

Eesti kõnekeele korpuse kogumine ja translitereerimine 2004-2008

Tiit Hennoste

Tartu ülikool

Ajalugu ja üldist

- 1997-2000 suulise keele korpuse kogumine ETF-i granti abil (ca 300 000 üksust)
 - 2000-2003 raha ei eraldatud
 - 2004-2008 käesolev projekt
 - Eesti keelestrateegia eesmärgist: 5 aastaga 2 000 000 tekstisõnalise korpuseni jääb puudu, kuna raha on eraldatud oluliselt vähem kui küsitud
 - Korpusematerjalid võrgus:
<http://www.cl.ut.ee/suuline/>
-

Korpuse olemus

- ❑ keelekorpus (linguistic corpus)
 - ❑ suulise eesti keele (kõnekeele) suhtlustekstid loomulikult esinevates suhtlussituatsioonides (=situatsioon ilma keeleuurimise eesmärkideta)
 - ❑ avatud korpus: laiendatav lõputult
 - ❑ universaalne korpus: ei ole teoreetiliselt ette määratud rangelt eri situatsioonitüüpe
 - ❑ 3 alusliigendust: argi- ja institutsionaalsed suhtlused, dialoogid ja monoloogid, silmast-silma, telefoni- ja meediasuhtlused
 - ❑ lisaks arvestatud: suhtlejate vanus, sugu, haridus, sotsiaalne staatus
-

-
- põhiliselt audiolindistused, vähe videot
 - korpuse ainulaadsus: samaaegselt suur, mitmekesine, translitereeritud ja taustadokumenteeritud põhjalikult
 - translitereeritud ja varustatud suhtlussituatsiooni taustakirjeldustega
 - CA transliteratsioonimärgid ja reeglid: märgivad suhtluses olulisi parameetreid
 - taustakirjeldus: situatiivsed jooned, millel on leitud seosed keele varieerumisega
-

Konversatsioonianalüüsi transkriptsioon (kesksed märgid)

Sõnad

häälduspäraselt: *sis, vä, kule, onju, õõ, mhmh, mqm*

si- – sõna jääb pooleli

Intonatsioonilausungid

. – langev intonatsioon

, – poollangev intonatsioon

? – tõusev intonatsioon

Pausid

(.) – mikropaus (0,2 sekundit või vähem)

(1.2) – pikem paus, pikkus kümnenndiksekundites

Rõhk ja intonatsioonitõus

` *võimalik* – rõhk või intonatsiooni tõus

Kõnetempo ja hääle tugevus

>.....< kiirendatud lõik

<.....> aeglustatud lõik

*.*****.* ümbritsevast kõnest vaiksem jutt

AHA – ümbritsevast kõnest valjem jutt

Naer, Köhatatus, Venitus

hehe, mhemhe – naer

s(h)õna – naerdes öeldud sõna

t:ere, ma:terjalid – häälikute venitamine

-
- **Sisse- ja väljahingamine**
 - *.hhh* – häälekas sissehingamine
 - *.jaa* – sisse hingates lausunud sõna
 - **Pealerääkimised**
 - [– pealerääkimise algus
 -] – pealerääkimise lõpp
 - *tulin=ja* – kaks sõna on hääldatud kokku
 - **Ebaselgused ja kommentaarid**
 - {*või*} – ebaselgelt kuulnud sõna
 - {*--*} – väljakuulmatu sõna
 - (*(tuleb laua juurde)*) – litereerija kommentaar
-

Tekstinäide

- 261
 - Saade "Öös on inimesi", Kuku-raadio
18.11.1998
 - 3. telefonikõne
 - EB - Erki Berends
 - MM - helistaja, 50a mees
 - Litereeritud Liina Lindström 18. 02.
1999, 21.03.2000
-

