EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 111–130

SUULISE EESTI KEELE KORPUS JA INIMESE SUHTLUS ARVUTIGA

Tiit Hennoste, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson

Ülevaade. Tartu Ülikoolis kogutakse suulise eesti keele korpust ja (selle alamosana) dialoogikorpust, et uurida inimestevahelist suhtlust. Kaugem eesmärk on luua kasutajaliideseid, mis võimaldaksid inimkõne vahendusel suhelda elektrooniliste andmebaasidega. Suhtluse modelleerimine on edukam, kui selle aluseks võetud inimsuhtluse uurimine tehakse suurel eri allkeeli sisaldaval korpusel ning viiakse läbi nii kvantitatiivne kui ka kvalitatiivne analüüs. Artikkel tutvustab Tartu Ülikooli suulise eesti keele korpuse ehitust, transkribeerimise põhimõtteid ja dialoogiaktide annoteerimiseks kasutatavat tüpoloogiat. Rakendusena vaadeldakse ametlikes telefonikõnedes esinevate päringute keelevaliku seletusi.*

Võtmesõnad: suulise keele korpus, dialoogikorpus, transkriptsioon, dialoogiaktid, märgendamine, suuline suhtlus, eesti keel

1. Sissejuhatus

Tänapäeval muutuvad järjest populaarsemaks nn intelligentsed kasutajaliidesed, mis võimaldavad inimkõne vahendusel suhtlemist elektrooniliste andmebaasidega. Väga lihtsa suhtluse modelleerimine (nt telefoninumbrite küsimine) ei vaja kuigivõrd tegeliku suulise keele ja selle kasutuse analüüsi. Kui aga soovitakse luua arvutiprogrammi, mis suudaks dialoogis osaleda inimesele võrdväärse partnerina, siis ei ole see ilma inimeste suhtlust analüüsimata võimalik.

Suulise keele analüüs nõuab korpust. Enamasti on kasutajaliideste loomiseks kasutatud piiratud korpusi, mis sisaldavad suhtlust kindlate ülesannete raames ja mis on tihti kogutud rollimängude abil. Näiteks korpus COCONUT[1] koosneb inimestevahelistest arvuti kaudu vahendatud dialoogidest, milles osalejad teevad koostööd

* Tööd toetavad Eesti Teadusfond (grant nr 7503) ning Haridus- ja Teadusministeerium (sihtfinantseeritav teema "Loomulike keelte arvutitöötluse formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine" ja riiklik programm "Eesti keele keeletehnoloogiline tugi"). Täname anonüümseid retsensente.

[1] http://www.pitt.edu/~coconut/ (29.09.2008).

kodu möbleerimisel. Korpus VERBMOBIL[2] sisaldab kakskeelseid dialooge, mis on salvestatud rollimängudes, kus kooskõlastatakse kokkusaamisi, broneeritakse tuba hotellis, koostatakse sõiduplaani jne.

Meie kaugem eesmärk on modelleerida arvutil suhteliselt keerukaid dialooge võimalikult loomulikult, s.t arvuti kui dialoogis osaleja peab järgima inimestevahelise suhtluse põhimõtteid, nii palju kui see on võimalik ja vajalik (Koit 2007). Selleks vajame inimsuhtluse analüüsiks piisavalt suurt dialoogikorpust, kus osalejate rollid on lähedased dialoogsüsteemi ja tema kasutaja rollidele.

Igasugune keelekasutus varieerub. Varieerumise mõjutajad võivad olla inimesekesksed (haridus, elukoht jms), situatiivsed (argine/avalik, dialoog/monoloog jms) ja interaktsioonilised (eri grammatilised konstruktsioonid on seotud kindlate suhtlusfunktsioonidega). Mõned keelelised variandid on sagedased, teised haruldased. On tavaline, et haruldased variandid jäetakse suhtluse modelleerimisel kõrvale kui ebaolulised. Suhtluse mikroanalüüs aga näitab, et ka neil on kindel kasutusala ning neid ei saa kasutajaliidest tehes ignoreerida. Seega on suhtluse uurimine ja modelleerimine edukamad, kui on olemas suur korpus, mis sisaldab erinevaid allkeeli.

Käesolev artikkel annab kokkuvõtliku ülevaate Tartu Ülikooli suulise eesti keele korpusest, keskendudes selle ühele allkorpusele – dialoogikorpusele.[3] Tutvustatakse dialoogikorpuse märgendamisel kasutatavat dialoogiaktide tüpoloogiat ja esitatakse üks näide märgendatud dialoogikorpuse võimaliku kasutuse kohta suhtluse modelleerimisel.

2. Tartu Ülikooli suulise eesti keele korpus

Suulise keele korpust[4] on Tartu Ülikooli (TÜ) suulise kõne uurimisrühm kogunud alates 1997. aastast, kaasates selleks ka suulise kõne ja eesti keele allkeelte kursustel osalevaid üliõpilasi.[5] Korpuse koostamise esimene küsimus on tekstiliikide valik. TÜ korpus on planeeritud avatud korpusena, s.t ei ole ette kindlaks määratud, kui palju ja mis liiki tekste ta peaks sisaldama. Korpus peab hõlmama suulise kõne erinevaid allkeeli, et tema abil saaks uurida kogu suulisele kõnele ühiseid jooni, eri allkeelte erijooni ning ka võrrelda suulist ja kirjalikku keelekasutust. Samuti peab uurimusi saama teha kõigi keeletasandite kohta (leksika, fonoloogia, morfoloogia, süntaks, semantika, pragmaatika). Iga uurija, kes vajab tasakaalustatud korpust (nt sellist, milles on võrdselt kindlat tüüpi situatsioone), saab selle ise suure korpuse alusel koostada.

2.1. Korpuse liigendus

Korpus on liigendatud kolme keelekasutust mõjutava parameetrirühma alusel (Hennoste 2003). Esimese annavad suhtlejate sotsiaalne ja dialektiline taust. Teise moodustavad suhtluse omadused, mis annavad kokku registrid: dialoog või monoloog, kõne spontaansuse aste, kontakti iseloom (silmast-silma, telefonisuht-

[2] http://verbmobil.dfki.de/ (29.09.2008).

[3] Korpuse põhialustest ja varasemast seisust on antud ülevaade artiklites Hennoste 2000, 2003, Hennoste jt 2000. Käesolevas kordame osalt põhifakte ja toome välja korpuse praeguse seisu ning muudatused.

[4] Varem oleme kasutanud segamini suulise keele ja suulise kõne korpuse nime. Praeguseks oleme sõnast *kõnekorpus* loobunud, kuna kõnekorpused (ingl *speech corpora*) märgivad arvutilingvistikas foneetiliseks kasutuseks mõeldud spetsiaalseid korpusi. Meie korpus on keelekorpus (*linguistic corpus*), mis kitsamalt piiritletuna võiks kanda ka pragmaatilise korpuse nime.

[5] http://www.cl.ut.ee/suuline/ (29.09.2008).

lus, meediasuhtlus), suhtluse argisus/institutsionaalsus.[6] Viimase puhul teeme omakorda vahet nelja alammõõtme vahel: osalejatevahelised suhted (tuttavad või võõrad), osalejate rollid suhtluses (eraisik või ametiasutuse esindaja), suhtluse ruum (eraruum või ametiruum), suhtluse põhieesmärk (osalemine või info vahetamine). Kolmanda rühma annavad erinevad suhtlusvaldkonnad, mis on seotud kindlate situatsioonitüüpidega (teenindus, lobisemine jms).

