doi:10.5128/ERYa10.02

DIALOOGIDE PRAGMAATILISE ANALÜÜSI TARKVARA

Sven Aller, Olga Gerassimenko, Tiit Hennoste, Riina Kasterpalu, Mare Koit, Krista Mihkels, Kirsi Laanesoo, Andriela Rääbis

Ülevaade. Meie uurimisobjektiks on eestikeelsed dialoogid (inimestevaheliste suuliste dialoogide transkriptsioonid ja inim-masindialoogide logifailid). Eesmärgiks on määrata, missuguseid dialoogiakte ja missuguseid suhtlusstrateegiaid kasutavad dialoogis osalejad ning missugustest struktuuriosadest dialoogid koosnevad, s.o läbi viia pragmaatilist analüüsi. Selleks, et lihtsustada dialoogide pragmaatilist analüüsi, arendame tarkvara, mis võimaldab tuvastada ja märgendada dialoogi tekstis dialoogiakte, suhtlusstrateegiaid ja dialoogi struktuuriosi (sh alamdialooge). Dialoogiaktide tuvastamiseks rakendame andmepõhist meetodit, samas suhtlusstrateegiate ja dialoogi struktuuri määramine põhineb reeglitel. Tarkvara kasutavad lingvistid dialoogide uurimisel, mille kaugem eesmärk on dialoogsüsteemi arendamine, mis suhtleks kasutajaga loomulikus keeles, järgides inimestevahelise suhtluse norme.*

Võtmesõnad: dialoog, dialoogiakt, suhtlusstrateegia, struktuur, eesti keel

1. Sissejuhatus

Traditsioonilise käsitluse kohaselt eelnevad seotud teksti pragmaatilisele analüüsile teksti koosseisus olevate lausete morfoloogiline, süntaktiline ja semantiline analüüs, kus iga järgmine analüüsietapp kasutab eelmiste etappide tulemusi. Meie püüame läbi viia pragmaatilist analüüsi nii, et see ei eeldaks eelmisi analüüsietappe (morfoloogiline jne), sest nende läbiviimiseks ei tarvitse veel leiduda (piisavalt töökindlat) tarkvara. Seega on pragmaatilise analüüsi sisendiks puhas tekst, st eeltöötlemata tekstifail. Esmalt määrame selles dialoogiaktid (DA), kasutades statistilist meetodit. Seejärel toimub dialoogistrateegiate ja dialoogi struktuuri tuvastamine, kasutades selleks reegleid, mis põhinevad DA-del.

* Artikli valmimist on toetanud Euroopa Regionaalarengufond Eesti Arvutiteaduse Tippkeskuse kaudu, Eesti Teadusagentuur (projektid SF0180078s08 "Loomulike keelte arvutitöötluse formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine", ETF9124 "Suhtlusagendi modelleerimine ja Eesti dialoogikorpus", ETF8558 "Eestikeelse spontaanse dialoogi struktuuri loomise keelelised vahendid") ning Haridus- ja Teadusministeerium (projekt EKT5 "Eestikeelse dialoogi pragmaatika analüsaator").

Meie eesmärgiks on luua tarkvara, mida saaksid kasutada keeleteadlased dialoogide märgendamisel, nende uurimisel ja struktuuri võrdlemisel.

Maailmas on välja töötatud mitmeid erinevaid DA-de tüpoloogiaid (nt Sinclair, Coulthard 1975, Stenström 1994, Allwood jt 2001). Praegu on kujunenud standardiks DAMSL (Allen, Core 1997), mis võimaldab dialoogides märgendada lausungeid erinevatel tasemetel ja ühtlasi määrata kõneaktide vahelisi seoseid. Tartu Ülikoolis on välja töötatud vestlusanalüüsil (Hutchby, Wooffitt 1998) põhinev tüpoloogia, mida kasutame oma dialoogikorpuses DA-de märgendamisel.

DA-de automaatseks tuvastamiseks dialoogides on kasutatud mitmesuguseid erinevaid andmepõhiseid meetodeid: n-gramme, Markovi peitmudeleid, Bayesi klassifikaatoreid, tehisnärvivõrke, otsustuspuid, transformatsioonipõhist masinõpet, mälupõhist õpet jne (Reithinger, Maier 1995, Wright jt 1999, Keizer jt 2002, Grau jt 2004, Levin jt 2003, Samuel jt 1998, Fernandez jt 2005, vt ka Fishel 2007a). Meie kasutame Naiivse Bayesi klassifikaatorit.

Suhtlusstrateegiaid on märgendatud ja analüüsitud infodialoogides (Jokinen 1996, Eskor 2006, 2005, 2004) ja argumenteerimisdialoogides (Eskor 2007, Georgila jt 2011). Viimasel juhul on kasutatud hüvitusega masinõpet. Meie lähtume oma tarkvara arendamisel Jokineni konstruktiivses dialoogimudelis (*Constructive Dialogue Model*, CDM) toodud suhtlusstrateegia mõistest (Jokinen 2009, 1996). Lausungitele suhtlusstrateegiate omistamisel kasutame reegleid, mis põhinevad DA-del. Dialoogi struktuuri tuvastamine toimub samuti reeglipõhiselt.

Artikli ülesehitus on järgmine. Osas 2 tutvustame oma empiirilist materjali: Eesti dialoogikorpust ja dialoogiaktide tüpoloogiat. Osad 3 kuni 5 kirjeldavad meie tarkvara võimalusi: poolautomaatne DA-de tuvastamine, suhtlusstrateegiate ja dialoogi struktuuri automaatne tuvastamine. Osas 6 teeme kokkuvõtteid.

