

TARTU ÜLIKOOL  
MATEMAATIKA-INFORMAATIKATEADUSKOND  
Arvutiteaduse Instituut  
Keeletehnoloogia õppetool  
Informaatika eriala

Margus Treumuth  
**Eesti dialoogikorpus ja selle töötlemise tarkvara**  
Magistritöö

Juhendaja: Mare Koit

Tartu 2004

# Sisukord

Sissejuhatus .....	4
1. Mis on dialoogikorpused ja milleks teda vaja on? .....	6
1.1. Dialoogikorpused maailmas .....	7
1.2. Korpused märgendamine .....	7
2. Eesti dialoogikorpused koosseis .....	11
2.1. Suulised dialoogid .....	11
2.2. Arvutisimulatsioonides kogutud dialoogid .....	12
2.3. Inimese ja arvuti vahelised dialoogid .....	13
3. Eesti dialoogikorpused märgendamine .....	15
3.1.1. Transkriptsioon .....	15
3.1.2. Dialoogiaktide märgendus .....	18
3.1.3. Morfoloogiline ja süntaktiline märgendus .....	22
4. Eesti dialoogikorpused tarkvara – dialoogikorpused tööpink .....	24
4.1. Dialoogikorpused tööpingi ülesehitus .....	25
4.1.1. Dialoogide lisamine korpusedesse .....	28
4.1.2. Transkriptsioonielementide ja dialoogiaktide loendamine .....	29
4.1.3. Dialoogiaktide järgnevuste sagedustabeli leidmine .....	32
4.1.4. Otsing .....	35
4.1.5. Aruanded .....	37
4.1.6. Dialoogi puhastamine morfoloogilise analüüsi tarvis .....	39
4.1.7. Dialoogi paigutamine ajateljele .....	41
4.1.8. Dialoogi viimine XML kujule .....	44
4.2. Dialoogiüksuste eraldamine .....	48
4.3. Dialoogikorpused tööpingi kasutusjuhend .....	51
4.4. Konverteerimismoodulite automaatne loomine .....	53
5. Dialoogsüsteem Reisiagent .....	55
5.1. Dialoogsüsteemi Reisiagent moodulite kirjeldused .....	56
6. Võimalikud edasiarendused .....	61
6.1. Dialoogikorpused tööpingi edasiarendus .....	61
6.2. Dialoogsüsteemi Reisiagent edasiarendus .....	63
Kokkuvõte .....	64
Abstract .....	65
Kasutatud kirjandus .....	66
Lisa 1. Transkriptsioonimärgid .....	67
Lisa 2. Taustakirjelduse näide .....	68
Lisa 3. Eesti dialoogikorpused märgendamisel kasutatavate dialoogiaktide tüpoloogia .....	71

## Joonised

Joonis 1. <i>Dialoogi transkribeerimine Crimson Editori aknas</i> .....	18
Joonis 2. <i>Avaleht, kus vaikimisi avaneb analüüsimoodulite alamleht</i> .....	25
Joonis 3. <i>Dialoogikorpuse tarkvara moodulid (sisendiks üks või enam dialoogi)</i> .....	26
Joonis 4. <i>Moodulid, mille sisendiks saab anda ühe dialoogi</i> .....	28
Joonis 5. <i>Dialoogide lisamine korpusesse</i> .....	29
Joonis 6. <i>Otsingu aken</i> .....	35
Joonis 7. <i>Dialoogikorpuse lihtsustatud andmemudel</i> .....	36
Joonis 8. <i>Aruannete aken</i> .....	37
Joonis 9. <i>Aruande mooduli poolt genereeritud tabel voorudest, mis sisaldavad mitut akti</i> ...	38
Joonis 10. <i>Dialoogi puhastamine morfoloogilise analüüsi tarvis</i> .....	40
Joonis 11. <i>Dialoogi paigutamine ajateljele</i> .....	42
Joonis 12. <i>Dialoog ajateljel</i> .....	42
Joonis 13. <i>Dialoogi viimine XML kujule</i> .....	44
Joonis 14. <i>XML dokumendi DTD skeemi diagramm</i> .....	45
Joonis 15. <i>XMLSPY aken XML dokumendi töötlemiseks</i> .....	47
Joonis 16. <i>Dialoogsüsteemi Reisiagent tööaken</i> .....	55
Joonis 17. <i>Dialoogsüsteemi Reisiagent vastuse leidmise algoritm kasutaja küsimusele</i> .....	58
Joonis 18. <i>Dialoogikorpuse laiendatud andmemudel</i> .....	62

## Sissejuhatus

Magistritöö käsitleb olulist keeletehnoloogilist ressursi – dialoogikorpust. Töö koosneb antud ainevaldkonna ülevaatest ning toimiva tarkvara – eesti dialoogikorpuse tööpingi loomisest.

Eesti dialoogikorpusega tutvudes ning selle uurijate ja kasutajatega suheldes selgus vajadus tarkvara järele, mis aitaks nii korpuse kogujaid kui ka keeleteadlasi, kes korpust oma uurimistöös kasutavad. Magistritöö raames loodud tarkvara kasutatakse dialoogikorpuse analüüsimisel ja dialoogidele mõningate teisenduste rakendamisel (dialoogi paigutamine ajateljele, dialoogi puhastamine morfoloogilise analüüsi tarvis, dialoogi viimine XML kujule). Seeläbi on TÜ arvutilingvistika uurimisrühmas paranenud suulise kõne häälduspärase üleskirjutuse ehk transkribeerimise kvaliteet, täpsustunud transkriptsioonireeglid ning dialoogiaktide märgendusjuhend. Loodud tarkvara kasutuselevõtt on mõjunud innustavalt kõigile uurimistöös osalejatele. Käesolevas töös antakse ülevaade dialoogikorpuse tarkvara loomisest ning selle funktsionaalsusest.

Magistritöö raames on arendatud edasi ka autori bakalaureusetöös [1] esitatud ideid. Nimelt on bakalaureusetöös loodud kasutajaga eesti keeles suhtlev dialoogsüsteem Reisiagent, mis vastab küsimustele Tallinna lennujaamast väljuvate lennukite kohta, nüüd realiseeritud veebipõhise süsteemina ning selle funktsionaalsust on täiendatud morfoloogilise sünteesi ja morfoloogilise ühestamise kasutuselevõtuga.

Eesti dialoogikorpus ja dialoogsüsteem Reisiagent on omavahel tihedalt seotud. Kuigi dialoogsüsteemiga suhtlemine suulise kõne vahendusel veel ei toimi, sest puudub kõnetuvastus, on dialoogsüsteem Reisiagent siiski üks eksperimentaalne väljund eestikeelse dialoogi modelleerimise alasele uurimustööle. Lisaks pakub dialoogsüsteem ka omamoodi huvitava võimaluse dialoogikorpuse täiendamiseks – nimelt annavad Reisiagendiga salvestatud dialoogid materjali uurijatele, kes võrdlevad arvutiga peetud dialooge inimestevaheliste dialoogidega.

Töö lõpus on toodud edasiarenduskava nii dialoogikorpuse tarkvara kui ka dialoogsüsteemi Reisiagent osas.

Töö lisades on toodud:

- suulise kõne transkriptsioonimärkide loend,
- suulise dialoogi nn taustakirjelduse näide,
- eesti dialoogikorpuse märgendamisel kasutatavate dialoogiaktide tüpoloogia.

Tööga on kaasas CD, millel asuvad:

- eesti dialoogikorpuse tööpingi paigaldusfailid,
- dialoogsüsteemi Reisiagent paigaldusfailid,
- paigaldusjuhendid,
- kasutusjuhendid.