Korpus koosneb kahest poolest: salvestused ja transkribeeritud materjal. Korpuses on valdavalt audiosalvestused, videosalvestusi on vähe (telesaated, lastekeel, koolitunnid). Materjal salvestatakse digitaalselt (vanem osa on audiokassettidel, mida pidevalt digitaliseeritakse) foneetilist analüüsi võimaldavasse wav-heliformaati. Materjal on litereeritud Wordis, transkribeeringud on olemas doc-failidena ja txt-failidena. Transkribeerimisel kasutame vabavaralisi abiprogramme VoiceWalker (helifaili sees liikumiseks) ja Praat või CLAN (helimaterjali täpsemaks analüüsiks).[7]

Korpuse tegelikud kasutuspiirid määrab põhjalikult transkribeeritud tekstide hulk. Põhjaliku transkriptsiooni all mõistame sellist taset, kus on üles märgitud suurem osa olulisi parameetreid, mis aitavad keele ja selle kasutuse analüüsi interpreteerida. Seisuga 30. september 2008 on TÜ korpuses 2011 põhjalikult transkribeeritud teksti, kokku 1 333 300 tekstiüksust (sõna, üneemi ja pausi).

Tüüpilised transkribeeritud argivestluste lõigud ja pikemad institutsionaalsed dialoogid on viie kuni viieteistkümne minuti pikkused. Lühemad institutsionaalsed dialoogid on transkribeeritud täielikult. Korpus jaguneb telefonikõnedeks, silmast-silma vestlusteks ning meediasuhtluseks (tabel 1).

Peaaegu kõik korpuse salvestused on dialoogid. Monoloogid on nt ettekanded, loengud, jutlused. Tekstide hulga poolest on rohkem institutsionaalset suhtlust, aga kuna argisuhtluse transkribeeringud on oluliselt pikemad, siis sõnade arvu järgi on rohkem argisuhtlust.

2.2. Transkriptsioon

Helikandjatele salvestatud korpus tuleb uurimiseks transkribeerida. Ükski transkriptsioonisüsteem ei kajasta kõiki kõnes esinevaid lingvistilisi nähtusi, sest see muudaks transkribeerimise ülimahukaks tööks. Meil kasutatav transkriptsioonisüsteem (vt lisa) pärineb vestlusanalüüsist ja on olemuselt pragmaatiline (vrd Jefferson 2004). Kuna vestlusanalüüs keskendub suhtluse arenemisele ja kujunemisele kõneluse käigus vestluskaaslaste koostöö tulemusena, märgitakse täpsemalt suhtlusnähtusi ja näiteks foneetikale pööratakse vähem tähelepanu. Transkriptsioonis tuuakse välja 7 nähtuste rühma (vt Hennoste 2000: 98–100):

- suhtlusüksused (vooruehitusüksused, meie terminoloogias lausungid), mis lõpevad potentsiaalsetes voorusiirdekohtades ja mille keskseks piiritlejaks on intonatsioon (Hennoste, Rääbis 2004: 27–30),
- sõnad ja suhtlushäälitsused (*ee*, *mhmh*). Sõnad kirjutatakse vastavalt häälduse, kuid tavalises ortograafias,
- mõõdetud pikkusega pausid,
- kõne prosoodilised ja paralingvistilised omadused (intonatsioon, venitused, katkestused, rõhud jne),

[6] Varem oleme kasutanud ka väljendit *avalik suhtlus* institutsionaalse suhtluse asemel, vt Hennoste 2003: 492.
[7] http://www.linguistics.ucsb.edu/projects/transcription/tools.html; http://www.fon.hum.uva.nl/praat/; http://childes.psy.cmu.edu/clan/ (29.09.2008).

- üksteisele peale- ja otsarääkimised,
- transkribeerija kahtlused (halvasti kuuldud sõnad jne),
- nähtuste kirjeldused, mille kohta puudub märk või mille transkribeerimist ei peeta analüüsi seisukohalt vajalikuks (kõrvalised hääled, nutt vms).

Tabel 1. TÜ suulise eesti keele korpuse koosseis

Salvestuse liik	**Transkribeeritud tekstide arv**				
Telefonisuhtlus	1297	argisuhtlus	176		
		institutsionaalne suhtlus	1121	infotelefon	555
				reisibüroo	93
				polikliiniku registratuur	99
				teenindus	80
				müügivestlused	48
				kolleegidevahelised vestlused	40
				kauplus	28
				takso tellimine	23
				bussiinfo	20
				ülikooliinfo	18
				küsitlus	16
				raamatukogu	10
				muud vestlused	91
Silmast-silma suhtlus	562	argisuhtlus	184		
		institutsionaalne suhtlus	378	kauplus	101
				ettekanded ja loengud	50
				teenindus	30
				intervjuud	25
				teeküsimine tänaval	20
				arst ja patsient	17
				reisibüroo	15
				koolitunnid	12
				koosolekud	11
				muud vestlused	98
Meedia	152	raadio	90		
		televisioon	62		
Kokku	2011				

Meie praegu kasutatav transkriptsioonivariant erineb traditsioonilisest vestlusanalüüsi transkriptsioonist eeskätt mõne märgi poolest, põhjuseks on asjaolu, et materjal peab olema arvuti poolt loetav. Nii kasutame allajoonimise asemel rõhu märkimiseks graavist (`) ja halvasti kuuldud lõigud paigutame loogelistesse sulgudesse (vrd varasemat kasutust Hennoste 2000: 99). Transkribeeritud teksti esitab näide (1).

(1) H: okei=okei suva.
(.)
T: vot=ja:=ah (.) mina=ei=tea (.) igasugused `bioloogid olid (.) põhiliselt.

(1.0)
H: ahah
(0.8)
T: [> a `mina=ei=tea <]
H: [{ahah. kus} te] `magasite seal.
(.)
T: mh, `mattide `peal. (1.0) > ja mingid õudselt head < `dušširuumid ja=värgid olid kõik `kasutada * meil. *
H: kus=kohas te `olite * mis `majas. * ((sööb samal ajal))
T: koolimajas.
H: mm.
(0.5)
T: ja=se=on `Kärdla ainu- või=tähendab=see (.) `Hiiumaa ainus `keskkool.
(0.8)
H: mhmh.

2.3. Taustakirjeldus

Oluline osa korpuse dokumentatsioonist on iga teksti taustakirjeldus, mis lubab uurida suhtlejate, suhtlussituatsiooni ja keele seoseid. Meie maksimaalses taustakirjelduse mudelis on välja toodud 44 situatsioonifaktorit, mis on leitud mõjutavat keelekasutust (Hennoste 2000: 100–105). Praktiliselt kasutame kirjelduse lühiversiooni, milles on 23 tegurit. Taustakirjelduse põhiosad on järgmised.

0. Tehniline info salvestamise ja litereerimise kohta.
1. Situatsioon (aeg, koht, suhtlussfäär, suhtlusnormid jms).
2. Suhtlejad, nende omadused ja omavahelised suhted (nimi, vanus, sugu, haridus, kodukant, sotsiaalne staatus jne).
3. Ainestik ja teema.
4. Tekst ja suhtlus (dialoog/monoloog/polüloog, teksti planeerituse ja fikseerituse aste jm).
5. Keel ja keelekasutus (dialekt, register jms).
6. Lisainfo.

Osa punkte on esitatud loendina, milles täitjal tuleb valitud variant alla joonida. Osa punkte on avatud, neisse tuleb info lisada. Näide (2) esitab väljavõtte näite (1) taustakirjeldusest.