2. Korpus ja dialoogiaktide tüpoloogia

Eesti dialoogikorpus sisaldab kolme liiki dialooge. Korpuse esimese osa moodustavad inimestevahelised suulised vestlused, mis on salvestatud autentsetes tingimustes ja transkribeeritud, kasutades vestlusanalüüsi (VA) transkriptsiooni. Selliste dialoogide koguarv on üle 1000. Dialoogide salvestamise põhieesmärk on olnud uurida inimestevahelist suhtlust. Seetõttu on korpuses erinevat tüüpi dialooge: telefonikõnesid (helistamine infotelefonile, reisibüroosse, bussijaama, polikliinikusse) ja silmast-silma vestlusi (poes, teeninduses, reisibüroos, teejuhatamised tänaval jne). Selline mitmekesisus teeb dialoogide automaatse analüüsi muidugi raskemaks, aga on vajalik inimsuhtluse igakülgsel uurimisel. Korpus on avatud ja kasvav, järjest lisanduvad uued dialoogid. Enamuse korpusest moodustavad siiski institutsionaalsed infotelefonikõned (Hennoste jt 2009).

Korpuse teine osa on kogutud võlur Ozi meetodil, kus arvuti rolli mängib kasutaja eest varjatult teine inimene (Kullasaar 2001, Pärkson 2011). Eksperimentideks kasutasime veebipõhist tarkvara (Treumuth 2011, Käbin 2011). Kasutaja sisestab oma teksti (infosoovi) klaviatuurilt ja loeb võluri vastuseid ekraanilt. Selliste dialoogide arv on umbes 100.

Kolmanda osa korpusest moodustavad dialoogid kahe veebis kasutatava dialoogsüsteemiga. Neist üks annab infot kinodes linastuvate filmide kohta ja teine

hambaraviinfot. Nagu võlur Ozi eksperimentides, sisestab ka siin kasutaja oma eestikeelse teksti klaviatuurilt ja näeb "arvuti" vastuseid ekraanilt (Treumuth 2011). Selliste dialoogide arv on praegu samuti umbes 100.

Tarkvara arendamiseks on kasutatud ühte osa korpuses leiduvatest suulistest infodialoogidest. Samas on tarkvara kavandatud kogu järjest täieneva korpuse automaatseks analüüsiks.

Dialoogiaktide märgendamiseks erinevates korpustes on varem välja töötatud mitmeid erinevaid tüpoloogiaid, millest tuntuim on DAMSL (Allen, Core 1997). Selle tüpoloogia põhieesmärk on esitada iga lausungi kõikvõimalikud funktsioonid, mida lausung saab kanda, ja seosed erinevate kõneaktide vahel.

Kuna meie tegeleme põhiliselt inimestevahelise suhtluse uurimisega, siis oleme aastatel 2001–2004 välja töötanud omaenese DA-de tüpoloogia, mis põhineb VA printsiipidel ja mida me kasutame DA-de märgendamiseks oma korpuses. Meie tüpoloogias jagatakse kõik DA-d kahte rühma: naabruspaariaktid, kus esiliige ootab teatud kindlat järelliiget (nagu küsimus – vastus), ja üksikaktid, mis ei oota reaktsiooni (nt vastuvõtuteade). Teisalt on kõik DA-d jagatud infoaktideks (nt erinevad küsimuste tüübid) ja suhtluse juhtimise aktideks (nt partneri algatatud parandused). Iga DA nimi koosneb kahest osast, mis on teineteisest eraldatud kooloniga (nt KY: suletud kas, VR: vastuvõtuteade): esimene osa näitab aktiklassi nime (nt KY – küsimused, VR – vabatahtlikud reaktsioonid) ja teine osa on akti pärisnimi (nt suletud kas, vastuvõtuteade). Aktide koguarv on meie tüpoloogias 126. Võrreldes mõne muu tüpoloogiaga (nt Allwood jt 2001) on meie omas rohkem tagasisideakte, sh nii üksikaktid (nagu vabatahtlikud reaktsioonid) kui ka naabruspaariaktid (nt partneri algatatud parandused). DA-de täieliku nimekirja võib leida nt teoses Hennoste, Rääbis 2004.

Näide (1) kujutab endast suulise dialoogi transkriptsiooni, kus on märgendatud DA-d meie tüpoloogia kohaselt (H – helistaja ehk klient, V – vastaja ehk ametnik). DA-de nimed on püstkriipsude vahel. Mõnel lausungil on mitu märgendit, st need lausungid kannavad korraga mitut funktsiooni. Näites on kasutatud VA transkriptsiooni.

(1)
((numbri valimine, kutsung)) | RIE: KUTSUNG |
V: info | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
tere | RIE: TERVITUS |
H: .hh tere | RIJ: VASTUTERVITUS |
ma paluks (.) rahandusülikon- õõ rahandusosakonnast=õ (.) `Siili telefoni-
numbrit. | DIE: SOOV |
(...)
V: Anne `Siil on see.= | PPE: ÜMBERSÕNASTAMINE | | KYE: VASTUST
PAKKUV |
H: =jah.= | PPJ: LÄBIVIIMINE | | KYJ: JAH |
V: =`kolm seitse viis üks kolm `kolm. | DIJ: INFO ANDMINE |
(0.5)
H: kolm seitse `viis üks kolm `kolm. | VR: NEUTRAALNE VASTUVÕTU-
TEADE |
aitäh? | RIE: TÄNAN |
(.) nägemist | RIE: HÜVASTIJÄTT |

Siiani oleme järginud sellist lähenemisviisi, kus kaks lingvisti märgendavad teineteisest sõltumatult DA-d käsitsi, kasutades tarkvara, mis lihtsustab dialoogide valikut korpusest ja DA-de valikut nimestikust (Vutt 2001), ja seejärel ühtlustab kolmas isik märgendused. Automaatne märgendamine teeks selle töö tunduvalt hõlpsamaks.