Alternatiivina tarkvara paigaldamisele on võimalus dialoogikorpuse jaoks loodud tööpingiga ja dialoogsüsteemiga Reisiagent võimalik tutvuda aadressil: <http://www.ut.ee/~treumuth/>

# 1. Mis on dialoogikorpus ja milleks teda vaja on?

Viimasel aastakümnel võib täheldada järjest kasvavat huvi dialoogsüsteemide väljatöötamise vastu. Euroopas, Ameerikas, Jaapanis luuakse juba kommertssüsteeme, mis suhtlevad kasutajaga loomuliku kõne vahendusel ja aitavad telefoni teel näiteks pileteid broneerida, reisi kavandada, ühiskondliku transpordi või toitlustusteenuste kohta infot hankida jms. Sellega seoses on tähtsale kohale tõusnud vahendid ja meetodid niisuguste süsteemide kiireks ja efektiivseks ülesehitamiseks, sealhulgas märgendatud dialoogikorpused, olemasolevate korpuste taaskasutuse, andmete korpustest automaatse hankimise, statistiliste mudelite treenimise ja testimise viisid.

Dialoogsüsteemi all mõistetakse programmi, mis suudab inimesega loomulikus keeles kas kõne või teksti vahendusel suhelda. Eristatakse kolme liiki dialoogsüsteeme: 1) lihtsad küsimus-vastussüsteemid (esimesed sellised valmisid juba 1970ndatel aastatel), 2) andmebaaside liidesed, mille üks põhifunktsioone on tõlkimine loomulikust keelest andmebaasisüsteemi päringukeelde ja tagasi, 3) ülesannete lahendamiseks ettenähtud dialoogsüsteemid ehk ekspertsüsteemide liidesed. Viimased luuakse tavaliselt kindla ülesannete klassi jaoks (nt reisi või sõjalise operatsiooni planeerimine, lepingute sõlmimine). Arvuti lahutab ülesande osaülesanneteks ning määrab rollide jaotuse süsteemi ja kasutaja vahel, lahendades oma osaülesandeid kas iseenda teadmiste ja tuletusmehhanismide abil või pöördudes selleks dialoogsüsteemi koosseisu mittekuuluvate rakendusprogrammide poole. [2]

Selleks, et luua inimesega loomulikus keeles suhtlevaid dialoogsüsteeme ja modelleerida dialoogi arvutil, on vaja uurida nii inimestevahelist kui ka inimese ja arvuti vahelist suhtlust. Sel eesmärgil kogutakse dialooge ja moodustatakse neist nn korpused – elektroonilisel kujul säilitatavad tekstikogumid. Keeleteaduses mõeldakse sõna 'korpus' all tavaliselt keeleainese kogumikku, mida kasutatakse uurimistöö materjalina. Arvutiajastul on korpusena hakatud mõistma peamiselt polüfunktsionaalseid elektroonilisel kujul olevaid tekstikogusid, millesse kuuluvad tekstid on valitud eesmärgipäraselt nii, et nendest koosnev tervik annaks tõepärase pildi kogu keelest. [3]

Dialoogikorpus pole üksnes keeletehnoloogiline ressurss, vaid ta pakub huvitavat materjali ka keele ja suhtluse uurijatele. Lisaks loomuliku keele mõistmise ja genereerimise

modelleerimisele, saab korpuste analüüsi tulemusi kasutada kõnesünteesis, kõnetuvastuses ja sõnastike loomisel.

## 1.1. Dialoogikorpusi maailmas

Järgnevalt loetleme mõned tuntumad dialoogikorpused maailmas. [4] Esimene korpus on moodustatud suulisi vestlusi lindistades, kaks järgmist korpust on kogutud Võlur Ozi (Wizard of Oz) meetodil (mis seisneb selles, et arvutit jäljendab inimene, aga dialoogsüsteemi kasutaja ei tea, et ta vestleb arvuti asemel inimesega) ning viimane korpus on saadud arvuti ja inimese vaheliste (tõeliste) dialoogide salvestamisel. Ka eesti dialoogikorpuses, mida käsitletakse peatükis 2, on need kolm dialoogide liiki esindatud.

- <http://www.cs.rochester.edu/research/speech/monroe/> - The Monroe Corpus. Korpus koosneb 20 suulisest dialoogist, mille keskmine pikkus on 20 minutit.
- <http://www.cs.rochester.edu/research/speech/trains.html> - TRAINS spoken dialog corpus. Sellelt lingilt leiab ühe CD tutvustuse, millele on salvestatud 98 dialoogist koosnev korpus. CD valmis projekti TRAINS raames, mille käigus koguti inimese ja arvuti vahelisi dialooge Võlur Ozi meetodil.
- <http://www.ida.liu.se/~nlplab/dialogues/corpora.html> - The Linköping Dialogue Corpus. Korpus koosneb 60 dialoogist ja on koostatud samuti Võlur Ozi meetodil. Poolte dialoogide korral osaleja arvas, et ta suhtleb arvutiga, aga pooltel kordadel öeldi talle ette, et see on simulatsioon.
- <http://www.tei-c.org/Applications/apps-da01.html> - Danish Spoken Language Dialogue Systems Project: User Test Corpus 1995. Testi käigus pidid dialoogis osalejad helistama süsteemile ning nende vestlustest moodustati korpus. Korpus koosneb 57 dialoogist.

## 1.2. Korpuste märgendamine

Korpus võib olla märgendatud või märgendamata sõltuvalt sellest, kas sõnadele või lausetele on lisatud lingvistilisi andmeid. Näiteks võib igale sõnale olla lisatud märgend selle kohta,

millisesse sõnaliiki ta kuulub või millise lauseliikmena ta esineb. Suured märgendatud korpused võivad oluliselt hõlbustada paljude lingvistiliste probleemide lahendamist. [6]

Korpusi võib märgendada mitmel tasemel, sh ortograafilise transkriptsiooni, morfoloogia, süntaksi, prosoodia, semantika ja pragmaatika tasemel.

Märgendamise aluseks on kodeerimisskeem, mille loomine ja arendamine sõltub omakorda märgendamise eesmärgist ning lingvistilistest nähtustest, mida analüüsida soovitakse. Erinevaid projekte korpuste märgendamiseks ja ühiste standardite loomiseks on läbi viidud juba mitmeid aastaid. Projekti DRI (Discourse Resource Initiative, 1996-1997) eesmärgiks oli välja töötada standardid märgendatud korpuse semantiliste, pragmaatiliste ja diskursuse tunnuste jaoks. Projekti EAGLES (Expert Advisory Group on Language Engineering Standards, 1998) ülesandeks oli koostada juhised dialoogide märgendamiseks ortograafilise transkriptsiooni, morfosüntaksi, süntaksi, prosoodia ja pragmaatika tasemel. Standardite loomise asemel kirjeldasid nimetatud projektid siiski üksnes enamkasutatavaid skeeme, märgenduskeeli ja -süsteeme. [2]

### **1.1. Märgendusstandardi poole: projekt MATE**

Euroopa Liidu projekt MATE (Multilevel Annotation, Tools Engineering, 1998-2000) [9] seadis oma eesmärgiks märgenduskeemide standardi väljatöötamise kuuel tasemel: prosoodia, morfosüntaks, kaasviitamine (co-reference), dialoogiaktid, suhtlusprobleemid, tasemetevahelised probleemid. Projekti tulemusena valiti iga taseme olemasolevate märgendamisskeemide hulgast parimad – teatud kriteeriumidele vastavad skeemid, mis võivad edaspidi kujuneda standardiks. [2]

**1.1.1. Prosoodianähtuste** hulka kuuluvad intonatsioon, rõhk, rütm, toon, pikkus. Selliste tunnuste varieerumine väljendabki erinevate prosoodiaüksuste funktsioone: nt langev intonatsioon on omane jutustavale, tõusev intonatsioon küsilausele jne. Ühe, ainuõige märgendamissüsteemi loomine ja arendamine on prosoodia vallas tänapäeval veel liiga keeruline, kuna prosoodia märgendamise eesmärgid võivad olla väga erinevad. Seetõttu on MATE raames välja pakutud neljatasemeline prosoodia märgendusstruktuur: 1) foneetiline transkriptsioon; 2) intonatsiooni foneetiline esitus; 3) intonatsiooni fonoloogiline esitus; 4) nn prosoodiliste fraaside leidmine (lausungid jagatakse osadeks lähtuvalt tooniüksustest, intonatsioonigruppidest jms).



