjärgus, eesti keele võõrkeelena osakonnas (B. Klaas)

TLÜ lastekeele korpus (Reili Argus)

Kokkuvõtteks

- Korpused on arvutiajastul keeleuurijale äärmiselt vajalik materjaliallikas.
- Korpuste eelis on nende polüfunktsionaalsus: need ei ole koostatud vaid ühe konkreetse nähtuse uurimiseks, vaid võimalikult paljude nähtuste uurimiseks. Erinevad keeleuurimisalad võivad aga siis vajada põhimõtteliselt erinevaid korpusi (nt foneetikakorpused).
- Keeleuurija vajab kõige enam märgendatud korpust, sellise korpuse tegemine on aga suur töö.
- Keelekogud on korpuste koostamise eeletapp. Iga korpuse koostamiseks peab olema juba teatav keelekogu.