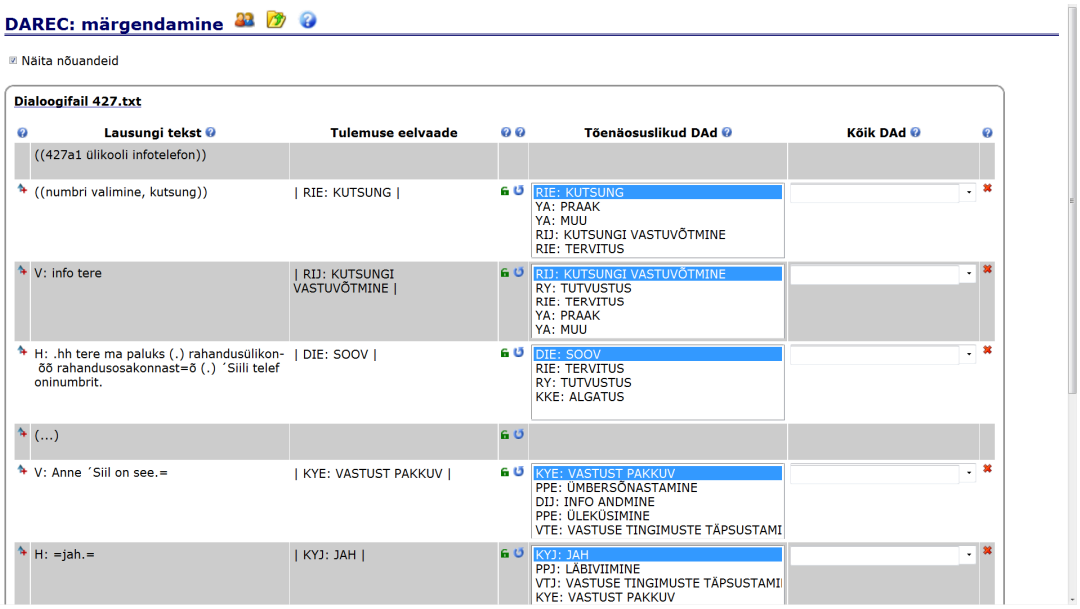
3. Dialoogiaktide tuvastamine

Nagu eespool mainitud, on DA-de tuvastamiseks dialoogitekstides kasutatud mitmesuguseid andmepõhiseid meetodeid. Ka meie otsustasime valida mõne andmepõhise meetodi, eelistades seda reeglite koostamisele, sest tuvastatavaid DA-sid on meie tüpoloogias väga palju.

Eestikeelsetes dialoogides DA-de klassifitseerimiseks meie tüpoloogia alusel on testitud mitmeid erinevaid meetodeid: (mitmekihilised) tajurid, otsustuspuud, sufiksipuud, Naiivse Bayesi klassifikaator (Fishel 2007a, Kikas 2007, Fišel, Kikas 2006). Kahjuks ei osutunud ükski meetod piisavalt heaks, et seda rakendada DA-de täiesti automaatseks tuvastamiseks. See on tingitud vähemalt kahest asjaolust: DA-de tüpoloogia keerukus ja meie (suhteliselt väikese) korpuse mitmekesisus, mis teeb raskeks meetodite treenimise. Seetõttu otsustasime oma tarkvaras realiseerida DA-de poolautomaatse märgendamise. Realiseerimiseks valisime kõige tõrkekindlama ja lihtsama meetodi: Naiivse Bayesi.

Programm tükeldab dialoogi teksti lausungiteks ja määrab igale lausungile kuni viis tõenäolisemat DA märgendit. Seejärel võib inimene-märgendaja parandada vigu ja vajadusel korrata automaatset märgendamist. Sisendiks on .txt fail: dialoogi tekst, milles voorud (aga mitte lausungid) asuvad igaüks erinevas reas. Väljund on .txt fail, kus voorud on tükeldatud lausungiteks, mis on paigutatud erinevatesse ridadesse, ja lausungitele on määratud DA-d Bayesi klassifikaatoriga. Katsetuste teel on klassifikaatoris valitud kasutamiseks järgmised tunnused, mis andsid parimaid tulemusi: sõnade trigrammide tõenäosus, lausungi pikkus ja DA märgendite tõenäosuste geomeetriline keskmine (vt Fishel 2007b).

Märgendaja töötab kahes faasis: treenimine ja märgendamine. Treenimisele võib eelneda ristvalideerimine. Treeningu alguses initsieeritakse uus sessioon, luuakse mudel ja kasutatakse seda uute andmete märgendamisel. Klassifikaator on realiseeritud programmeerimiskeeles Perl. Treenimiseks valiti eesti dialoogikorpusest 800 infodialoogi ja kasutati 10-kordset ristvalideerimist. Klassifikaatori saagis on 64,7% ja täpsus 33,0%. Need tulemused tunduvad olevat kehvad, kuid tuleb arvesse võtta, et arvutused tehti iga lausungi jaoks kõige tõenäolisemat märgendit kasutades, samas kui programm pakub kuni viis märgendit tõenäosuste kahanemise järjekorras. Inimene-märgendaja ei pea enam otsima sobivat märgendit kogu DA-de nimestikust, vaid enamasti leiab selle nende viie hulgast, mida klassifikaator pakkus.