**1.1.2. Morfosüntaktilisel** tasemel käsitletakse probleeme, mis on seotud morfoloogilise ja süntaktilise märgendamise (analüüsiga). Põhiliseks probleemiks on selles valdkonnas suulise ja kirjaliku kõne suured erinevused.

**1.1.3. Kaasviitamist** kasutatakse anafooride ja muude viitamisvõimaluste tähistamiseks. Erinevates keeltes on selleks erinevad vahendid, kuid ka ühes ja samas keeles esineb erinevaid süntaktilisi võimalusi ja piiranguid. MATE projekti eesmärgiks oli luua standardiseeritud märgenduskeem, kus kasutaja saaks ise valida, milliseid süntaktilisi piiranguid kasutada konkreetse korpuse märgendamisel.

**1.1.4. Dialoogiaktide** märgenduskeeme kasutatakse lausungite oluliste tunnuste tähistamiseks. Sellised märgendid näitavad lausungi rolli dialoogis (näiteks küsimus, info andmine, tervitus jne) ning lausungitevahelisi suhteid. Suurem osa tänapäeva dialoogsüsteemidest keskendub mingile kindlale teemale või valdkonnale. Nii on võimalik läbi saada väiksema märgendite hulgaga ja tõsta süsteemi kiirust. Sellistes dialoogsüsteemides kasutatava info saab jagada kaheks: teemaga seotud ning suhtlusprotsessiks vajalik info. MATE projektis pöörati dialoogiaktide märgenduskeemide hindamisel tähelepanu märgendusjuhendi olemasolule, märgendajate arvule, märgendatud dialoogide/segmentide/lausungite arvule ning dialoogide teemale. Oluliseks peeti näidete olemasolu, kasutatud märgenduskeelt, märgenduskeemi taaskasutatavust.

**1.1.5. Suhtlusprobleemid** tekivad sageli just inimese ja arvuti vahelises suhtluses, arvuti piiratud võimaluste tõttu. Kui inimestevahelises vestluses tekib probleem (nt kuulaja ei saa aru, mis kõneleja ütles), siis see lahendatakse: kõigepealt selgitatakse probleem välja ning seejärel tehakse parandus, kusjuures parandusprotsessi võib alata nii kuulaja kui ka kõneleja. Arvutiga, eriti suulise kõne süsteemiga suheldes on asi keerulisem, sest kõiki inimestevahelise suhtluse parandusprotsessi võimalusi pole suudetud veel realiseerida. Ideaalis peaksid dialoogsüsteemid olema võimalikult koostöövalmid, sest see muudab suhtluse sujuvamaks. Pealegi saab niiviisi ära hoida kasutaja algatatud parandusprotsesse (üleküsimisi, mittemõistmisi jms) ja kasutaja niisugust käitumist, millest süsteem aru ei saa ning mille lahendamise ta seetõttu toime ei tule.

**1.1.6. Tasanditevaheline** märgendamine tähistab MATE projektis kõiki nähtusi, mida iseloomustatakse teistele elementidele viitamisega. Vaadeldavad elemendid võivad kuuluda erinevatele tasanditele (nt sõna ja foneem) või samale tasandile: lause ja lause (võib vaadelda näiteks sõnade järjekorda või leksikaalseid suhteid).

Ka eespool (**1.1** Dialoogikorpuse maailmas) loetletud korpused on märgendatud. Kõigis suulise dialoogi korpustes on märgendatud prosoodianähtused, osades ka morfoloogilised nähtused. Monroe korpuses on märgendatud ka dialoogiaktid. Viimases neist, Taani korpuses, hoitakse dialooge XML kujul ning ka seda ideed on eesti dialoogikorpuse tööpingis kasutatud.

## 2. Eesti dialoogikorpuse koosseis

Nagu eespool antud ülevaatest selgus, võib dialooge koguda mitmel viisil. Eesti dialoogikorpuses on kolme liiki materjali:

1. inimestevahelised suulised dialoogid, mis on lindistatud ja seejärel häälduspäraselt üles kirjutatud (transkribeeritud);
2. inimese ja arvuti vahelised simuleeritud dialoogid, mis on kogutud Völur Ozi meetodil, kus arvuti rolli mängib kasutaja teadmata teine inimene;
3. inimese ja arvuti vahelised tegelikud dialoogid, mis on kogutud käesolevas töös realiseeritud veebipõhise programmi Reisiagent abil.

Vaatleme lähemalt neid erinevat liiki dialooge.

### 2.1. Suulised dialoogid

Tartu Ülikooli suulise kõne uurimisrühmas on juba alates 1997. aastast lindistatud eestikeelset suulist kõnet, enamasti dialooge. Eesti suulise kõne korpus moodustub peamiselt argivestlustest, kaubandusvestlustest, teenindusvestlustest. Lisaks dialoogidele sisaldab korpus ka tele- ja raadiosaateid, koolitunde, jutlusi, loenguid, koosolekuid. [7]

Suulise kõne korpuse maht ulatub praeguseks u 600 000 sõnani. (Vahel kasutatakse ka terminit *tekstisõna*, mis tähendab, et sõnade all mõeldakse sõnavorme, nii nagu nad tekstis esinevad, mitte aga sõnade algvorme.)

Suulise kõne korpuse abil uurivad keeleteadlased, kuidas suuline keel erineb kirjalikust keelest, millised on vestlusstrateegiad, millised struktuurid esinevad vestlustes (nt dialoogiaktide naaberpaarid, nagu küsimus ja vastus või tervitus ja vastutervitus).

Dialoogide üleskirjutust nimetatakse litereerimiseks. Kui litereerimine on häälduspärane, siis nimetatakse seda transkribeerimiseks. Dialoogid on reeglina litereeritud tervikuna. Litereerimisel on kasutatud konversatsioonianalüüsi transkriptsiooni (transkriptsioonimärgid vt Lisa 1). Iga teksti juurde kuulub taustakirjeldus (vt Lisa 2), milles on andmed lindistamise, litereerimise, situatsiooni, osalejate, teema, tekstitüübi jms kohta. Toome näitena (Näide 1)

telefonikõne hambakliinikusse. Osalejad on H (helistaja ehk klient) ja V (vastaja ehk ametnik).

**Näide 1.** *Telefonikõne hambakliinikusse.*

H: ((kutsung))  
V: Hansluks=hambaravi=tere.  
H: mt=.hh tere. (.) kas teil 'üliõpilastele mingit 'hinnasoodustust ka on=h.  
V: mt=.hh=ää < tähendab (.) ää 'haigekassaga on 'nimodi ett=o (.) mt=ö nende > (.) 'täidise hinnad tuleb 'ikkagi ise maksta, ['kõigil.] (.) .hh noh (.) 'haigekassa maksab muidugim 'osa.  
H: [jah]  
H: jah  
V: aga selle: suhtes ei 'ole üliõpilastele mingit 'erisoodustust. 'muidu kui noh 'lihtsalt 'ravi on kui 'plommi ei=ole=vel, (0.5) sis sellest 'ravist=e võetakse 'tunnitasu, mitte 'nimodi=neid igat asja 'eraldi.  
H: ahah. ja 'kui palju see tunnitasu 'on.=  
V: =tunnitasu on 'kakssada 'krooni. (.)  
H: selge. (0.5) aitäh?  
V: jah palun.  
H: 'nägemist  
V: nägemist.