(2) /---/
1. Situatsioon ja olukord
1.1. Aeg ja koht
- päev, kuu, aasta: **15. jaan. 1997 kell 18–18.30**
vahetu suhtlus
- koht (**linn**, maakond, vald, küla, talu): **Tartu**
- **kodu** (eramaja, korter) / ametiasutus (kontor, kauplus jne): **ühiselamutuba**
/---/

1.4. Osalejate asetus ruumis (**istuvad** / seisavad; laua, toolide jm esemete kasutamine – kirjeldada): **istuvad laua ümber**
suhtlusdistants (alla poole meetri, **pool kuni poolteist**, pikem):
/---/
1.6. Situatsiooni kultuuriline määratlus:
- vestluse põhitüüp (**argisituatsioon** / avalik situatsioon, **eravestlus** / ametialane vestlus):
- nimetus võimalikult täpselt: **söömine, argivestlus. Tiina on külas oma sõbrannal Heleril, vestlusringis osaleb ka Heleri toakaaslane Lea**.
/---/
1.9. Situatsiooni suhe suhtlejatega
- **tuntud reeglitega** / võõraste reeglitega:
- **mugav** / ebamugav:
- esmakordne / **mitmes kord**:
- ootamatu / **kavandatud** / kokku lepitud:
/---/
1.11. Situatsioonis suhtlust häirivad või seda positiivselt mõjutavad situatsioonivälised faktorid (telefonikõne, võõra tulek, toidu toomine, kohvi pakkumine, teadmine, et tuleb lindistada jne): **söömine, toidu pakkumine**
2. Suhtlejad, nende omadused ja omavahelised suhted
2.1. Konkreetsed suhtlejad
- nimed: **Tiina, Heleri, Lea**
- suhtlejate hulk (kaks inimest / **väike rühm kuni kümmekond inimest** / suur rühm):
- rollid: suhtlejad: **Tiina, Heleri, Lea**
juuresolijad: **ei ole**
- konkreetsed sotsiaalsed rollid:
2.2. Suhtlejate sotsiaalbioloogilised omadused
2.2.1.
- nimi (roll): **Tiina**
- sugu: **naine**
- vanus või sünniaeg: **21-a**
- haridus (alg / põhi / **kesk** / kõrg):
- rahvus / hõim: **eestlane**
- kodukant / lapsepõlvekodu: **pärit Tallinnast**
- sotsiaalne staatus (tööline / talupoeg / teenindaja / intelligent / ärimees / pensionär / teenistuja / kodune / töötu / ärijuht / ametnik / keskastmejuht / kõrgema astme juht / müügiinimene / õpilane / **üliõpilane**):
- kõnet mõjutavad füüsilised puuded või väljapaistvad omadused (kõneanne *etc.*): **ei ole**
/---/
2.7. Suhtlejate omavahelised suhted üldse ja konkreetses situatsioonis
- võõras / tuttav / **lähedane** (kirjeldada): **sõbrannad**
- staatussuhted (**võrdne** / alluv / kõrgem; – lisada, kumb on kumb):
/---/

4. Tekst ja suhtlus
/---/
4.3. Teksti planeerituse aste
- varem / samal ajal: **pole planeeritud**
- planeerimissügavus (teema / eesmärgid / struktuur / märksõnad / sõnavara / süntaks / intonatsioon):
4.4. Teksti fikseeritus: **ei ole**
- paberil / peas:
- kogu tekst / osa / **mitte midagi**:
/---/

2.4. Korpuse kasutamine

Suulise keele korpuste kasutamine erineb tüüpiliste kirjaliku keele korpuste kasutamisest. Viimased sisaldavad tekste, mis on juba varem avalikkuses ringelnud. Suulise keele korpused sisaldavad selliseid väga vähe. Argitekstid on originaalis määratud ainult kõnelejatele enestele. Ka suur osa institutsionaalseid tekste on määratud väikesele hulgale kindlate omadustega osalejatele (nt loengud üldjuhul aine üliõpilastele) või sisaldavad tundlikku materjali (nt arsti ja patsiendi vestlused). See toob kaasa vajaduse tekste teatud moel muuta ja nende kasutamist piirata. Meie oleme kasutanud järgmisi piiranguid.

Kõik nimed, telefoninumbrid, aadressid jm identifitseerimist võimaldavad andmed transkriptsioonides asendatakse rütmiliselt võrdväärsete asendajatega (nt *Tiina* > *Liina*). Õigeid andmeid leiab vaid taustakirjeldusest.

Korpus jaguneb eri piirangutasemetega alaosadeks. Osalejate nõusolekul saavad uurijad kasutada salvestusi teadus- ja õppe-eesmärkidel. Osa salvestusi on suhtlejad andnud uurijate isikliku vastutuse alla ja neid kasutatakse vaid suulise keele uurimisrühma piires.

Kõik korpuse kasutajad peavad kirjutama alla konfidentsiaalsuskohustusele ja piirama avalikult esitatavad tsitaadid kõnelejate identifitseerimist mittevõimaldava mahuni.

3. Eesti Dialoogikorpus EDiK

Suulise eesti keele korpuse juurde kuulub dialoogikorpus EDiK.[8] See on koostatud spetsiaalselt inimestevahelise institutsionaalse suhtluse uurimiseks, et selle alusel modelleerida inimese ja arvuti vahelist dialoogi (Hennoste jt 2002a). See jaguneb kolmeks alaosaks:

- suulise eesti keele korpusest valitud 1137 dialoogi, kokku 210000 tekstisõna, mis jagunevad järgmiselt: 1012 telefonikõnet (infotelefon, reisibüroo, bussijaam, polikliiniku registratuur, kauplused, taksodispetšer jt) ja 125 silmast-silma vestlust (kaubandus, teenindus, reisibüroo, teejuhatamine jt);
- võlur Ozi meetodil kogutud kirjalikud reisiinfodialoogid (Valdisoo jt 2003);

[8] http://math.ut.ee/~koit/Dialoog/EDiC.html (29.09.2008).

- inimese ja arvuti vahelised dialoogid, mis on kogutud kahe lihtsa küsimus-vastussüsteemi Reisiagent[9] ja Teatriagent[10] arendamise käigus (vt ka Treumuth jt 2006).

Osa suulistest dialoogidest EDiK-is on analüüsitud ja märgendatud morfoloogiliselt, süntaktiliselt ning dialoogiaktiliselt.

3.1. Morfoloogiline analüüs

Suulise keele korpuse morfoloogiliseks analüüsiks kasutasime analüsaatorit ESTMORF (Kaalep 1997, 1998)[11]. ESTMORF on programm, mis võrdleb jooksvas tekstis leiduvaid sõnesid sõnastikus olevate lekseemide kombinatsioonidega. Töö tulemusena esitab ta iga sõna kohta sõnaliigi nime ja andmed muude grammatiliste kategooriate kohta, mis antud sõnaga seostuvad (arv, kääne, isik jms). ESTMORF on mõeldud eesti kirjakeele jaoks. Et kohandada kirjaliku keele analüsaatorit suulise keele analüüsiks, tegime katse, toomaks välja probleemid, millega algne analüsaator hakkama ei saanud (Hennoste jt 2002b). Katse tulemusena ilmnes vajadus lisada kaks uut märgendit: B partikli ja T tundmatu sõna märkimiseks. Partiklid on suulises keeles kasutatavad sõnad, millel on eeskätt pragmaatiline roll suhtluses (*ahah*, *mhmh*, *noh* jms, vt Hennoste 2002). Teiseks tuli lisada analüsaatorisse rida sõnu, mida kirjalik keel ei tunne (nt argisõnad).

Dialoogikatkend näites (3) on morfoloogiliselt analüüsitud näites (3a) (K – klient, A – ametnik). Korrektne morfoloogiline analüüs on tähistatud miinusmärgiga, nt *tere* on partikkel B, mitte aga substantiiv S.

(3) K: teated tere
A: tere ma sooviksin teada.

(3a) **K**

```
<s>
teated
teade+d //_S_ com pl nom //
tere
- tere+0 //_B_ //
tere+0 //_S_ com sg gen //
tere+0 //_S_ com sg nom //
</s>
```

A

```
<s>
tere
- tere+0 //_B_ //
tere+0 //_S_ com sg gen //
tere+0 //_S_ com sg nom //
ma
mina+0 //_P_ pers ps1 sg nom //
sooviksin
soovi+ksin //_V_ main cond pres ps1 sg ps af //
```


[9] http://www.dialoogid.ee/reisiagent/ (29.09.2008).
[10] http://www.dialoogid.ee/teatriagent/ (29.09.2008).
[11] http://www.cl.ut.ee/korpused/morfliides/; http://www.filosoft.ee (29.09.2008).

teada
tead+a //_V_ main inf //

3.2. Süntaktiline analüüs

Süntaktiliseks analüüsiks on kasutatud kirjaliku eesti keele süntaksianalüsaatorit, mis töötati välja aastatel 1996–2001 Tartu Ülikoolis (Muischnek jt 2000, Roosmaa jt 2003). Analüsaator põhineb kitsenduste grammatikal (ESTKG), kus analüüsi alguses lisatakse igale sõnavormile kõik võimalikud analüüsivariandid ja seejärel hakatakse konteksti mittesobivaid eemaldama. Eemaldamine toimub vastavalt kitsenduste grammatika reeglitele ehk kitsendustele, millest igaüks esitab mõnda spetsiifilist keelereeglilaadset fakti. ESTKG-s on 1118 süntaktiliste märgendite eemaldamise reeglit. Süntaktilise analüüsi protsess on selles jaotatud kaheks osaks. Morfoloogiline ühestaja tegeleb kontekstiinfo põhjal morfoloogiliselt mitmese analüüsiga sõnavormile õige morfoloogilise kirjelduse väljavalimisega, süntaksianalüsaator leiab sõnavormi süntaktilise funktsiooni lauses.