Joonis 1. Lõik märgendatavast dialoogist märgendamislehel (vt ka näide 1)

Veebiliides käivitab kasutajaga suheldes serveris Perli skripte. Liides koosneb viiest lehest: 1) kasutajate haldus, 2) kasutajafailide haldus, 3) sisendfaili toimetamine, 4) märgendamine ja 5) väljundfailide kuvamine. Liides on realiseeritud keeles PHP, kasutajamugavuse ja efektiivsuse tõstmiseks on lisatud mõned skriptid keeles JavaScript (joonis 1, vt ka Aller 2012).

4. Suhtlusstrateegiate tuvastamine

Teadaolevalt on suhtlusstrateegiaid märgendatud ja uuritud infodialoogides ja kokkuleppimisdialoogides. Viimastes on tuvastatud argumenteerimisstrateegiaid ja kasutatud selleks hüvitusega masinõpet (Georgila jt 2011). Meie lähtume infodialoogide märgendamisel suhtlusstrateegia mõistest, mille on sisse toonud Kristiina Jokinen oma konstruktiivses dialoogimudelis (Jokinen 1996). Suhtluses osaleja kasutab suhtlusstrateegiat oma järgmise lausungi ülesehitamisel reaktsioonina partneri lausungile. Mudel arvestab suhtlusstrateegia valikul nelja binaarsete väärtustega kontekstifaktorit:

1) ootused – kas partneri lausung on ootuspärane (1) või mitte (0),
2) teema – kas partneri lausung jätkab sama teemat (1) või mitte (0),
3) initsiatiiv – kas kõnelejal on initsiatiiv (1) või mitte (0),
4) eesmärgid – kas kõneleja eesmärgid on täidetud (1) või mitte (0).

Seega on võimalikke suhtlusstrateegiaid 16 (tabel 1, vt ka Eskor 2005).

Tabel 1. Suhtlusstrateegiad konstruktiivses dialoogimudelis vastavalt neljale kontekstifaktorile (strateegiate nimetused on ingliskeelsed ja paksus kirjas, antud on ka eestikeelsed tõlked)

Ootused	Teema	Eesmärk	Initsiatiiv kõnelejal	Initsiatiiv kuulajal
oodatud	seotud	täitmata	***backto*** tagasi eelmise juurde	***follow-up-old*** jätka eelmisega
		täidetud	***finish/start*** eelmine lõpetatud, alusta uuega	***follow-up-new*** jätka uuega
	mitteseotud	täitmata	***repeat-new*** korda uut	***new question*** uus küsimus
		täidetud	***specify*** täpsusta eelmist	***new request*** uus soov
mitteoodatud	seotud	täitmata	***subquestion,X*** lisaküsimus	***continue*** jätka
		täidetud	***new dialogue*** uus dialoog	***somethingelse*** muu
	mitteseotud	täitmata	***object,X*** vaidle vastu	***notrelated*** teemaga mitteseotud
		täidetud	***specify-new*** paku uus	***new st-request*** uus väide

Igale strateegiale võib vastavusse seada kontekstifaktorite väärtustest moodustatud vektori (tabel 2).

Tabel 2. Suhtlusstrateegiatele vastavad vektorid (kontekstifaktorite väärtused: oodatud – seotud – kõneleja initsiatiiv – eesmärk)

Suhtlusstrateegia	Vektor
notrelated	0000
new-st-request	0001
object,X	0010
specify-new	0011
continue	0100
somethingelse	0101
subquestion,X	0110
new-dialogue	0111
new question	1000
new-request	1001
repeat-new	1010
specify	1011
follow-up-old	1100
follow-up-new	1101
backto	1110
finish/start	1111

Märgendasime käsitsi suhtlusstrateegiad 50 infodialoogis, mis olid juhuslikult valitud eesti dialoogikorpusest ja märgendatud DA-dega. Nende dialoogide uurimise käigus sõnastasime reeglid, mis kasutavad suhtlusstrateegia määramiseks DA-sid ja osalejatunnuseid. Selle tulemuseks oli järgmine algoritm (silumise käigus on seda hiljem täpsustatud).[1]

Ei märgenda

1. Rituaalseid akte üldiselt ei märgenda, v.a
   RIE: lõpusignaal – *specify-new*.
   RIJ: lõpetamise vastuvõtmine – *follow-up-new*.
   RIJ: lõpetamise tagasilükkamine – *somethingelse*.
2. Praaki (PRAAK) ei märgenda.

Naabruspaariaktid

Direktiivid, küsimused, arvamused

Esiliikmed

3. Esimene H: DIE/KYE on *finish/start*.
4. Hilisem H: DIE/KYE (ainsa märgendiga)
   a. Kui ei eelne V: DIJ/KYJ, siis *backto*.
   b. Kui eelneb V: info puudumine või teemavahetus, siis (algab uus teema) *specify-new*.
   c. Kui eelneb V: info andmine, siis oli saadud info puudulik, jätkub sama teema: *new-dialogue*.
5. V: KYE: alternatiiv/jutustav kas või DIE: pakkumine on *new-dialogue*.
6. H: DIE+TVE on *specify-new*.
7. H: DIE+PA on *specify-new* (vahetab eneseparandusega teemat).
8. SEE: arvamus on *new-dialogue*.

Järelliikmed

9. DIJ/KYJ: info andmine/nõustumine on *follow-up-old*.
10. DIJ/KYJ: info puudumine/mittenõustumine on *continue*.
11. DIJ: nõustuv ei on *continue* (nagu info puudumine).
12. SEJ: muu on *continue* (nagu mittenõustumine).