□ EB: (---) `joviaalsevõitu=hh (.) laulukene (.) tuli
mei:l: väikeste lõõtspillide=h (.) `ühingult=ja=h (.)
noh `armastuse teema oli=sin `kindlasti `mängus
aga (.) selge (.) `sina (.) e vormis: (.) ja kuigi meile
helistas ka üks `daam kes=ee Jüri Aarma (.) poolt
välja öeldule pisut `vastu vaidles (.) teades:
`kindlasti ühte sellist kena eestikeelset
`armastus` laulukest (.) kus on kasutatud ka `teie
vormi. (.) nii=et aitäh `tähelepaneku eest. (.) .hh me
`räägime aga nüüd `järgmise (.) mm Kuku raadio (.)
stuudio (.) .hh `külalisega läbi `eetri=hh, (.) läbi
`telefonieetri, (.) ja:(.) m tere õhtust `teilegi.

□ (.)

-
- MM: e tere õhtust.
 - EB: oleme `teie või `sina peal.
 - (.)
 - MM: mmh (.) no=ma arvan=et `teie peal.
 - (.)
 - EB: no oleme [`teie peal.]
 - MM: [ma=olen] n:oh natuke niuke
`vanem=tead (.) `viiskend `vana.
 - (.)
 - EB: siis=[on]=se kõigiti põhjendatud.
 - MM: [(mh)]
-

-
- MM: et (.) et tundub niimodi=et `noortel `inimestel on
`tõesti nagu veidi `lihtsam see `sina peale `minek,
(.) et (.) et mul=ühe teise `asutuse (.) `noorte
`daamidega on `töösuhe. (.) ja nemad on hakanud
mind `kokku `leppeta noh (.) ütleks (.) `häbematult
`sinatama. (.) ja: ja mul ei=ole selle vastu `midagi,
mul=on `hia `meel. (.) ja mina ei: `julge neid `nii (.)
õ kohe vastu `sinatada=s ma katsun `mõeld- (.)
`mööda `minna=õ `sõnastusega. (.) ütlen=et mul
oleks `vaja või=et (.) et=ma tülitan `jälle=ja (.)
kuidagi `niimodi, (.) noh=et `vältida seda: noh sina
`ütlemist. (.) kuna meie `meid pole `mitte `keegi
isegi `tutvustanud. (.) ja=ma=i suuda nende
`nimesid `meelde `jätta aga=nad kõik on `nii
`sõbralikud. (.) ni=et et noh (.) `ilmselt on see:: noh
veidi `vanuse `vahe ka=see=asi. (.) [mis `mõjutab.]
-

Taustakirjalduse komponendid (lühendatud põhiskeem)

- ❑ **0. Tehniline info**
 - ❑ **1 Situatsioon ja olukord**
 - ❑ 1.1 Aeg ja koht
 - ❑ 1.4 Osalejate asetus ruumis
 - ❑ 1.6 Situatsiooni kultuuriline määratlus
 - ❑ 1.11 Situatsioonis suhtlust häirivad või seda positiivselt mõjutavad situatsioonivälised faktorid
 - ❑ **2 Suhtlejad, nende omadused ja omavahelised suhted**
 - ❑ 2.1 Konkreetsed suhtlejad
 - ❑ 2.2 Suhtlejate sotsiaalbioloogilised omadused
 - ❑ 2.5 Suhtlejate selle hetke omadused
 - ❑ 2.7 Suhtlejate omavahelised suhted üldse ja konkreetses situatsioonis
-

-
- **3 Ainestik ja teema**
 - 3.1 Konkreetne teema või teemad.
 - 3.5. Teemasündmuse seotus suhtlemisolukorraga ja ümbrusega
 - **4. Tekst ja suhtlus**
 - 4.1. Teksti ja suhtluse liik: dialoog/monoloog/polüloog
 - 4.2. Tekstiosa/teemaosa retooriline tüüp
 - 4.3. Teksti planeerituse aste
 - 4.4. Teksti fikseeritus
 - **5. Keel ja keelekasutus**
 - 5.5. Võti: energilisus, toon
 - **6. Lisa**
 - Vabas vormis ülevaade tekstis ja situatsioonis silma torganud huvitavate joonte kohta.
-