Analüsaator kohandati suulise keele analüüsiks (Müürisep jt 2006, Müürisep, Nigol 2008). Selleks tuli lisada uusi reegleid osalausepiiride tuvastamiseks ja muuta mitmeid süntaktilisi kitsendusi. Töö süntaksianalüsaatori arendamisega jätkub.

Näites (4) on süntaktiliselt analüüsitud lausung: *Se veranda on minu meelest maailma kihvtim asi.*

(4) Se #
see+0 //_P_ dem sg nom // **CLB @NN>
veranda #
veranda+0 //_S_ com sg nom // @SUBJ
on #
ole+0 //_V_ main indic pres ps3 sg //@+FMV
minu #
mina+0 //_P_ pers ps1 sg gen // @P>
meelest #
meelest+0 //_K_ post #gen // @ADVL
maailma #
maa_ilm+0 //_S_ com sg gen // @NN>
kihvtim #
kihvti=m+0 //_A_ comp sg nom // @AN>
asi #
asi+0 //_S_ com sg nom // @PRD
$.
. //_Z_ Fst //

3.3. Dialoogiaktide analüüs

Morfoloogiline ja süntaktiline analüüs on olulised ka iseseisvana, kuid suhtluse modelleerimisel on nad üksnes abitegevused, olles vaid vahendid suhtluse tarvis. Suhtlemine ise seisneb selles, et inimesed teevad keele abil erinevaid tegevusi – küsivad, vastavad jne. Selliseid tegevusi nimetatakse suhtlus- ehk dialoogiaktideks.

Suhtluse modelleerimiseks on tarvis luua dialoogiaktide tüpoloogia, analüüsida dialoogid aktideks ja seejärel leida, kuidas on aktid seotud keeleliste üksuste morfoloogiliste ja süntaktiliste omadustega.

Dialoogiaktide praktilise määramise probleeme on käsitlenud viimastel aastakümnetel nii korpuslingvistid, diskursuse analüüsijad kui ka keeletehnoloogid (nt Stolcke jt 2000, Allwood jt 2001, Jokinen jt 2001). Praktiliselt tegeleb dialoogiaktide analüüsiga kogu pragmaatika, kuigi mitte alati dialoogiakti mõistet kasutades.

Dialoogiaktide tüpoloogiaid on loodud maailmas mitmeid, kuid ühist standardit olemas ei ole. Meie tüpoloogia on üldine ja oma põhiosas kooskõlas hästi tuntud tüpoloogiatega (nt DAMSL, SWBD-DAMSL, vt Koit 2003). Mille poolest meie tüpoloogia erineb teistest (vt ülevaadet Hennoste, Rääbis 2004)?

Meie tüpoloogia põhineb vestlusanalüüsi printsiipidel. Vestlusanalüüs on vestlusandmete empiiriline, induktiivne mikroanalüüs (Hutchby, Wooffitt 1998, Kasterpalu, Gerassimenko 2006). Selle aluseks on idee, et vestlus on osalejate koostöö, mis põhineb kolmel mehhanismil: vooruvahetus (vooru ehitamine ja vooru jaotamine), voorujärjestus (eelistused ja naaberpaarid) ja parandus. Samas, kuigi vestlusanalüüs tegeleb kõnelejate tegevusega, ei ehita ta põhimõtteliselt dialoogiaktide tüpoloogiaid.

Meie tüpoloogia on empiiriline ja avatud. Me eeldame, et dialoogiakt on empiiriline nähtus ja võimatu on teoreetiliselt ette määrata kõikvõimalikke akte. Seetõttu sisaldab iga aktiklass alamklassi "Muu". Sinna paigutatakse märgendamisel need aktid, mida tüpoloogias ei ole (veel) määratletud või mis on tehnilistel põhjustel mitteanalüüsitavad. Vajaduse korral defineeritakse selle alamklassi baasil uusi aktirühmi.

Meie aktitüpoloogia põhialused on järgmised (Hennoste, Rääbis 2004: 15–37).

1. Tüüpiline aktide analüüs lähtub sellest, et kõnelejal on olemas plaan või strateegia, mida ta soovib läbi viia, ning ta valib oma aktid vastavalt sellele. Vestlusanalüüs eeldab, et kõneleja kohandab oma suhtlusvoore jooksvalt vestluspartneri eelnevate vorudega. Iga voor ennustab mingil määral, milline jätk tuleb tema järel, ja on ise sobitatud eelneva vooruga. Seetõttu on keskne kahe järjestikuse vooru aktide omavaheliste suhete analüüs. Siin jagatakse aktid kahte rühma. Mõned nõuavad enda järel kindlat tüüpi akti kindlas positsioonis, ideaalis järgnevas voorus. Selliseid aktipaare nimetatakse naaberpaarideks (ingl *adjacency pairs*, nt tervitus–vastutervitus, küsimus–vastus; Schegloff, Sacks 1973). Oodatud akti puudumine või mõne muu dialoogiaktiga reageerimine on tajutav ebaootuspärasena (nt vastusest põiklemine või vastamine küsimusele küsimusega). Teine osa akte on sellised, mille omavahelised suhted on vabamad. Suhtluspartner (ka arvuti) peab suutma vahet teha naaberpaariakti ja üksikakti vahel. Sellest lähtudes jagame aktid naaberpaari- ja üksikaktideks. Igal naaberpaaril on esi- ja järelliige (vrd DAMSL-i edasi- ja tagasivaatav funktsioon).
2. Kõneleja võib reageerida eelnevale aktile ootuspäraselt või mitteootuspäraselt. Viimane tekitab suhtluses probleemi. Kuna probleemid on suhtluses pidevad, siis peavad keeles leiduma vahendid, mis neid signaliseerivad ja lahendada aitavad. Kõik aktitüpoloogiad sisaldavad probleemide lahendamise akte, aga tüüpiliselt on need paigutatud laiali erinevate dialoogi juhtimise ja tagasisideaktide alla ega moodusta terviklikku süsteemi (nt

Bunt 1999). Näiteks DAMSL-is esindab üks akt (ingl *Abandoned* – loovutatud) suhtlusstaatust, aga enamik parandusakte kuulub tagasivaatavate funktsioonide klassi (*Signal-non-understanding* – mittemõistmise signaal, *Completion* – viimistlemine, *Correct-misspeaking* – valesti öeldu korrigeerimine, *Repeat-rephrase* – kordamine-ümbersõnastamine).

Vestlusanalüüs lähtub sellest, et on olemas omaette probleemide lahendamise mehhanism (parandusmehhanism). Emanuel A. Schegloff (1979) toob välja nelja liiki parandusi: enese algatatud eneseparandus (ingl *self-initiated self-repair*), partneri algatatud partneriparandus (*other-initiated other-repair*), partneri algatatud eneseparandus (*other-initiated self-repair*) ja enese algatatud partneriparandus (*self-initiated other-repair*).