Topeltmärgenditega (järel- ja esiliige):

13. KYJ+KYE on *new-question*.
14. DIJ+DIE on *new-st-request*.

Kontakti kontroll

15. KKE on *specify-new*.
16. KKJ on *follow-up-new*.

Alamdialoogid

17. VTE ja PPE on *subquestion,X*.
18. VTJ ja PPJ on *follow-up-old*.

Üksikaktid

Infolisad

19. H: IL: täpsustamine on (tema DIE/KYE täpsustamine) *backto*.

[1] Kasutame siin suhtlusstrateegiate ingliskeelseid nimetusi, nagu need esinevad konstruktiivses dialoogimudelis.

20. V: IL: täpsustamine/põhjendamine/ülerõhutamine sama osaleja info andmise järel on *follow-up-new*.

Vabatahtlikud reaktsioonid

21. H: VR: neutraalne/hinnanguline jätkaja on *continue*.
22. Kõik ülejäänud VR on *somethingelse*, v.a
    a. VR: neutraalne/hinnanguline piiritleja.
        i. Kui järgneb info andmine, siis ta on *object,X*.
        ii. Kui ei järgne info andmist (dialoogi lõpus), siis *specify*.
    b. VR: neutraalne/hinnanguline info osutamine uueks on *repeat-new*.

Üksikaktid

23. YA: info andmine on *follow-up-old*.
24. YA: muu on *somethingelse*.

Automaatse märgendaja sisendiks on dialoogi tekst (.txt), milles on märgendatud DA-d, ja väljundiks tekstifail, kus on märgendatud suhtlusstrateegiad ning lisatud neile vastavad vektorid, mis esitavad nelja kontekstifaktori väärtusi (0 või 1). Mõned lausungid jäävad märgendita, sest strateegiad on seotud eeskätt info soovimise ja info andmisega (näide 2 – programmi väljund; vrd ka näide 1). Programmeerimiskeeleks on PHP.

(2)
((numbri valimine, kutsung)) | RIE: KUTSUNG
V: info | RIJ: KUTSUNGI VASTUVÕTMINE | RY: TUTVUSTUS
tere | RIE: TERVITUS
H: .hh tere | RIJ: VASTUTERVITUS
ma paluks (.) rahandusülikon- õõ rahandusosakonnast=õ (.) `Siili telefoninumbrit. | DIE: SOOV **[finish/start] [1111]**
V: Anne `Siil on see.= | PPE: ÜMBERSÕNASTAMINE | KYE: VASTUST PAKKUV **[subquestion,X] [0110]**
H: =jah.= | PPJ: LÄBIVIIMINE | KYJ: JAH **[follow-up-old] [1100]**
V: =`kolm seitse viis üks kolm `kolm. | DIJ: INFO ANDMINE **[follow-up-old] [1100]**
H: kolm seitse `viis üks kolm `kolm. | VR: NEUTRAALNE VASTUVÕTUTEADE **[somethingelse] [0101]**
aitäh? | RIE: TÄNAN
(.) nägemist | RIE: HÜVASTIJÄTT

5. Infodialoogi struktuuri tuvastamine

Tüüpiline kõne infotelefonile koosneb kolmest osast: 1) rituaalne algus, 2) põhiosa, kus esitatakse infosoov ja saadakse vastus, ja 3) rituaalne lõpp.

[RITUAALNE ALGUS]
H: ((kutsung)) RIE: Kutsung
V: RIJ: Kutsungi vastuvõtmine RY: Tutvustus [V esitleb teenusepakkujat]
(RY: Tutvustus [V esitleb ennast])
(RIE: Tervitus)
H: RIJ: Vastutervitus
[PÕHIOSA]
H: DIE: Soov / KYE: Avatud/ Jutustav kas
([INFOJAGAMISDIALOOG]
--> V: VTE: Vastuse tingimuste täpsustamine
<-- H: VTJ: Vastuse tingimuste täpsustamine
)
([PARTNERI ALGATATUD PARANDUS]
--> V/H: PPE: Ümbersõnastamine/ Üleküsimine/Mittemõistmine
H/V: PPJ: Läbiviimine
<-- (V/H: VR: Paranduse hindamine)
)
V: (VR: Neutraalne vastuvõtuteade DIJ/KYJ: Keeldumine) DIJ/KYJ Info andmine
([PARTNERI ALGATATUD PARANDUS]
--> H/V: PPE: Ümbersõnastamine/ Üleküsimine/Mittemõistmine
V/H: PPJ: Läbiviimine
<-- (H/V: VR: Paranduse hindamine)
)
(H: VR: Neutraalne vastuvõtuteade / Neutraalne piiritleja/ Neutraalne info osutamine uueks
)
[RITUAALNE LÕPP]
H: RIE: Tänan (RIE: Hüvastijätt)
(V: RIJ: Palun RIJ: Vastuhüvastijätt)

Joonis 2. Infodialoogi osad: rituaalne algus, põhiosa, rituaalne lõpp. Alamdialoogi algust tähistab --> ja lõppu <--. Ümarsulgudes on DA-d(e järjendid), mis võivad puududa (vt ka Koit 2012). Suhtlejad on H (helistaja) ja V (vastaja)

Põhiosa tuumaks on naabruspaar direktiiv – direktiivi täitmine või küsimus – vastus. Põhiosa koosseisus võib esineda alamdialooge: helistaja infosoovile võib järgneda vastaja täpsustav küsimus ja helistaja vastus sellele või algatab üks osalejatest paranduse, mille viib läbi partner.