Korpuse kasutamine

- vajalik kõigi projektide jaoks, mis analüüsivad ja modelleerivad suulist keelt ja selle kasutust
 - kasutatav tegeliku eesti suhtluskeelega analüüsimiseks
 - morfoloogia, süntaks, leksika, semantika ja pragmaatika
 - eesti kõnekeel erineb süntaksi, leksika, semantika ja kasutuse poolest oluliselt kirjakeelest
 - lindistuste kvaliteet ei luba korpust kasutada foneetiliseks uurimiseks
-

-
- telefonipõhised infosüsteemid
 - suulise teksti refereerimise ja sisukokkuvõtete programmid
 - interaktiivsed keeleõpperogrammid, mille abil õpetatakse tegelikku kõnekeelt.
 - suulise keele erisõnastikud
 - kõnepuudega inimeste korpus võimaldab uurida seda kõnet ja on abiks suhtluspuude leevendamise vahendite väljatöötamisel
-

Projekti tegevus

- Korpuse üldküsimumuste lahendamine
 - Korpuse kasutamise juriidilised probleemid
 - Korpuse kogumine ja translitereerimine
 - Korpuse digitaliseerimine
 - Korpuse kasutamise võimaldamine
 - Korpuse tutvustamine konverentsidel
-

Täitjad ja koostöö

- projekti juht Tiit Hennoste
 - doktorandid Olga Gerassimenko, Riina Kasterpalu, Andriela Rääbis, Krista Strandson
= suulise keele ja suhtluse uurimise mitteformaalne tööühm
 - kaksikprojekt Mare Koidu projektiga Eestikeelne infodialoog arvutiga (2006-2008)
 - koostöö: Konversatsiooniangendi modelleerimine: eestikeelse dialoogi automaattöötuse teoreetilised ja rakenduslikud probleemid (2004-2007, Mare Koit, ETF)
-

Korpuse üldküsimumuste lahendamine

- ❑ mahtude ja struktuuri, translitereerimisjuhendi täiendamine
 - ❑ korpuse kasutamise materjalide võrgus kättesaadavaks tegemine
 - ❑ taustakirjelduste automaatanalüüsile üleviimine (taustakirjelduse korrastamine, automaatanalüüsi programm, juhendid)
 - ❑ eestikeelsed ja eesti materjalile kohandatud juhendid: CLAN (litereerimine)
 - ❑ Transana (videolindistuste litereerimise programm)
 - ❑ translitereerijate koolitamine (kursused)
-

Korpuse kasutamise juriidiliste ja eetiliste probleemide lahendamine

- juriidilised ja eetilised probleemid kasutamisel ja materjalile viitamisel
 - valdav osa lindistusi ei ole loomu poolest avalikud tekstid ega internetis kasutatavad
 - Korpus jaguneb eri piirangutasemetega alaosadeks
 - kitsaim: kasutatavad vaid oma uurijarühma piires (arstivestlused, telefonimüügivestlused)
 - kõik soovijad peavad sõlmima lepingu korpuse kasutamise ja konfidentsiaalsuse kohta
 - ülikooli jurist on heaks kiitnud NÄIDE
-

Korpuse täiendamine

- audio- ja videomaterjali lindistamine
 - Tekstide lindistamisel kaks strateegiat
 - üksikud lindistajad ja litereerijad: odavam, kitsam materjal, aeglane
 - sotsioloogilise uurimise asutus teeb katva materjalikogumise, palgatud litereerijad teevad pealiskaudse litereeringu, asjatundjad täpsustavad: kiire ja laiapõhjalist materjali andev, väga kallis
-

-
- lindistamine: aastas lisandub ca 100 000 üksust
 - translitereerimine
 - taustakirjelduste koostamine
 - varasemate translitereeringute kontrollimine ja täpsustamine
 - osa uusi lindistusi videos viimasel aastal (klassidialoogid)
-

Suulise keele korpus oktoober 2007

(<http://www.cl.ut.ee/suuline/>)