Parandusaktide eristamine on oluline ka seetõttu, et paljudel juhtudel kasutatakse infoaktide ja parandusaktide ehitamiseks samu vahendeid (nt suurem osa partneri poolt parandusi algatavaid akte on küsimused). Ka arvuti peab aru saama, millal on tegu infoküsimusega ja millal parandusalgatusega. Eelnevast lähtudes oleme toonud välja omaette rühmana suhtlusprobleemide lahendamise aktid ehk parandusaktid.

3. Dialoogis kasutatavad aktid jagatakse traditsiooniliselt kaheks: infoaktid (nt küsimused) ja dialoogi juhtimise aktid (tagasiside). Vestlusanalüüs lähtub sellest, et sellist jaotust pole olemas. Iga akt annab mingil kombel infot ja iga akt juhib suhtlust (küsimus ei ole pelgalt infoakt, vaid juhib ka suhtlust, määrates järgneva akti tüübi ja suures osas selle võimaliku sisu). Samas ei ole kõik infoaktid ühesugused. Naaberpaariaktid jagunevad küsimusteks, direktiivideks ja seisukohavõttudeks ja teiselt poolt vastusreaktsioonideks nendele aktidele.

   Üksikaktid jagunevad kahe parameetri järgi. Esiteks, ühed aktid annavad primaarselt infot, teised on vastukajad saadud infole, aga ei ole samal määral kohustuslikud nagu naaberpaaride järelliikmed. Teiseks, ühed aktid on seotud eelneva vooru aktidega, teised aga samas voorus oleva primaarse aktiga. Neist parameetritest lähtudes eristame kolme liiki üksikakte. Primaarsed üksikaktid annavad teavet, võtavad seisukohti jne. Nad kannavad infot ega sõltu samas voorus olevast teisest aktist. Infolisad lisavad uut infot sama vooru eelmisele infoaktile kõneleja enda initsiatiivil. Vabatahtlikud reaktsioonid (traditsiooniliselt tagasiside tuum) on reaktsioonid partneri eelmisele voorule.
4. Lisaks on olemas dialoogi juhtimise aktid, mis juhivad kogu dialoogi, kuigi suunavad ka järgmist vooru (tervitamine jms rituaalid, teemavahetus). Kokkuvõttes on meie tüpoloogias 127 akti, mis koonduvad 12 aktirühma.

I. Naaberpaariaktid

DIALOOGI JUHTIMISE AKTID

1. Rituaalsed aktid (tervitamine, tänamine jne).
2. Teemavahetuse aktid, mida kasutatakse uue (alam)teema alustamiseks.

PARANDUSAKTID

3. Parandused, mida algatavad ja viivad läbi erinevad osalejad.
4. Kontakti kontrolli aktid (nt *kas sa kuuled*, *halloo*).

INFOAKTID
5. Direktiivid ja reaktsioonid (soov, ettepanek, pakkumine jne).
6. Küsimused ja vastused.
7. Seisukohavõtud ja reaktsioonid (väide, arvamus jms).

II. Üksikaktid

DIALOOGI JUHTIMISE AKTID
1. Rituaalsed aktid (kontakteerumine, tutvustamine jms).
PARANDUSAKTID
2. Parandused, mida algatab ja viib läbi sama osaleja (eneseparandused).
INFOAKTID
3. Primaarsed üksikaktid (eelteade, lubadus, referaat jms).
4. Infolisad (täpsustamine, pehmendamine jms).
5. Vabatahtlikud reaktsioonid (jätkaja, info vastuvõtuteade jms).

Igal aktil on kaheosaline nimi, mis koosneb akronüümist ja pärisnimest. Akronüümi kaks esimest tähte annavad rühmanime (nt IL = infolisa, KY = küsimused, RI = rituaalid, DI = direktiivid, VR = vabatahtlikud reaktsioonid). Naaberpaariaktidel on ka kolmas täht, mis osutab, kas tegemist on esi- või järelliikmega (KYE = küsimuse esiliige, KYJ = küsimuse järelliige). Akronüümi järel olev sõna (akti pärisnimi) toob välja akti semantilise/funktsionaalse sisu (KYE: AVATUD; KYE: JUTUSTAV KAS).

Näide (5) esitab märgendatud dialoogi (K – klient, A – ametnik).

(5) K: mt=.hh tere, RIJ: TERVITUS
öelge=palun: `pensioniameti `telefoni (.) .h `number (.) `Tartus. DIE: SOOV
(...)
A: ee `number on `seitseeli=`neli? DIJ: INFO ANDMINE
(0.5)
K: jah?= VR: NEUTRAALNE JÄTKAJA
A: =seitse `neli `kolm `kuus. DIJ: INFO ANDMINE
(2.5)
K: aitäh? RIE: TÄNAN
A: palun RIJ: PALUN

Iga dialoogi märgendavad teineteisest sõltumatult kaks lingvisti ja kolmas ühtlustab märgenduse. Iga lausungi märgendamise aluseks on lausungi mikroanalüüs, mis põhineb vestlusanalüüsil ja suhtluslingvistikal (vt ülevaadet Kasterpalu, Gerassimenko 2006 ja sealseid viiteid).

3.4. Korpuse tarkvara

Dialoogikorpusega töötades on vaja tarkvara, mis aitaks nii korpuse kogujaid kui ka uurijaid, kes korpust kasutavad. Seepärast on arendamisel nn korpuse tööpink, mis võimaldab erinevaid tegevusi (Treumuth 2005).[12] Tarkvara on realiseeritud vabavaralisel platvormil ning kättesaadav veebis (kaitstud parooliga), võimaldades kõigil uurimistöös osalejail valida ja töödelda endale vajalikke alamkorpusi.

[12] http://lepo.it.da.ut.ee/~treumuth/ (29.09.2008).

Tööpink võimaldab teha erinevaid statistikaid ja otsinguid. Põhivõimalused on dialoogide lisamine ja eemaldamine korpusest, transkribeeritud elementide (sõnad, pausid jm) ja aktimärgendite loendamine, dialoogiaktide järgnevuste sagedustabeli koostamine, otsing dialoogis esineva teksti (sõne) või aktimärgendi järgi jne.

Samuti võimaldab tööpink dialooge teisendada ühelt kujult teisele. Siia kuuluvad nt dialoogi paigutamine ajateljele, puhastamine morfoloogilise analüüsi tarvis, viimine XML-kujule. Dialoogikorpuse tööpingis on tekstitoimeti, mis annab kasutajale tagasisidet transkriptsioonis esineda võivate vigade kohta ja võimaldab parandada transkriptsioonide loetavust.

Dialoogiaktide märgendamine on seni toimunud käsitsi, kasutades abivahendina programmi (autor Evely Vutt, täiendanud Maret Valdisoo), mis hõlbustab sobiva akti valikut ja selle paigutamist märgendatavas tekstis vajalikku kohta. Testversioonis on valminud tarkvara (autor Mark Fishel), mis jagab teksti lausungiteks ja märgendab automaatselt dialoogiaktid, kasutades masinõpet (Bayesi liigitajat), pakkudes igale lausungile kuni viis märgendusvarianti (vt ka Fishel 2007). Lingvist saab seejärel nende variantide hulgast (või ülejäänud aktipuust) sobiva(d) valida, samuti vajaduse korral muuta lausungipiire.

4. Mis kasu on korpusest?

Arvuti ja inimese suhtluse üks rakendusi on arvutiprogramm, mis vastab inimese küsimustele ja soovidele (annab telefoninumbreid, vahendab taksotellimusi vms). Niisuguse suhtluse tüüpiline mudel on selline, kus inimene helistab ja esitab oma küsimuse või soovi, arvuti aga peab sellele reageerima. Kõneleja võib põhimõtteliselt esitada oma akti kas küsimuse või direktiivi vormis ning valida ka eri küsimuseliikide vahel.

Dialoogi modelleerimise seisukohast on oluline teha selgeks, mille alusel kõnelejad valivad, kas kasutada direktiivi või küsimust (ning nt kas üld- või eriküsimust). Kui need valikud on sotsiaalselt tingitud, siis saaks arvuti seda teadmist kasutada, kui ta suudab analüüsida suhtluskaaslase sotsiaalseid omadusi. Kui näiteks naised kasutavad ühte ja mehed teist varianti, siis saab arvuti arvesse võtta, kes on tema partner, ja valida oma reaktsioonid vastavalt sellele. Kui valik on määratud situatiivselt, siis peab arvuti suutma eri tüüpi situatsioonides reageerida erinevalt.