## 2.2. Arvutisimulatsioonides kogutud dialoogid

22 kirjalikku dialoogi (vt Näide 2) on saanud Võlur Ozi tehnikat rakendades: katseisikutel palutakse justkui testida üht arvutiprogrammi, mis annab nende poolt esitatud küsimustele korrektseid ja informatiivseid vastuseid. Tegelikuses on aga suhtluspartneriks arvutivõrgu vahendusel teine inimene (antud juhul oli selleks M. Kullasaar-Valdisoo). Dialoogide teemaks oli valitud reise planeerimine, lubatud olid infopäringud bussifirma Sebe busside, Tartu-Helsingi lennukite ja Tallinn-Helsingi laevade kohta. Läbiviidud katseid, tekkinud probleeme ja nende lahendusi on lähemalt analüüsitud magistritöös [8].

Sellise nn simuleerimise teel kogutakse konkreetse valdkonna dialooge, mis eeldatavasti inimese ja arvuti vahelises suhtluses aset leida võivad, ja moodustatakse neist dialoogikorpused, mida hiljem kasutatakse dialoogsüsteemi programmeerimisel. Nii viisi kogutud korpused maht on praegu u 2500 sõna. Dialoogi keskmine pikkus on 250 sõna (võrreldav inimestevahelise suulise dialoogi pikkusega).

**Näide 2.** *Katkend Võlur Oz" meetodil kogutud dialoogist.*

Infoklient: Kuidas sõita Tartust Pärnu enne kella 12 hommikul  
Arvuti: Buss väljub kell 05.00  
Arvuti: Buss väljub kell 08.00

Arvuti: Kas Teid huvitavad ka saabumisajad?  
Infoklient: Jah  
Arvuti: Üks hetk, palun!  
Arvuti: Kell 05.00 saabub 08.20  
Arvuti: Kell 08.00 saabub 11.40  
Infoklient: Kas kella 12 ajal läheb ka praam Kuressaarde  
Arvuti: Buss väljub kell 08.20  
Infoklient: Kas Saaremaale saab kuidagi  
Arvuti: Mis peatusest Te sõita soovite?  
Infoklient: Valjala  
Arvuti: Täpsustage sihtpeatuse nimi, palun!  
Infoklient: Pärnu  
Arvuti: Selline reis meie andmebaasis puudub!  
Infoklient: Tahaks sõita pühapäeval Paidest Narva  
Arvuti: Kas olete nõus Tallinnas ümber istuma?  
Infoklient: jah  
Arvuti: üks hetk, palun!  
Arvuti: Buss Paidest väljub kell 10.35  
Arvuti: Buss Tallinnasse saabub kell 12.00  
Arvuti: Kas Te soovite lähimat bussi Tallinnast edasi liikumiseks?  
Infoklient: jah  
Arvuti: Buss Tallinnast väljub kell 13.00  
Arvuti: Buss Narvasse saabub kell 16.25

### **2.3. Inimese ja arvuti vahelised dialoogid**

Kolmas liik dialooge eesti dialoogikorpuses on inimese ja arvuti vahel tegelikult asetleidnud dialoogid, mis on kogutud käesoleva töö autori poolt loodud veebipõhise dialoogsüsteemiga Reisiagent. Selle dialoogsüsteemi loomist alustati autori bakalaureusetöös [1]; magistritöö raames realiseeriti mõned edasiarendusideed.

Bakalaureusetöös loodud programm ei töötanud veebis, vaid ta tuli kasutaja arvutisse installeerida, magistritöös realiseeriti see süsteem veebipõhisena ja täiendati ka tema funktsionaalsust morfoloogilise sünteesi ja morfoloogilise ühestamise kasutuselevõtuga. Lähemalt on seda kirjeldatud magistritöö 5. peatükis.

Süsteem on käivitatav magistritöös loodud dialoogikorpuse tööpingi avalehelt ning asub aadressil: <http://www.aprote.ee/reisiagent/>. Nii dialoogsüsteem Reisiagent (vt 5 ptk Dialoogsüsteem Reisiagent) kui ka dialoogikorpuse tööpink (vt 4 ptk Eesti dialoogikorpuse tarkvara – dialoogikorpuse tööpink) paiknevad füüsiliselt AS Aprote serveris, sest realisatsioon nõudis PHP lisamooduleid, mida vaikumisi Tartu Ülikooli serveris ei olnud. Edasiste plaanide hulka kuulub ka läbirääkimine TÜ administraatoritega teatavate lisamoodulite paigaldamiseks ülikooli serverisse ning tööpingi üleviimine AS Aprote serverist ülikooli serverisse.

Käesoleval ajal on dialoogsüsteemi kasutajateks vaid need isikud, kellel autor on palunud seda süsteemi testida. Ka reisibüroo Estravel töötajad on seda proovinud kasutada.

Veebipõhise dialoogsüsteemiga Reisiagent kogutud dialoogide arv on praegu 14 ning sõnade arv umbes 6000, vt Näide 3).

### **Näide 3.** Dialoog Reisiagendiga

Küsimus: kas helsingi kaudu viini saab lennata

Vastus: Helsingisse saab lennata teisipäeval kell 06:25, kell 09:10, kell 11:30, kell 14:50, kell 15:25, kell 17:40 ja kell 21:30.

Küsimus: millal lennuk helsingist tagasi tuleb

Vastus: Helsingisse saab lennata teisipäeval kell 06:25, kell 09:10, kell 11:30, kell 14:50, kell 15:25, kell 17:40 ja kell 21:30.

Küsimus: kas teil tallinnna saabuvate lendude kohta ka infot on

Vastus: Tean väljumisi järgnevatesse linnadesse: Amsterdam, Kopenhaagen, Frankfurt, Hamburg, Helsingi, Kiiev, London, Moskva, Oslo, Riia, Stockholm, Vilnius, Varssav, Tampere, Praha, Pariis, Kârdla, Kuressaare, Berliin.

Küsimus: kui palju maksab pilet moskvasse

Vastus: Moskvasse saab lennata kolmapäeval kell 06:50.

Küsimus: palju pilet maksab

Vastus: Pileti saate Estraveli kassast.

Küsimus: kus see asub

Vastus: Kahjuks ei saanud teist aru.

Küsimus: kus asub estraveli kassa

Vastus: Kassa asub lennujaamas.

Küsimus: mis kell kassa avatakse

Vastus: Kassa asub lennujaamas.

Küsimus: kus lennujaam asub

Vastus: Lennujaam asub Tallinnas. Tulge kohale tund enne väljumist.

### 3. Eesti dialoogikorpuse märgendamine

Nagu eelmises peatükis antud ülevaatest nähtub, moodustavad olulisima osa eesti dialoogikorpusest suulised dialoogid. Need on praegu märgendatud neljal tasemel:

- transkriptsioon,
- dialoogiaktid,
- morfoloogia (100 000 sõna)
- süntaks (proovitud mõnedes dialoogides).

Võlur Ozi meetodil kogutud dialoogides on märgendatud ainult dialoogiaktid.

Vaatame lähemalt neid märgendustasemeid.

#### 3.1.1. Transkriptsioon

Transkriptsioonis märgendatakse [7]:

- sõnad ja mitmesugused suhtlushäälitused,
- suhtlusüksused,
- pausid lausungite sees ja nende vahel,
- kõne omadused (intonatsioon, venitused, katkestamised, rõhud, valjus),
- peale- ehk koosrääkimised ja haakumised: juhtumid, kus kaks või enam inimest kõnelevad korraga või kus ühe kõneleja jutu lõpp ja teise algus haakuvad tihedalt kokku,
- transkribeerija kahtlused (halvasti kuulnud sõnad vms),
- kirjeldused nähtustest, mille kohta puudub transkriptsioonimärk või mida transkribeerija ei taha transkribeerida, kuid mis on vajalikud ära näidata (kõrvalised hääled, nutt)

(vt Näide 1 ja Lisa 1).

Eelmainitud näites on kõnevoor ühel real ning algab osalejatunnusega V, millele järgneb osaleja poolt lausutud tekst (lausung):

**V: Hansluks=hambaravi=tere.**

Tekstis on kolm kokkuhääldatud sõna, selle väljendamiseks kasutatakse võrdusmärki. Tekst lõpeb punktiga, mis tähistab langevat intonatsiooni.