Korpuse analüüsi tulemusel leidsime, et dialoogi struktuursete osade ja alamdialoogide tuvastamiseks saab kasutada DA-de naabruspaare kui põhilisi märguandeid.

Rituaalse alguse ja lõpu saab tuvastada rituaalsete naabruspaariaktide ja üksikakti RY: Tutvustus alusel dialoogi alguses või vastavalt lõpus. Põhiosa algab kohe pärast rituaalset algust soovi või küsimusega ja kestab kuni rituaalse lõpu alguseni. Põhiosas sisalduvaid alamdialooge saab tuvastada topeltmärgendite abil: täpsustav küsimus kannab alati lisaks küsimuse märgendile ka märgendit VTE: Vastuse tingimuste täpsustamine ja partneri algatatud parandus kas märgendit PPE: Mittemõistmine, PPE: Üleküsimine või PPE: Ümbersõnastamine. Naabruspaaride

järelliikmed algavad siis vastavalt märgenditega VTJ: Vastuse tingimuste täpsustamine ja PPJ: Parandus (vt joonis 2).

Näites (3) (programmi väljund) on märgendatud dialoogi osad (RI – rituaalne algus või lõpp, PRIM – põhiosa, RP – parandus). Põhiosas sisaldub alamdialoog: kliendi H algatatud parandus, mille viib läbi ametnik V (vrd ka näide 1).

(3) /Rituaalne algus (RI)/
RI ((numbri valimine, kutsung)) | RIE: KUTSUNG |
RI V: info | RIJ: KUTSUNGI VASTUVÕTMINE | |RY: TUTVUSTUS|
RI tere | RIE: TERVITUS |
RI H: .hh tere | RIJ: VASTUTERVITUS |
/Põhiosa (PRIM)/
PRIM ma paluks (.) rahandusülikon- õõ rahandusosakonnast=õ (.) `Siili telefoninumbrit. | DIE: SOOV |
/Paranduse alamdialoog (RP) (algatab V)/
RP V: Anne `Siil on see.= | PPE: ÜMBERSÕNASTAMINE | |KYE: VASTUST PAKKUV|
RP H: =jah.= | PPJ: LÄBIVIIMINE | |KYJ: JAH|
/Põhiosa (PRIM)/
PRIM V: =`kolm seitse viis üks kolm `kolm. | DIJ: INFO ANDMINE |
PRIM H: kolm seitse `viis üks kolm `kolm. | VR: NEUTRAALNE VASTUVÕTUTEADE |
/Rituaalne lõpp (RI)/
RI aitäh? | RIE: TÄNAN |
RI (.) nägemist | RIE: HÜVASTIJÄTT |

Dialoogi struktuuri märgendaja saab sisendiks dialoogi teksti, milles on märgendatud DA-d (.txt fail), ja kasutab dialoogi osade äratundmiseks reegleid. Väljund antakse kahes formaadis: .txt ja .xml. Programmeerimiskeeleks on PHP.

6. Kokkuvõte ja edasine töö

Artiklis esitleti pragmaatilise analüüsi tarkvara, mis on kavandatud eestikeelsete dialoogide uurimise abivahendiks. Tarkvara võimaldab tuvastada ja märgendada dialoogiakte, suhtlusstrateegiaid ja dialoogide struktuuri. Esialgu kavandasime täiesti automaatset DA-de märgendamist meie tüpoloogia kohaselt ja testisime mitut erinevat andmepõhist meetodit. Kahjuks ei osutunud ükski testitud meetod piisavalt töökindlaks, sest tüpoloogia on keeruline (sisaldab 126 DA-d, mille tuvastamisel pole ka lingvistid alati üksmeelel) ja korpus ei ole homogeenne, sisaldades erinevat liiki dialooge, mida soovitakse kasutada inimestevahelise ning inimese ja arvuti vahelise suhtluse analüüsimisel ja võrdlemisel. Seetõttu on meie tarkvaras realiseeritud DA-de poolautomaatne märgendamine: programm jagab dialoogi voorud lausungiteks ning omistab igale lausungile kuni viis DA märgendit, kasutades selleks Naiivse Bayesi klassifikaatorit. Seejärel võimaldab kasutajaliides lingvistil vajadusel parandada vigu. Kui DA-d on märgendatud, siis võib lasta dialoogis

määrata suhtlusstrateegiad ja ühtlasi märgendada dialoogi struktuuri, leides alguse, põhiosa ja lõpu ning põhiosas sisalduda võivad alamdialoogid.

Pragmaatilise analüüsi tarkvara on kasutatav internetis. Kasutaja saab valida korpusest märgendamata dialoogi ja lasta selles märgendada DA-d (kus tuleb seejärel võib-olla parandada märgendusvigu) või valida dialoogi, kus on juba märgendatud DA-d ja lasta märgendada dialoogi struktuuri ja/või suhtlusstrateegiad.

Olgugi et nii suhtlusstrateegiate kui ka dialoogi struktuuri tuvastamine toimub praegu eeldusel, et sisendiks on dialoog, milles on märgendatud DA-d meie tüpoloogia kohaselt, usume, et meie tarkvara aluseks olevaid ideid võib rakendada ka mõne muu tüpoloogia puhul (näiteks eelnevalt seades teise tüpoloogia igale DA-le vastavusse DA meie tüpoloogias). Poolautomaatset DA-de tuvastajat võib treenida ka mõnes muus, eesti keelest erinevas keeles üleskirjutatud dialoogidel ja sel viisil märgendada DA-sid meie tüpoloogia järgi.