Käesoleva artikli jaoks analüüsisime ametiasutustesse helistavate klientide esmaseid päringuid – selliseid soove ja küsimusi, mis on kõneleja helistamise eesmärgiks, s.t pole eelpäringud, nt *kas teil aega on* ega ka vestluse käigus sündinud teisesed (alam)päringud. Valitud alamkorpuses on esindatud kolm situatsioonitüüpi: infotelefon, polikliiniku registratuur ja takso tellimine. Uurimisküsimuseks on, mis määrab, millises vormis inimene oma päringu esitab.

4.1. Direktiivid ja küsimused

Direktiivid ja küsimused on teineteisega tihedalt seotud aktirühmad, nii et mõned tüpoloogiad käsitlevadki neid ühe rühmana. Näiteks on DAMSL-is olemas aktirühm *Info-requests* (infopäringud), kuhu kuuluvad aktid, mis seavad kuulajale kohustuse

anda infot. Meie arvates on selline lähenemisviis liiga üldine. Kui keeles on olemas eri vormid suuresti samade tegevuste jaoks, siis on tõenäoline, et neil on ka erinevad kasutusvõimalused.

Mõned aktitüpoloogiad eristavad direktiive ja küsimusi selle alusel, kas kasutaja vajab infot (*mis kell läheb buss?*) või soovib mõjutada kuulaja mittekommunikatiivseid tegevusi (*too vett!*). Esimest vaadatakse küsimusena ja teist direktiivina. Meie väidame, et dialoogi jätkumise seisukohalt pole oluline, kas kuulaja peab väljaspool dialoogi midagi tegema või ei. Ta peab igal juhul reageerima nii küsimusele kui ka direktiivile, sest mõlemad on naaberpaaride esiliikmed.

Meie eristame direktiive ja küsimusi nende vormi alusel. Küsimused on päringud, millel on spetsiifilised keelelised tunnused (küsisõna, teatud sõnajärg jms), direktiividel selliseid spetsiifilisi tunnuseid ei ole. Küsimusi võib omakorda liigitada oodatava reaktsiooni alusel: 1) infot ootavad avatud küsimused (nt *millal väljub viimane buss?*), 2) alternatiivküsimused (*kas rong saabub esimesele või teisele teele?*), 3) suletud *kas*-küsimused, 4) jutustavad *kas*-küsimused, 5) vastust pakkuvad küsimused. Suletud *kas*-küsimused ootavad vastuseks *jah* või *ei*. Jutustavaid *kas*-küsimusi väljendatakse eesti keeles samade vahenditega nagu suletud *kas*-küsimusi, kuid *jah*-vastuse asemel oodatakse info andmist nagu avatud küsimuste puhul (*kas te saaksite mulle öelda X telefoni?*). Vastust pakkuv küsimus sisaldab küsija oletusi õige või sobiva vastuse suhtes (*pluss maksud, jah?*; vt Hennoste, Rääbis 2004, Hennoste jt 2003, 2004, 2008, Koit jt 2006, 2008, Gerassimenko jt 2007).

4.2. Päringute analüüs

Analüüsitud korpuses esitati esmased päringud keeleliselt neljal erineval viisil: direktiiv, avatud küsimus, jutustav *kas*-küsimus, suletud *kas*-küsimus.

Tabel 2. Ülevaade analüüsitud alamkorpusest

Situatsiooni tüüp	Dialooge kokku	Helistaja esimesed päringud (%)				
		Direktiive	Jutustavaid *kas*-küsimusi	Avatud küsimusi	Suletud *kas*-küsimusi	Muid dialoogiakte
Polikliiniku registratuur	26	50%	31%	4%	4%	11%
Takso tellimine	22	77%	13%	–	5%	5%
Infotelefon	60	62%	17%	21%	–	–
Kokku	108	62%	19%	13%	2%	4%

Tabel 2 näitab, et enamik päringutest on direktiivid. Kui rahulduda sellega, võiksime järeldada, et keskne on direktiiv ja muud on sekundaarsed. Samas peab aga arvuti aru saama ka ebatüüpiliselt vormistatud päringutest. Teiseks, mõnikord on direktiivide ja küsimuste valikut seletatud viisakusega (küsimus on väidetavasti viisakam vorm) (vt nt Brown, Levinson 1987). Meie analüüs ei näidanud otseseid ühemõttelisi seoseid viisakuse ja küsimuse vormis esitatud päringu vahel. Eesti keeles on viisakuse väljendamiseks ka muid vahendeid (tingiv kõneviis, viisakussõnad). Ainult seitse päringut (5 infotelefoni- ja 2 taksokõnet) on ilma nende markeriteta. Kolmandaks, on näha, et direktiive ja küsimusi kasutatakse erinevates situatsioo-

nides erineva sagedusega. See tõstatab probleemi, kas nende kasutus pole seotud eri situatsioonides otsitava info tüübiga. Avatud küsimused erinevad ülejäänutest selle poolest, et nende abil küsitakse küsisõna abil määratletud infot. Avatud küsimused algavad eri küsisõnadega, mille äratundmine ei valmista arvutile raskusi. Suletud *kas*-küsimusi ei kasutata päringute esitamiseks, vaid n-ö eelküsimusteks, tingimuste kaardistamiseks.

Võrdleme omavahel direktiivide (DIE: SOOV) ja jutustavate *kas*-küsimuste (KYE: JUTUSTAV KAS) kasutamist, mida mõlemat kasutatakse samatüübiliste päringute esitamiseks. Meetodiks on vestlusanalüüs.

Takso tellimine on suhtlus, milles helistaja ootab küll vastust (takso tuleb teile), aga keskne on dispetšeri eeldatav tegevus (takso saatmine). Tellimised esitatakse põhiliselt direktiivi abil (näide 6).

(6) ma palun `taksot `Ringtee `kuuskend kaheksa `bee. DIE: SOOV

Jutustavat *kas*-küsimust kasutati analüüsitud materjalis ainult kolmel korral. Kõiki neid ühendab helistaja ebakindlus päringu täidetavuse suhtes, sest see on ebatüüpiline (soovitakse kas kahte autot või ebatüüpilist autot, näide 7).

(7) `on teil `kahte autot `Lossi `kolmteist saata KYE: JUTUSTAV KAS

Võime üldistada, et siin väljendavad helistajad direktiiviga oma õigustatud ootust, et soov täidetakse (vrd inglise keele kohta Curl, Drew 2008).

Helistamised **polikliiniku registratuuri** sisaldavad erinevaid päringuid. Kõige sagedamini soovitakse reserveerida arsti vastuvõtuaega. Ka siin ootab helistaja eeskätt registraatori tegevust. Ka arsti vastuvõtuaja kokkuleppimiseks kasutavad helistajad enamasti direktiivi (näide 8).

(8) ma sooviks doktor `Vaheri juurde `aega. DIE: SOOV

Erinevalt takso tellimisest on siin mõned ettenägematud olukorrad (arstil ei pruugi olla soovitud ajal vastuvõttu jms). Siiski kasutati ka neil juhtudel enamasti direktiive. Küsimusi kasutatakse siis, kui päringu täitmine ei paista olevat garanteeritud (näide 9)

(9) kas `teie=juurde `lapsi saab ka regist`reerida=vel `vana aasta sees=hh. KYE: JUTUSTAV KAS

Lisaks küsivad helistajad ka infot nt soodustuste kohta. Erinevalt eelnevast on see otsene infosaamisele orienteeritud päring. Sellised päringud on korpuses väga harvad ja ka need vormistatakse küsimusena.