Teiselt realt algab uus kõnevoor, mida alustab osaleja H.

**H: selge. (0.5) aitäh?**

Osaleja teeb kahe sõna vahel pausi, mis kestab 0.5 sekundit. Lausung lõpeb küsimärgiga, mis tähistab tõusvat intonatsiooni.

Eesti suulistes dialoogides on siiani transkriptsioonimärkide lisamine toimunud käsitsi (s.t suulise kõne rühma liikmed kuuluvad lindistatud kõnet ja püüavad seda kuulmise järgi üles kirjutada). Hiljuti hakati suuliste dialoogide kirjapanekuks ja transkriptsioonimärkide lisamiseks katsetama tarkvara CLAN (Computerized Language Analysis) [10], mis lisaks transkribeerimise hõlbustamisele (CLAN aitab mõõta pauside pikkusi ja määrata pealerääkimise piire) võimaldab ka näiteks sõnade ja foneemide sageduste leidmist, heli- ja videofailide sidumist transkriptsiooniga (need tööd on selle tarkvara arendamisel aga alles algusjärgus).

Kuna CLAN ei paku süntaktiliste konstruktsioonide eristamiseks värvidega esiletoomise võimalust, on magistritöö raames kohaldatud üks vabavaraline tekstitoimeti nõnda, et see annaks kasutajale võimaluse transkriptsioonimärkide lisamisel näha teatavaid konstruktsioone värvidega esiletõstetuna.

Nii nagu suvalises programmeerimiskeeles programmeerimisel on hea omada tekstitoimetit, mis suudab vastava keele süntaktilisi konstruktsioone värvidega eristada, tehes seeläbi juba ka pindmist süntaksikontrolli ning parandades koodi loetavust, on vaja ka dialoogide transkribeerimisel analoogset abivahendit.

Selleks sobib vabavarana kättesaadav tekstitoimeti Crimson Editor. Antud tekstitoimeti jaoks on võimalik defineerida suvalise keele puhul võtmesõnade ja konstruktsioonide hulk, millele Crimson Editor oskab rakendada värvide abil esiletõstmist.

Magistritöö raames on Crimson Editori jaoks defineeritud keele Dialog konstruktsioonide hulk järgnevalt:

```
$DELIMITERS=( ) | [ ] = { } '
$KEYWORDPREFIX==
$VARIABLEPREFIX= '
$QUOTATIONMARK1= |
$BLOCKCOMMENTON=( (
$BLOCKCOMMENTOFF= ) )
$HIGHLIGHTON=(
```



```
$HIGHLIGHTOFF= )
$RANGE1BEG= [
$RANGE1END= ]
$RANGE2BEG= {
$RANGE2END= }
```

Toodud definitsioonides on Crimson Editori võimaluste piires suudetud määratleda keele Dialog järgmised konstruktsioonid:

#### **rõhulised sõnad**

```
$VARIABLEPREFIX= '
```

#### **aktimärgendid**

```
$QUOTATIONMARK1= |
```

#### **kommentaariid**

```
$BLOCKCOMMENTON= ( (
$BLOCKCOMMENTOFF= ) )
```

#### **pausid**

```
$HIGHLIGHTON= (
$HIGHLIGHTOFF= )
```

#### **pealerääkimised**

```
$RANGE1BEG= [
$RANGE1END= ]
```

#### **kokkuhääldamised**

```
$KEYWORDPREFIX==
```

#### **ebaselgused**

```
$RANGE2BEG= {
$RANGE2END= }
```

Võtmesõnade hulka tuli lisada võrdusmärk (keeles Dialog on see kokkuhääldamise märk), mis on eelnevalt defineeritud kui prefiks, sest muidu ei suuda Crimson Editor võrdusmärgi teksti seest leida (näiteks: *tere=siis*), vaid peab seda teksti osaks:

```
[keywords1:global]
```

```
=
```

Vastavalt defineeritud konstruktsioonidele moodustati uus süntaktiline menüüvalik - Dialog, mida kasutaja saab tekstile rakendada. Failidele, mis on txt laiendiga (sellise laiendiga eesti dialoogikorpuses dialoogid ka on), rakendub süntaktiline valik Dialog automaatselt.

Töö tulemusena on nüüd võimalik kasutada tekstitoimetit, mis suudab kasutajale pakkuda tagasisidet transkribeerimise süntaksi osas ning parandada transkriptsioonide loetavust. Järgneval joonisel on toodud ekraanipilt Crimson Editor'is töödeldavast dialoogist:

```

1 ((347b7 telefonikõne eraisik - kino))
2 ((ühtlustas Andriela Rääbis 30.09.2003))
3 ((kutsung)) | RIE: KUTSUNG |
4 U: kino Sa'fiir | RIJ: KUTSUNGI VASTUVÕTMINE |
5 tere | RIE: TERVITUS |
6 H: .hh (1.5) tere. | RIJ: VASTUTERVITUS |
7 .hh e selline 'küsimus | YA: EELTEADE |
8 < kas Jää'aeg on > 'inglise[keelne.] | KYE: ALTERNATIIV |
9 U: [eesti]'keelne | KYJ: ALTERNATIIV: ÜKS | ((tüdinult))
10 H: kõik. | PPE: ÜMBERSÕNASTAMINE | | KYE: VASTUST PAKKUV |
11 U: {jah}. | PPJ: LÄBIVIIMINE | | KYJ: JAH |
12 H: selge. | UR: NEUTRAALNE PIIRITLEJA |
13 aitäh? | RIE: TÄNAN |
14 U: palun=siis, | RIJ: PALUN |
15 nä[gemist] | RIE: HÜVASTIJÄTT |
16 H: [näge]'mist | RIJ: VASTUHÜVASTIJÄTT |
17

```

Joonis 1. Dialoogi transkribeerimine Crimson Editori aknas

Joonisel on näha, et dialoogiaktide märgendid on eristatud lilla värviga, kommentaarid (kahekordsetes ümarsulgudes) rohelisega, pealerääkimised (nurksulgudes) kollase taustaga, rõhulised (apostroofiga algavad) sõnad sinisega, pausid (ümarsulgudes) punasega, kokkuhääldamised punase värviga. Kui transkribeerija teeb trükkimisel vea ja unustab näiteks lõpetava sulu, siis on see viga nüüd kohe kergesti märgatav, sest trükitav tekst ei muuda värvi (kuigi peaks).

Kirjeldatud tekstitoimeti ei ole otseselt magistritöös loodud tööpingi osa (vt. 4 Eesti dialoogikorpuse tarkvara – dialoogikorpuse tööpink), kuid realiseerib funktsionaalsuse, mis oli esialgselt tööpink kavandatud. Crimson Editorile ja selle seadistusfailidele on viited tööpingi kasutusjuhendi lõpus.

### 3.1.2. Dialoogiaktide märgendus

Aktimärgendid näitavad, millist tähendust kannab lausung. Aktide märgendus võimaldab tuvastada naaberpaare (vt. Näide 5) ja vestluse stsenaariume.

Tartu Ülikoolis on välja töötatud dialoogiaktide tüpoloogia, mille esimeseks aluseks on konversatsioonianalüüsist pärit arusaam, et dialoogiaktid (neil küll kõnevoorud) jagunevad kahte rühma. Iga kõnevoor ennustab mingil määral, milline jätk tema järel tuleb ja iga voor on tehtud sobivaks eelneva kõnevooriga. Sealjuures mõned neist nõuavad kindlat teatud tüüpi järgmist vooru ja teised annavad ainult viiteid, milline voor on sobiv. Kahe vooru kogumeid, milles voorudevaheline side on eriti tugev ja konventsionaliseerunud, nimetatakse naaberpaarideks. Voorud (dialoogiaktid) jagunevad seega naaberpaare moodustavateks ja mittemoodustavateks [5].