Meie tulevane töö seisneb eestikeelsete vestluste analüüsimises kirjeldatud tarkvara abil. Ühtlasi täiendame tarkvara, võttes arvesse kasutajatelt saadud tagasisidet. Kaugem eesmärk on arendada dialoogsüsteeme, mis suhtlevad kasutajaga eesti keeles, järgides inimestevahelise suhtluse norme ja reegleid.

Viidatud kirjandus

Allen, James; Core, Mark 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/ (14.2.2014).

Aller, Sven 2012. Dialoogiaktide märgendamine Eesti dialoogikorpuses: ülevaade ressurssidest ja tarkvaraarendus. [Recognition of Dialogue Acts in the Estonian Dialogue Corpus: Overview of Resources and Software Development.] Magistritöö. Tartu Ülikool.

Allwood, J.; Ahlsen, E.; Björnberg, M.; Nivre, J. 2001. Social activity and communication act-related coding. – J. Allwood (ed.). Dialog Coding – Function and Grammar. Göteborg Coding Schemas. Gothenburg Papers in Theoretical Linguistics, 85, 1–28.

Daelemans, W.; Zavrel, J.; van der Sloot, K.; van den Bosch, A. 2004. TiMBL: Tilburg Memory-Based Learner Reference Guide. Technical Report ILK 04-02. Tilburg University and University of Antwerp.

Eskor, Liina 2004. Dialoogiaktid ja suhtlusstrateegiad: eesti dialoogikorpuse analüüs. [Dialogue acts and communicative strategies: analysis of Estonian dialogue corpus.] Magistritöö. Tartu Ülikool.

Eskor, Liina 2005. Dialoogiaktid ja suhtlusstrateegiad: eesti dialoogikorpuse analüüs. [Dialogue acts and communicative strategies: analysis of Estonian dialogue corpus.] – Keel ja Kirjandus, 9, 711–727.

Eskor, Liina 2006. Suhtlusstrateegiad infodialoogides. [Communicative strategies in information dialogues.] – Mare Koit, Renate Pajusalu, Haldur Õim (toim.). Keel ja arvuti. [Language and Computer.] Tartu: Tartu Ülikooli Kirjastus, 183–195.

Eskor, Liina 2007. Suhtlusstrateegiad ja -taktikad müügivestlustes. [Communicative strategies and tactics in telemarketing calls.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 3, 83–97. http://dx.doi.org/10.5128/ERYa3.06

Fernandez, R.; Ginzburg, J.; Lappin, S. 2005. Using machine learning for non-sentential utterance classification. – Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal, 77–86.

Fishel, Mark 2007a. Machine learning techniques in dialogue act recognition. – Eesti Rakenduslingvistika Ühingu aastaraamat, 3, 117–134. http://dx.doi.org/10.5128/ERYa3.08

Fishel, Mark 2007b. Complex taxonomy dialogue act recognition with a Bayesian classifier. – Proceedings: DECALOG'2007 Workshop on the Semantics and Pragmatics of Dialogue. Rovereto, Italy, 161–162.

Fišel, Mark; Kikas, Taavet 2006. Dialoogiaktide automaatne tuvastamine. [Automatic recognition of dialogue acts.] – Mare Koit, Renate Pajusalu, Haldur Õim (toim.). Keel ja arvuti. [Language and Computer.] Tartu: Tartu Ülikooli Kirjastus, 233–245.

Georgila, K.; Artstein, R.; Nazarian, A.; Rushforth, M.; Traum, D. R.; Sycara, K. 2011. An annotation scheme for cross-cultural argumentation and persuasion dialogues. – 12th Annual SIGdial Meeting on Discourse and Dialogue. Portland, Oregon, USA, 272–278.

Grau, S.; Sanchis, E.; Castro, M. J.; Vilar, D. 2004. Dialogue act classification using a Bayesian approach. – Proceedings of the 9th International Conference Speech and Computer, 495–499.

Hennoste, Tiit; Gerassimenko, Olga; Kasterpalu, Riina; Koit, Mare; Rääbis, Andriela; Strandson, Krista 2009. Suulise eesti keele korpus ja inimese suhtlus arvutiga. [Corpus of spoken Estonian and human-computer interaction.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 5, 111–130. http://dx.doi.org/10.5128/ERYa5.07

Hennoste, Tiit; Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. [Dialogue acts in Estonian information dialogues: a typology and analysis.] Tartu: TÜ Kirjastus. http://dspace.utlib.ee/dspace/handle/10062/18995 (10.12.2013).

Hutchby, Ian; Wooffitt, Robin 1998. Conversation Analysis. Principles, Practices and Applications. Cambridge, UK: Polity Press.

Jokinen, Kristiina 1996. Cooperative response planning in CDM: Reasoning about communicative strategies. – S. LuperFoy, A. Nijholt, G. Veldhuijzen van Zanten (Eds.). TWLT11. Dialogue Management in Natural Language Systems. Enschede: Universiteit Twente, 159–168.

Jokinen, Kristiina 2009. Constructive Dialogue Modelling: Speech Interaction and Rational Agents. John Wiley & Sons Ltd. http://dx.doi.org/10.1002/9780470511275

Keizer, S.; op den Akker, R.; Nijholt, A. 2002. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. – Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, USA, 88–94. http://dx.doi.org/10.3115/1118121.1118134

Kikas, Taavet 2007. Dialoogiaktide tuvastamine eestikeelsetes dialoogides sufiksipuude abil. [Recognition of dialogue acts using suffix trees.] Magistritöö. Tartu Ülikool. http://dspace.utlib.ee/dspace/handle/10062/2755 (10.12.2013).