Kõned **infotelefonile** erinevad nii takso tellimise kui ka registratuurikõnedest. Esiteks soovib helistaja siin alati saada infot, mitte ei oota tegevust. Teiseks, küsitud info on erinevat tüüpi. Meie alamkorpuses soovitakse enim telefoninumbreid (45 juhul ehk 75%). Vähem küsitakse aadresse, asutuste lahtiolekuaegu, ettevõtete tegevusalasid jms.

Peaaegu kõik telefoninumbrite päringud (37) on vormistatud direktiivina (näide 10). Enamasti on numbrisoov selgelt ja täpselt formuleeritud. Vähestel juhtudel on aga helistaja ebakindel täpse aadressi või mõne muu asjaolu suhtes (nt selles, kas eraisikute numbrid on andmebaasis).

(10) palun `Tallinna `Tõnismäe `hambapolikliinik. DIE: SOOV

Ka siin vormistatakse osa päringuid küsimusega. Mõnikord on tegemist üldise infosooviga, mõnikord küsitakse spetsiifilist infot. Aga üldisi andmeid ja eriinfot päritakse ka direktiivi vormis. Millal helistaja valib küsimuse? Siingi on keskseks määrajaks asjaolu, et helistaja pole kindel oma soovi täitmise võimalikkuses (näide 11).

(11) palun kas teil `on: `Vesseli kaupluse `numbrit `Elvas. KYE: JUTUSTAV KAS

Kokkuvõtteks näeme, et enamasti vormistatakse päring direktiivi abil. Päringu formuleerimise vorm ei sõltu sellest, kas soovitakse infot või tegevust. Küsimuste kasutust ühendab üks joon: peaaegu kõik need on seletatavad helistaja ebakindlusega selle suhtes, kas päringut on võimalik täita. Lisaks aga tuleb välja teine üldistus: sagedaste ja tüüpiliste päringute esitamiseks kasutatakse kõigil juhtudel direktiivi, haruldase päringu jaoks küsimust. Samas võime arvata, et kõneleja on kindel sagedaste ja tüüpiliste asjade soovimises ja ebakindel haruldastel juhtudel.

5. Kokkuvõte

Me uurime ja võrdleme inimestevahelisi dialooge, eesmärgiga luua intelligentseid kasutajaliideseid, mis suudaksid vastata kasutajale niisamuti, nagu seda teeb inimesest ametnik. Me väidame, et kõneliideste loomiseks andmebaasidele on vaja uurida erinevat liiki inimestevahelisi ametikõnesid.

Ühestainsast ametisuhtluse tüübist ei piisa, sest on tarvis teada, missugused nähtused on omased suulisele suhtlusele üldiselt, missuguseid keelelisi vahendeid kasutatakse ainult kindlates vestlustüüpides ja missugused on nendevahelised erinevused. Tuleb uurida suuri korpusi ja erinevaid allkeeli, et selgitada, kuidas ja miks inimesed kasutavad erinevate eesmärkide saavutamiseks erinevaid keelelisi vahendeid. Kindlasti on meil vaja piiratud alamkorpusi kindlateks ülesanneteks või uurimisvaldkondadeks, aga neid saab suure korpuse põhjal hõlpsasti moodustada.

Viidatud kirjandus

Allwood, Jens; Ahlsén, Elisabeth; Björnberg, Maria; Nivre, Joakim 2001. Social activity and communication act-related coding. – J. Allwood (Ed.). Dialog Coding – Function and Grammar. Cothenburg Papers in Theoretical Linguistics 85. Göteborg Coding Schemas. Göteburg, 1–28.

Brown, Penelope; Levinson, Stephen L. 1987. Politeness: Some Universals on Language Usage. Studies in Interactional Sociolinguistics 4. Cambridge: Cambridge University Press.

Bunt, Harry 1999. Dynamic interpretation and dialogue theory. – M. M. Taylor, F. Néel, D. G. Bouwhuis (Eds.). The Structure of Multimodal Dialogue II. Philadelphia, Amsterdam: John Benjamins Publishing Company, 139–166.

Curl, Traci S.; Drew, Paul 2008. Contingency and action: A comparison of two forms of requesting. – Research on Language and Social Interaction, 41, 1–25.

Fishel, Mark 2007. Machine learning techniques in dialogue act recognition. – Estonian Papers in Applied Linguistics, 3, 117–134.

Gerassimenko, Olga; Hennoste, Tiit; Kasterpalu, Riina; Koit, Mare; Rääbis, Andriela; Strandson, Krista; Valdisoo, Maret; Vutt, Evely 2007. Kliendi soovide automaatne tuvastamine eestikeelsetes infodialoogides. – Eesti Rakenduslingvistika Ühingu aastaraamat, 3, 134–154.

Hennoste, Tiit 2000. Eesti suulise kõne uurimine: transkriptsioon, taust ja korpus. – Keel ja Kirjandus, 2, 91–106.
Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – R. Pajusalu, I. Tragel, T. Hennoste, H. Õim (toim.). Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: TÜ Kirjastus, 56–73.
Hennoste, Tiit 2003. Suulise eesti keele uurimine: korpus. – Keel ja Kirjandus, 7, 481–500.
Hennoste, Tiit; Lindström, Liina; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Eesti suulise kõne korpus ja mõne allkeele võrdlemise katse. – T. Hennoste (toim.). Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: TÜ Kirjastus, 245–284.
Hennoste, Tiit; Koit, Mare; Kullasaar, Maret; Rääbis, Andriela; Vutt, Evely 2002a. Eesti dialoogikorpuse loomise probleemid. – R. Pajusalu, T. Hennoste (toim.). Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Tartu, 143–160.
Hennoste, Tiit; Lindström, Liina; Gerassimenko, Olga; Jansons, Airi; Rääbis, Andriela; Strandson, Krista; Toomet, Piret; Vellerind, Riina 2002b. Suuline kõne ja morfoloogiaanalüsaator. – R. Pajusalu, T. Hennoste (toim.). Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Tartu: TÜ Kirjastus, 161–171.
Hennoste, Tiit; Koit, Mare; Rääbis, Andriela; Strandson, Krista; Valdisoo, Maret; Vutt, Evely 2003. Directives in Estonian information dialogues. – V. Matoušek, P. Mautner (Eds.). Text, Speech and Dialogue. Proceedings of the 6th International Conference, TSD 2003, České Budějovice, Czech Republic, September 8–12, 2003. Lecture Notes in Com puter Science, 2807. Berlin, Heidelberg: Springer Verlag, 406–411. doi:10.1007/b13236
Hennoste, Tiit; Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: TÜ Kirjastus.
Hennoste, Tiit; Koit, Mare; Strandson, Krista; Rääbis, Andriela; Valdisoo, Maret; Vutt, Evely 2004. Küsimuste ja direktiivide märgendamine eestikeelsetes infodialoogides. – H. Metslang (koost.), M.-M. Sepper, J. Lepasaar (toim.). Toimiv keel II. Töid rakenduslingvistika alalt. Tallinna Pedagoogikaülikooli eesti filoloogia osakonna toimetised 3. Tallinn: TPÜ Kirjastus, 138–154.
Hennoste, Tiit; Gerassimenko, Olga; Kasterpalu, Riina; Koit, Mare; Rääbis, Andriela; Strandson, Krista 2008. From human communication to intelligent user interfaces: Corpora of Spoken Estonian. – Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008, European Language Resources Association (ELRA), Marrakech, Morocco, May, 28–30. http://www.lrec-conf.org/proceedings/lrec2008/summaries/518.html (29.09.2008).
Hutchby, Ian; Wooffitt, Robin 1998. Conversation Analysis. Principles, Practices and Applications. Cambridge, UK: Polity Press.
Jefferson, Gail 2004. Glossary of transcript symbols with an introduction. – G. H. Lerner (Ed.). Conversation Analysis. Studies from the First Generation. Amsterdam/Philadelphia: John Benjamins, 13–59.
Jokinen, Kristiina; Hurtig, Topi; Hynna, Kevin; Kanto, Kari; Kaipanen, Mauri; Kerminen, Antti 2001. Selforganizing dialogue management. – Proceedings of the Natural Language Pacific Rim Symposium (NLPRS). Workshop Neural Networks and Natural Language Processing, Tokyo, Japan.
Kaalep, Heiki-Jaan 1997. An Estonian morphological analyser and the impact of a corpus on its development. – Computers and Humanities, 31 (2), 115–133. doi:10.1023/A:1000668108369
Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus, 1, 22–29.