Põhimõtteks on, et naaberpaari teine pool esitatakse siis, kui esimene akt on lõppenud, mitte varem (nt küsimusele ei vastata kohe peale küsisõna esitamist) ega mitte hiljem. Ideaalsel juhul peaks naaberpaari järelliige vahetult järgnema esiliikmele. Tegelikus suhtluses ei tarvitse see nii olla. Loomulikult võib esineda ka juhtumeid, kus naaberpaari teist liiget lihtsalt ei tulegi.

Seega peab märgendamisel muuhulgas suutma tuvastada, millal on tegu naaberpaari esiliikmega (nt küsimusega, millele oodatakse vastust) ja millal pole (nt eristama arvamust, millele oodatakse vastust, arvamusest, millele vastust ei oodata) (vt Näide 4).

#### **Näide 4.** *Naaberpaarid*

oodatakse vastust esitatud arvamusele:

H: väljas on vist külm. | SEE: ARVAMUS |

V: jah | SEJ: NÕUSTUMINE |

siin aga ei oodata teist liiget:

M: oh sa pisikene hehe | SEE: ARVAMUS |

O: [head=aega] | RIE: HÜVASTIJÄTT |

M: [nägemist] | RIJ: VASTUHÜVASTIJÄTT |

Teiseks tüpoloogia lähtekohaks on tõdemus, et aktid saab jagada nende keskse funktsiooni põhjal kahte rühma: infoaktideks, millega antakse eeskätt infot, ja suhtluse juhtimise aktideks.

Teistes tuntud aktitüpoloogiates (vt nt [9]) on juhtimisaktid tavaliselt koondatud ühte rühma. Eesti tüpoloogias on konversatsioonianalüüsi alusel sisse toodud kaks eri rühma: sujuvat

suhtlust juhtivad aktid ja probleemide lahendamise aktid, mis kasutavad nn parandusmehhanismi. See on vajalik seetõttu, et inimesega võimalikult loomulikult suhtlev arvuti peab tundma ära signaalid, mis osutavad probleemi suhtluses ja suutma eristada rikkumise heastamist sujuvast suhtlusest.

Aktide põhitüpoloogia on järgmine:

## I. Naaberpaare moodustavad aktid

### I.1. Dialoogi juhtivad aktid

Sujuv vestlus

1. Rituaalid. Siia rühma kuuluvad põhiliselt vestluse alguses ja lõpus kasutatavad vormelid (tervitamine, hüvastijätt, tänamine, palumine, esitlemine jms).

2. Teemavahetus.

Probleemide lahendamine

3. Partneri algatatud parandused.

4. Kontakti kontroll.

### I.2. Infoaktid

5. Direktiivid, millega antakse edasi ja võetakse vastu soove, ettepanekuid ja pakkumisi.

6. Küsimused ja vastused.

7. Seisukohavõttud, milles üks pool esitab mingi seisukoha (väite, arvamuse jms) ning teine pool reageerib sellele.

## II Naaberpaare mittemoodustavad aktid e üksikaktid

### II.1. Dialoogi juhtivad aktid

Sujuv vestlus

8. Rituaalsed üksikaktid (kontakteerumine, kutsumine jms).

9. Vabatahtlikud reaktsioonid (jätkaja, vastuvõtuteade jms).

Probleemide lahendamine

10. Parandused.

### II.2. Infoaktid

11. Primaarsed üksikaktid (jutustamine, lubadus, retooriline küsimus jms).

12. Infolisad, mille abil täiendatakse antud infot kõneleja enda initsiatiivil (täpsustamine, pehmendamine jms).

Praeguses aktitüpoloogias on kokku 126 akti (vt Lisa 3 ja Näide 5).

Aktinimed on moodustatud järgmisel viisil.

Iga dialoogiakti nimetus (nt RIE: TERVITUS või IL: TÄPSUSTAMINE) koosneb kahest osast:

1. akronüüm, mis sisaldab kas kaks või kolm tähte, millest esimesed kaks näitavad aktirühma (nt RI = rituaalid, IL = infolisa) ning kolmas, mida kasutatakse ainult naaberpaare moodustavate aktide puhul, näitab, kas on tegu naaberpaari esi- või järelliikmega (vastavalt E või J)
2. akti nimi (nt TERVITUS, TÄPSUSTAMINE).

**Näide 5.** *Dialoog, kus on märgendatud dialoogiaktid.*

P: mis ´korrusel on: tualett´paberit näiteks. | KYE: AVATUD |  
M: meil ´siin. | PPE: ÜMBERSÕNASTAMINE | | KYE: VASTUST PAKKUV |  
P: jah. | PPJ: LÄBIVIIMINE | | KYJ: JAH |  
M: jah | VR: NEUTRAALNE VASTUVÕTUTEADE | | VR: PARANDUSE HINDAMINE |  
seal ´teisel pool kassade juures. | KYJ: INFO ANDMINE |  
P: ´ahah | VR: NEUTRAALNE INFO OSUTAMINE UUEKS |  
´kena | VR: NEUTRAALNE PIIRITLEJA |  
´aitäh. | RI: TÄNAN |

Selles dialoogis on kaks osalejat, voore on viis, kusjuures eelviimane voor on märgendatud kahe aktina ning viimane voor on märgendatud kolme aktina.

Lausungid selles dialoogis on:

- mis ´korrusel on: tualett´paberit näiteks.
- meil ´siin.
- jah. seal ´teisel pool kassade juures.
- ´ahah ´kena ´aitäh.

Antud dialoogis esinevad naaberpaarid:

- | KYE: VASTUST PAKKUV | + | KYJ: JAH | (vastust pakkuv küsimus, millele järgneb jaatav vastus)
- | KYE: AVATUD | + | KYJ: INFO ANDMINE | (avatud küsimus, millele järgneb info andmine)

Selles näites esinev akt | KYE: AVATUD | kuulub aktitüpoloogias küsimuste rühma ning on naaberpaari eesliige. Sellele aktile vastavaks järelliikmeks on | KYJ: INFO ANDMINE |, mis paikneb samuti küsimuste rühmas.

### 3.1.3. Morfoloogiline ja süntaktiline märgendus

Morfoloogiline märgendus (mis saadakse morfoloogilise analüüsi tulemusel) määrab sõnade struktuuri (nt. tüvi, järelliide, lõpp), sõnaliigi ja käände või pöörde. Morfoloogilise analüüsi väljund on omakorda sisendiks süntaktilisele analüüsile, mida eesti suulise kõne korpuse peal on käesolevaks hetkeks ka proovitud, kui veel vähestele tekstidele. Eesti suuliste dialoogide märgendamisel on kasutatud morfoloogilist analüsaatorit ESTMORF [17], mis algselt töötati välja kirjaliku keele analüüsiks ja mida on kohandatud suulise kõne vajadustest lähtuvalt. Süntaktilise analüüsi tegemisel on kasutatud Kaili Müürisepa poolt loodud süntaksianalüsaatorit [18].

Järgnevas näites (vt Näide 6) on morfoloogiliselt ja süntaktiliselt analüüsitud kõnevooru

V: 'Estmar='info 'Leenu=kuuleb

**Näide 6.** *Morfoloogiline ja süntaktiline märgendus.*

V

```
#####
**CLB @???
```

\$<s>

```
#####
Estmar
  Estmar+0 //_S_ prop sg nom #cap // **CLB @NN>
Info
  Info+0 //_S_ prop sg nom #cap // @NN> +=@SUBJ
Leenu
  Leenu+0 //_S_ prop sg nom #cap // @SUBJ
kuuleb
  kuul+b //_V_ main indic pres ps3 sg ps af #FinV #Part // @+FMV
$</s>
#####
```

Esmalt on tekst puhastatud transkriptsioonimärkidest (kokkuhääldamist näitavad võrdusmärgid ja rõhulisi sõnu tähistavad ülakomad). Seejärel on rakendatud morfoloogilist analüüsi, mis on iga sõna järele kirjutanud uuele reale sõna tüve ja lõpu ning märkide // ja // vahele sõnaliigi ja käände või pöörde [19][20]. Morfoloogiliselt analüüsitud tekst on sisendiks süntaktilisele analüsaatorile, mis on väljastanud analüüsi peale lõpetavat // märki.