Koit, Mare 2012. Towards automatic recognition of the structure of Estonian directory inquiries. – A. Tavast, K. Muischnek, M. Koit (Eds.). Proceedings of 5th International Conference on Human Language Technologies: the Baltic Perspective, Baltic HLT 2012. IOS Press, 120–128.

Kullasaar, Maret 2001. Eestikeelse dialoogikorpuse arendamine "võlur Ozi" tehnikaga. [Developing the Estonian dialogue korpus by „Wizard of Oz" technique.] Magistritöö. Tartu Ülikool.

Käbin, Tiit 2011. Eestikeelsete dialoogide kogumise veebirakendus. [Estonian dialogues capture web application.] Bakalaureusetöö. Tartu Ülikool. http://comserv.cs.ut.ee/forms/ati_report/index.php?year=2011 (10.12.2013).

Levin, L.; Langley, C.; Lavie, A.; Gates, D.; Wallace, D.; Peterson, K. 2003. Domain specific speech acts for spoken language translation. – Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan.

Pärkson, Siiri 2011. Võlur Ozi eksperimentide kogumine ja partneri algatatud paranduste analüüs. [Wizard of Oz experiments and analysis of collected dialogues.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 197–214. http://dx.doi.org/10.5128/ERYa7.12

Reithinger, Norbert; Maier, Elisabeth 1995. Utilizing statistical dialogue act processing in VERBMOBIL. – Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA, 116–121. http://dx.doi.org/10.3115/981658.981674

Samuel, Ken; Carberry, Sandra; Shanker-Vijay K. 1998. Dialogue act tagging with transformation-based learning. – Proceedings of the 17th International Conference on Computational linguistics. Vol. 2. Montreal, Quebeck, 1150–1156. http://dx.doi.org/10.3115/980432.980757

Sinclair, J.; Coulthard, R. M. 1975. Toward an Analysis of Discourse. Oxford: Oxford University Press.

Stenström, A.-B. 1994. An Introduction to Spoken Interaction. London, New York: Longman.

Treumuth, Margus 2011. A framework for asynchronous dialogue systems: concepts, issues and design aspects. Dissertationes Mathematicae Universitatis Tartuensis 72. Tartu: Tartu University Press. http://dspace.utlib.ee/dspace/bitstream/handle/10062/17522/treumuth_margus.pdf?sequence=1 (10.12.2013).

Vutt, Evely 2001. Eestikeelse dialoogikorpuse märgendamistarkvara. [Software for annotation of Estonian dialogue korpus.] Magistritöö. Tartu Ülikool.

Wright, H.; Poesio, M.; Isard, S. 1999. Using high level dialogue information for dialogue act recognition using prosodic features. – Proceedings of an ESCA Tutorial and Research Workshop on Dialogue and Prosody. Eindhoven, 139–143.

Sven Aller (Tartu Ülikool), magistrikraad informaatikas, on programmeerinud keeletöötlustarkvara.
Liivi 2, 50409 Tartu, Estonia
sven.aller@ut.ee

Olga Gerassimenko (Tartu Ülikool), magistrikraad eesti keele alal, on uurinud tagasisidepartikleid eesti ja vene keeles.
Jakobi 2, 51014 Tartu, Estonia
olga.gerassimenko@ut.ee

Tiit Hennoste (Tartu Ülikool), filosoofiadoktori kraad eesti keele alal, on uurinud suulist eesti keelt.
Jakobi 2, 51014 Tartu, Estonia
tiit.hennoste@ut.ee

Riina Kasterpalu (Tartu Ülikool) on uurinud tagasisidepartikleid eesti keeles.
Jakobi 2, 51014 Tartu, Estonia
riina.kasterpalu@ut.ee

Mare Koit (Tartu Ülikool), füüsika-matemaatikakandidaadi kraad, on uurinud dialoogi modelleerimist arvutil.
Liivi 2, 50409 Tartu, Estonia
mare.koit@ut.ee

Kirsi Laanesoo (Tartu Ülikool), magistrikraad eesti keele alal, on uurinud küsimusi suulises eesti keeles.
Jakobi 2, 51014 Tartu, Estonia
kirsi.laanesoo@ut.ee

Krista Mihkels (Tartu Ülikool), filosoofiadoktori kraad üldkeeleteaduse alal, on uurinud parandusi suulises eesti keeles.
Jakobi 2, 51014 Tartu, Estonia
krista.mihkels@ut.ee

Andriela Rääbis (Tartu Ülikool), filosoofiadoktori kraad üldkeeleteaduse alal, on uurinud suulist eesti keelt.
Jakobi 2, 51014 Tartu, Estonia
andriela.raabis@ut.ee

SOFTWARE FOR PRAGMATIC ANALYSIS OF DIALOGUES

Sven Aller, Olga Gerassimenko, Tiit Hennoste, Riina Kasterpalu, Mare Koit, Krista Mihkels, Kirsi Laanesoo, Andriela Rääbis
University of Tartu

We are investigating written Estonian dialogues – transcripts of human-human spoken dialogues as well as human-computer dialogues – with the aim of determining which dialogue acts are used in interaction, which communicative strategies are followed in order to achieve communicative goals, and which structural parts a dialogue includes. In order to simplify the pragmatic analysis of dialogues, we have developed software that makes it possible to recognise and annotate the dialogue acts, the communicative strategies and the structure of a dialogue. In recognition of dialogue acts, a data-driven method is implemented. Determination of the communicative strategies and the dialogue structure is based on rules. The software is used by linguists in dialogue studies the further aim of which is to develop a dialogue system that interacts with a user in Estonian and follows norms of human-human communication. The paper introduces the software tool and gives examples of its implementation.

Keywords: dialogue, dialogue act, communicative strategy, structure, Estonian