Kasterpalu, Riina; Gerassimenko, Olga 2006. Vestlusanalüüs. – I. Tragel, H. Õim (toim.). Teoreetiline keeleteadus Eestis II. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 7. Tartu: TÜ Kirjastus, 112–126.

Koit, Mare 2003. Märgendatud dialoogikorpus kui keeleressurss. – M. Langemets, H. Sahkai, M-M. Sepper (toim.). Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Tallinn: Eesti Keele Sihtasutus, 119–136.

Koit, Mare 2007. Arvuti suhtluses. – Eesti Rakenduslingvistika Ühingu aastaraamat, 3, 193–209.

Koit, Mare; Valdisoo, Maret; Gerassimenko, Olga; Hennoste, Tiit; Kasterpalu, Riina; Rääbis, Andriela; Strandson, Krista 2006. Processing of requests in Estonian institutional dialogues: Corpus analysis. – Petr Sojka, Ivan Kopecek, Karel Pala (Eds.). Text, Speech and Dialogue. 9th International Conference, TSD 2006. Brno, Czech Republic, September 11–15, 2006. Proceedings. Lecture Notes in Computer Science, 4188. Berlin, Heidelberg: Springer Verlag, 621–628. doi:10.1007/11846406_78

Koit, Mare; Gerassimenko, Olga; Rääbis, Andriela; Strandson, Krista 2008. Developing a dialogue system: How to grant a customer´s directive? – P. Sojka, A. Horak, I. Kopecek, K. Pala (Eds.). Text, Speech and Dialogue. Proceedings of the 11th International Conference, TSD 2008, Brno, Czech Republic, September 8–12, 2008. Lecture Notes in Computer Science, 5246. Berlin, Heidelberg: Springer-Verlag, 593–600. doi:10.1007/978-3-540-87391-4_75

Muischnek, Kadri; Müürisep, Kaili; Orav, Heili; Rääbis, Andriela; Uibo, Heli 2000. Süntaktiline märgendamine – arvutiga ja käsitsi. – T. Hennoste (toim.). Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: TÜ Kirjastus, 219–243.

Müürisepp, Kaili; Nigol, Helen; Uibo, Heli 2006. Eesti suulise keele korpuse automaatne pindsüntaktiline analüüs. – M. Koit, R. Pajusalu, H. Õim (toim.). Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Tartu: TÜ Kirjastus, 72–84.

Müürisep, Kaili; Nigol, Helen 2008. Where do parsing errors come from: The case of spoken Estonian. – P. Sojka, A. Horak, I. Kopecek, K. Pala (Eds.). Text, Speech and Dialogue. Proceedings of the 11th International Conference, TSD 2008, Brno, Czech Republic, September 8–12, 2008. Lecture Notes in Computer Science, 5246. Berlin, Heidelberg: Springer-Verlag, 161–168. doi:10.1007/978-3-540-87391-4_22

Roosmaa, Tiit; Koit, Mare; Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina; Uibo Heli 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? – Keel ja Kirjandus, 3, 192–209.

Schegloff, Emanuel A. 1979. The relevance of repair to Syntax-for-Conversation. – Talmy Givon (Ed.). Discourse and Syntax. Syntax and Semantics 12. New York: Academic Press, 261–288.

Schegloff, Emanuel A.; Sacks, Harvey 1973. Opening up closings. – Semiotica, 4, 289–327.

Stolcke, A.; Coccaro, N.; Bates, R.; Taylor, P.; Van Ess-Dykema, C.; Ries, K.; Shriberg, E.; Jurafsky, D.; Martin, R.; Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. – Computational Linguistics, 26 (3), 339–373. doi:10.1162/089120100561737

Treumuth, Margus 2005. A software tool for the Estonian Dialogue Corpus. – Proceedings of Second Baltic Conference on Human Language Technologies, Tallinn, 4–5 April, 341–346.

Treumuth, Margus; Alumäe, Tanel; Meister, Einar 2006. A natural language interface to a theater information database. – T. Erjavec, J. Žganec Gros (Eds.). Language Technologies, IS-LTC 2006: Proceedings of 5th Slovenian and 1st International Conference, 9–10 October, Ljubljana, Slovenia. Ljubljana, 27–30.

Valdisoo, Maret; Vutt, Evely; Koit, Mare 2003. On a method for designing a dialogue system and the experience of its application. – Journal of Computer and Systems Sciences International, 42 (3), 456–464.

Lisa. Kesksed transkriptsioonimärgid

Lausungid (vooruehitusüksused)

.	langev intonatsioon
?	tõusev intonatsioon
,	poollangev intonatsioon

Pausid

(.)	mikropaus: 0.2 sekundit või lühem
(0.8)	mõõdetud paus kümnendiksekundites

Prosoodilised ja paralingvistilised nähtused

`	graavis, rõhutatud sõna või silp
>... <	kiirem segment
<... >	aeglasem segment
*... *	vaiksem segment
AHA	valjem segment
mhemhe	naer
s(h)õna	naerdes hääldatud sõna
@...@	hääletooni muutus, nt imiteerimine
-	sõna poolelijäämine
:	hääliku venitamine
.hhh	häälekas sissehingamine
.jaa	sissehingamise ajal hääldatud sõna
=h	häälekas väljahingamine (sõna lõpul)

Pealerääkimine ja otsarääkimine

=	otsarääkimine (kahe üksuse vahel ei ole vaikust)
[	pealerääkimise algus
]	pealerääkimise lõpp

Kommentaarid

{--}	transkribeerimatu segment
(( ))	transkribeerija kommentaar

Tiit Hennoste (Tartu Ülikool) on uurinud suulist eesti keelt ja suhtlust.
tiit.hennoste@ut.ee

Olga Gerassimenko (Tartu Ülikool) on uurinud tagasisidevahendeid eesti ja vene suulises suhtluses.
olga.gerassimenko@ut.ee

Riina Kasterpalu (Tartu Ülikool) uurimisvaldkonnad on suuline suhtlus, dialoogi struktuur.
riina.kasterpalu@ut.ee

Mare Koit (Tartu Ülikool) on uurinud dialoogi modelleerimist arvutil.
mare.koit@ut.ee

Andriela Rääbis (Tartu Ülikool). Uurimisvaldkonnad on suuline kõne, telefonisuhtlus, infodialoogide struktuur.
andriela.raabis@ut.ee

Krista Strandson (Tartu Ülikool) on uurinud parandusi suulises eesti keeles.
krista.strandson@ut.ee

CORPUS OF SPOKEN ESTONIAN AND HUMAN-COMPUTER INTERACTION

**Tiit Hennoste, Olga Gerassimenko,
Riina Kasterpalu, Mare Koit,
Andriela Rääbis, Krista Strandson**
University of Tartu

We argue for the necessity of studying human-human spoken conversations of various kinds in order to create user interfaces to databases. An efficient human-computer dialogue system benefits from a well-organized corpus that can be used for investigating the strategies people use in conversations in order to be efficient and to handle the problems of spoken communication. For modelling natural behaviour and for testing the model we need a dialogue corpus where the roles of participants are close to the roles of a dialogue system and its user. For creating a user interface the corpus of one institutional conversation type is insufficient, since we need to know what phenomena are inherent to spoken language in general, what means are used only in certain types of conversations and what the differences are. For that reason, we collect and investigate the Corpus of Spoken Estonian and the Estonian Dialogue Corpus (a subcorpus of the former) as sources for investigating human-human interaction. The transcription conventions and annotation typology of spoken human-human dialogues in Estonian are introduced. Application of the Estonian Dialogue Corpus for investigating formal and functional characteristics of requests in information dialogues is presented

Keywords: corpus of spoken language, dialogue corpus, transcription, dialogue acts, annotation, spoken interaction, Estonian