Näites analüüsitud lauses on süntaktiline analüsaator leidnud järgnevad lauseliikmed:

@NN> - nimisõna eestäiendina

@SUBJ - alus

@+FMV - finiidne öeldis

Süntaktilist analüüsi on seni tehtud väga vähestele dialoogidele, sest suulise kõne lauseehitus erineb väga palju kirjaliku keele lauseehitusest – aga olemasolev süntaksianalüsaator on loodud kirjaliku keele põhjal.

## 4. Eesti dialoogikorpuse tarkvara – dialoogikorpuse tööpink

Käesolevas töös loodud dialoogikorpuse tööpink võimaldab dialooge/dialoogikorpust analüüsida, sooritada otsingut üle korpuse, rakendada dialoogidele mõningaid teisendusi (dialoogi paigutamine ajateljele, dialoogi puhastamine morfoloogilise analüüsi tarvis, dialoogi viimine XML kujule) ning aitab ühtlasi leida transkriptsiooni- ja märgendusvigu.

Tarkvara väljatöötamise jooksul on kogu aeg arvestatud tulevaste kasutajate – põhiliselt Tartu Ülikooli suulise kõne uurijate – soovide ja vajadustega. Osa mooduleid omabki mõtet ainult suulise kõne korral (dialoogi puhastamine morfoloogilise analüüsi jaoks, dialoogi paigutamine ajateljele), enamus neist aga on rakendatavad nii suulistele kui ka kirjalikele dialoogidele (nähtuste loendamine, aktide järgnevuste sagedustabeli leidmine, dialoogi viimine XML kujule, otsing ja aruanded).

Tarkvara on realiseeritud vabavaralisel platvormil ning veebipõhise kasutajaliidesega. Arenduskeeleks on PHP [11] ja andmebaasiserveriks MySQL [12]. Veebiserveriks on Apache HTTP Server [13]. Disainis on kasutatud on stiililehti (Cascading Style Sheets) [14]. Veebipõhine tarkvara on kasutajatele mugav. Kasutajad peavad omama ainult veebilehitsejat. Lisaks on kõik parandused/täiendused kasutajaile koheselt nähtavad ja kättesaadavad. Esmase ülevaate loodud tarkvara võimalustest annab avaleht (vt Joonis 2).





Joonis 2. Avaleht, kus vaikimisi avaneb analüüsimoodulite alamleht.

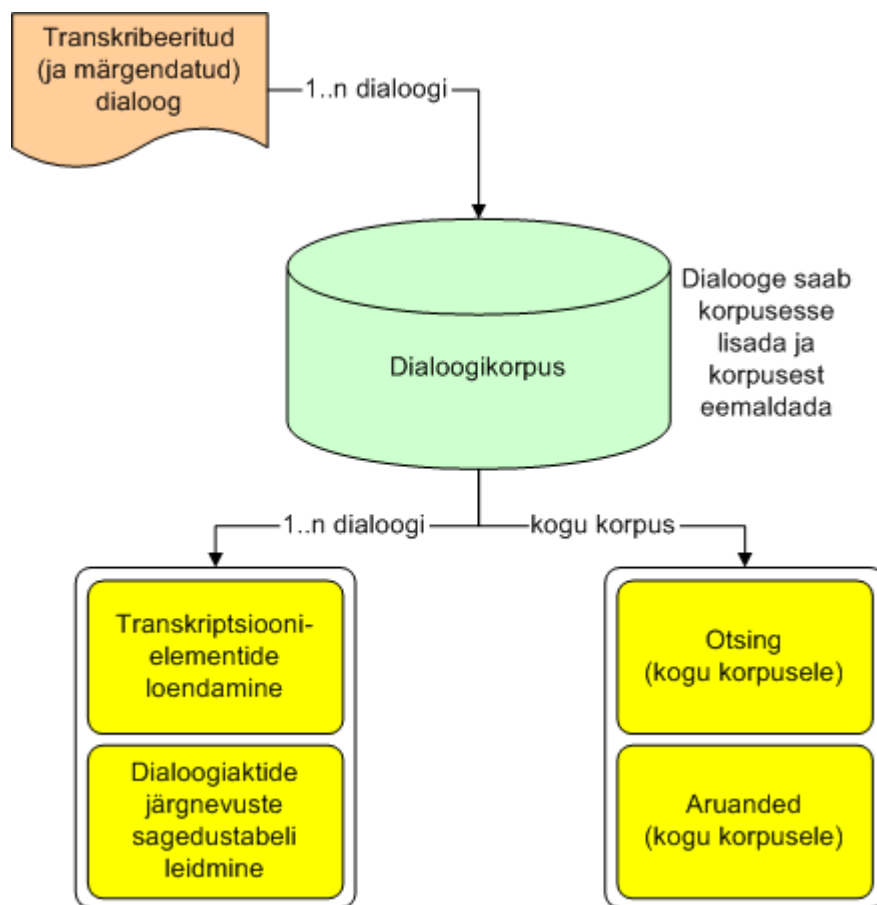
#### 4.1. Dialoogikorpuse tööpingi ülesehitus

Esialgselt oli planeeritud luua tööriistad, mis oleksid rakendatavad ühele dialoogile, st. sisendiks on üksainus dialoog. Töö käigus selgus aga vajadus rakendada mõningaid tööriistu korraka mitmele dialoogile või kogu dialoogikorpusele. Seejärel ongi osa mooduleid realiseeritud nõnda, et kasutaja saab valida analüüsimiseks kas ühe või mitu dialoogi. Osad moodulid aga võimaldavad endiselt tööd vaid ühe dialoogiga, sest näiteks polnud vajadust mitmele dialoogile korraka teisendusi rakendada.

Niisiis, loodud tarkvara koosneb moodulitest, millest ühed on rakendatavad korraka nii ühele kui mitmele dialoogile, mida saab valida korpusest, kuid teised moodulid on rakendatavad vaid ühele dialoogile korraka.

Esimene alaliik – moodulid, mis on rakendatavad korpuses leiduvatele dialoogidele nii ühe- kui mitmekaupana – pakub järgmised võimalused (vt ka Joonis 3):

- Dialoogide lisamine korpusesse/eemaldamine korpusest:  
Võimaldab korpusesse lisada dialooge ühe või mitme kaupa. Võimaldab kustutada kõik dialoogid korpusest. Ühekaupa kustutamist pole seni veel realiseeritud (vt 6.ptk. Võimalikud edasiarendused).
- Dialoogides esinevate nähtuste loendamine:  
Võimaldab loendada mitmesuguseid transkriptsioonelemente (sõnu, pause, kokkuhääldamisi jne), aktimärgendeid jms.
- Dialoogiaktide järgnevuste sagedustabeli leidmine:  
Võimaldab leida dialoogiaktide järgnevuste (bigrammide) sagedustabeli.
- Otsing  
Võimaldab teostada otsingut kahe erineva tunnuse järgi: dialoogis esinev tekst (sõne) ja aktimärgend.
- Aruanded  
Korpuses olevate dialoogide põhjal on võimalik genereerida mitmesuguseid aruandeid.