- 686 audiolinti
 - 23 videolinti

 - 1777 translitereeritud teksti
 - 1 171 800 tekstiüksust (sõna ja pausi)

 - Silmast-silma vestlused 29,3% (521)
 - argivestlused 31,5% (164)
 - institutsionaalsed vestlused 68,5% (324)
(kauplus, teenindus, arst, intervjuu, jutlus, loeng, koolitund, konverentsiettekanne, reisibüroo jm)
-

-
- Telefonivestlused 62,8% (1116)
 - argivestlused 14% (159)
 - institutsionaalsed vestlused 86% (945)
(infotelefon, müügipakkumine, reisibüroo,
polikliiniku registratuur, taksotellimine jms)
 - Meediasaated 7,8% (140)
-

Dialogikorpus (EDiC)

- <http://www.cs.ut.ee/~koit/Dialog/EDiC>
 - 1061 translitereeritud teksti
 - 178 100 tekstisõna
 - 945 telefonikõnet
 - 116 silmast-silma vestlust
-

Digitaliseerimine

- ❑ Korpuse vanemad osad (lindistatud enne 2004. aastat) on analoogkujul
 - ❑ üleviimine digitaalkujule
-

Korpuse kasutamise võimaldamine

- administreerimine
 - korpuse pidev korrashoidmine
 - lepingute sõlmimine kasutajatega
 - suhtlemine kogujate ja kasutajatega
-

Korpuse tutvustamine

- Osalemine konverentsidel ja workshoppidel, et tutvustada oma korpust, selle ehituspõhimõtteid ja selle põhjal tehtavat tööd uurimistööd
 - SIGdial 2004, LREC 2004
 - DISS'05, TSD 2005, SPECOM 2005
 - FinTAL 2006, TSD 2006, SPECOM 2006, International Conference of Conversation Analysis Helsinki 2006,
 - MegaLing 2007, RANLP 2007, IPRA International Pragmatic Conference, Göteborg 2007
 - Eesti rakenduslingvistika konverentsid, Balti keeletehnoloogia konverentsid jm
 - Artiklid konverentsikogumikes ja mujal
-

Korpusega seotud doktoritööd

- Tiit Hennoste, Eesti suulise keele korpus ja kõnekeele eneseaparandussüsteemi analüüs (Helsingi ülikool)
 - Andriela Rääbis, Telefonivestluse struktuur (Tartu ülikool)
 - Riina Kasterpalu, Eelkõneleja suhtluskontuuri järgimine ja murdmine partikliga jah/jaa eestikeelses suhtluses (Tartu ülikool)
 - Olga Gerassimenko, Tagasisidevahendid eesti ja vene telefonikõnedes (Tartu ülikool)
 - Krista Strandson, Klassisuhtluse jooni. Õpetaja ja õpilase kõne algkooli tunnis (Tartu ülikool)
-

Korpuseartiklid

- Hennoste, Tiit. Eesti suulise kõne uurimine: transkriptsioon, taust ja korpus. - Keel ja Kirjandus 2000, nr 2, lk 91-106.
 - Hennoste, Tiit. Suulise eesti keele uurimine: korpus. - Keel ja Kirjandus 2003, nr 7, lk 481-500.
 - Hennoste, Tiit, Mare Koit, Maret Kullasaar, Andriela Rääbis, Evely Vutt. Eesti dialoogikorpuse loomise probleemid. - Täheendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Renate Pajusalu, Tiit Hennoste. Tartu 2002, lk 143-160.
-

-
- Hennoste, Tiit, Liina Lindström, Olga Gerassimenko, Airi Jansons, Andriela Rääbis, Krista Strandson, Piret Toomet, Riina Vellerind. Suuline kõne ja morfoloogiaanalüsaator. - Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Renate Pajusalu, Tiit Hennoste. Tartu 2002, lk 161-171
 - Hennoste, Tiit. Suuline keel, dialoog ja arvuti. Sissejuhatuseks. – Keel ja arvuti. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 6. Toim Mare Koit, Renate Pajusalu, Haldur Õim. Tartu 2006, lk 126-142
-