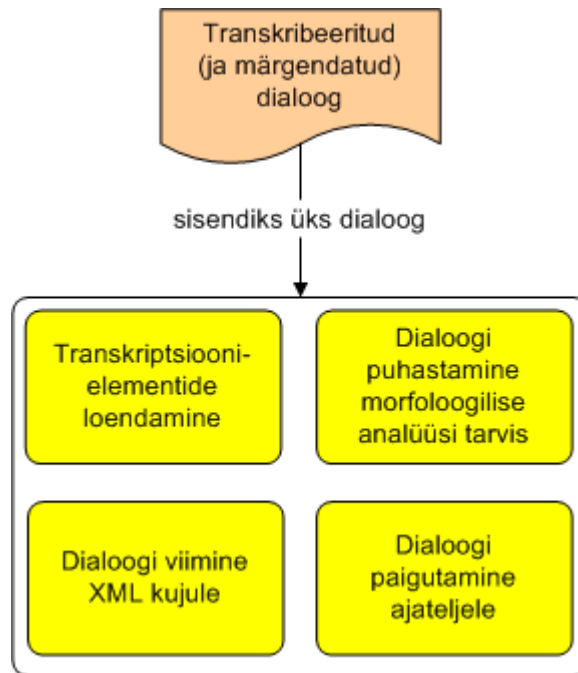
**Joonis 3.** Dialoogikorpuse tarkvara moodulid (sisendiks üks või enam dialoogi).

Teise alaliigi moodustavad teisendusmoodulid, mis on rakendatavad ühele dialoogile (vt Joonis 4). Teises alaliigis on ka analüüsimoodul transkriptsioonelementide loendamiseks, mis töötab ka esimeses alaliigis, kuid ei summeeritud tulemusi.

Kui esimese alaliigi moodulite jaoks on võimalik valida dialoogid, mis on eelnevalt korpusesse laetud, siis teise alaliigi moodulite sisendiks saab kasutaja valida vaid ühe dialoogi, mis asub kasutaja oma arvutis. Ka teise alaliigi funktsionaalsuse jaoks on plaanis realiseerida dialoogide valimine korpusest (vt 6. ptk Võimalikud edasiarendused).

Teise alaliigi moodulite poolt pakutavad funktsioonid on järgnevad:

- Dialoogis esinevate nähtuste loendamine:  
Võimaldab loendada sõnu, aktimärgendeid, pause jms.
- Dialoogi paigutus ajateljele:  
Võimaldab viia dialoogi ajateljele, edaspidi saab kasutada näiteks helifailiga sünkroniseerimiseks.
- Dialoogi XML kuju genereerimine:  
Võimaldab genereerida dialoogi XML kuju, kuhu kopeeritakse ka dialoogiaktide märgendustasand ning tehakse automaatselt morfoloogiline analüüs.
- Dialoogi puhastamine morfoloogilise analüüsi tarvis:  
Võimaldab puhastada dialoogi transkriptsioonimärgenditest, kommentaaridest jms., et dialoogile saaks rakendada morfoloogilist analüüsi (mis ei tohi saada sisendiks morfoloogia mõistes üleliigseid märke).



Joonis 4. Moodulid, mille sisendiks saab anda ühe dialoogi.

Järgnevalt vaatame lähemalt moodulite funktsionaalsust ja realisatsiooni.

#### 4.1.1. Dialoogide lisamine korpusesse

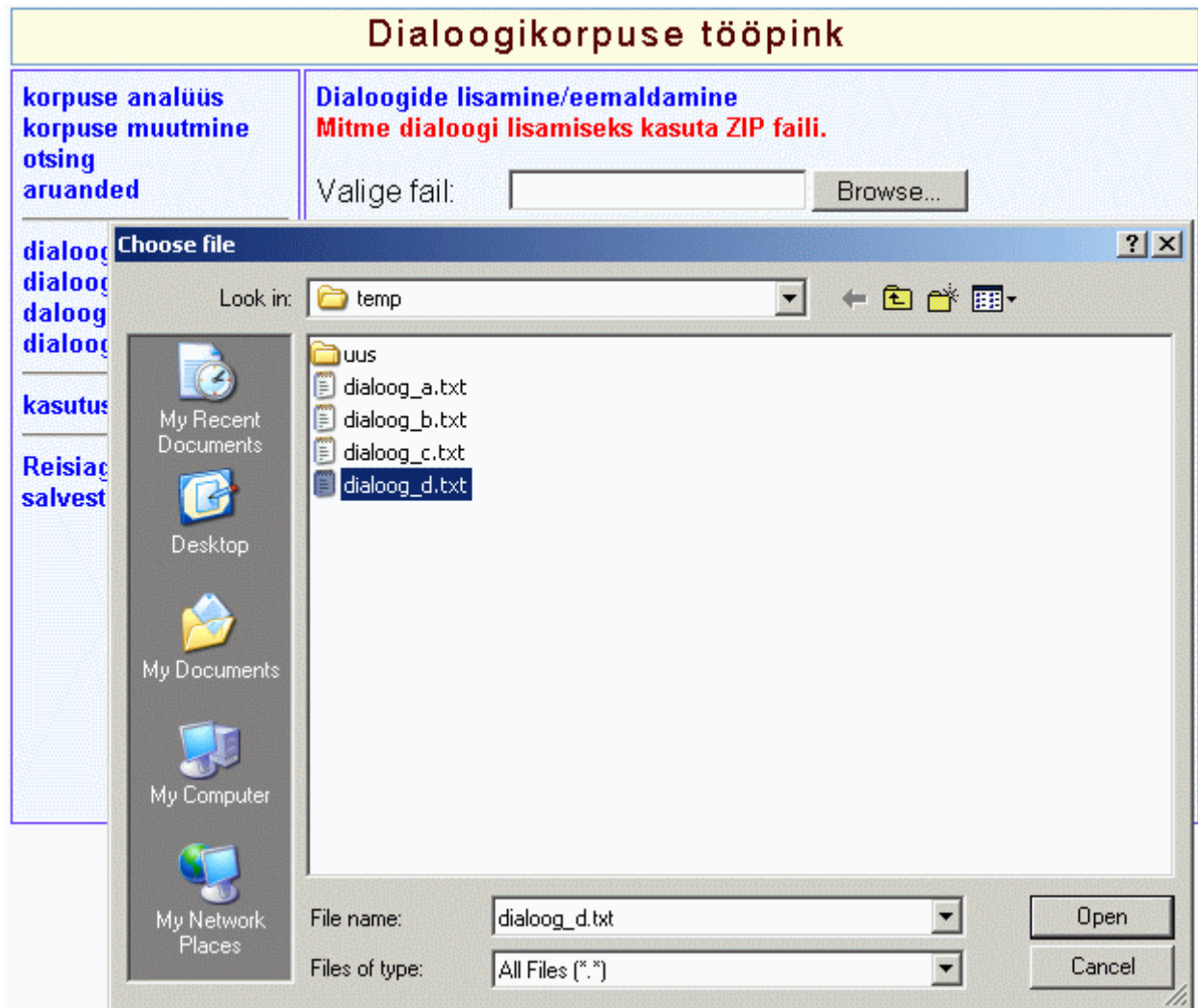
Dialoogide analüüsimiseks on vajalik lisada dialoogid korpusesse (vt Joonis 5). Korpus paikneb tööpingiga samas serveris MySQL andmebaasina.

Kuna korpus koosneb praegu paljudest dialoogidest, mis kõik asuvad erinevates tekstifailides (\*.txt), siis tuli luua võimalus mitme dialoogi üheaegselt lisamiseks. Mitme faili samaaegne lisamine veebi kaudu teostati ühe pakitud (ZIP formaadis) faili lisamisena, milles sisalduvad dialoogid. Süsteem loeb ZIP failist dialoogid ning lisab need kõik korraga relatsioonilisse andmebaasi. Andmebaasi ülesehitus on toodud jaotises 4.1.4, Otsingu mooduli tehnilises kirjelduses (vt Joonis 7).

Dialoogi lisamisel kirjutatakse samanimeline dialoog korpuses üle. (Kavas on lisada hoiatus ülekirjutamise kohta kasutajale vt. 6. ptk Võimalikud edasiarendused.) Kui dialoogid on ZIP failis ning paiknevad seal sisemistes kataloogides, siis ZIP faili lisamisel saavad kataloogide nimed failinimede prefiksiks. Lisatava dialoogifaili või ZIP faili suurim lubatud suurus on 500K. See on piisav, sest ükski praegu korpuses olev dialoog ei ületa realselt 500 kilobaiti. Isegi kogu korpus pakituna ZIP formaati ei ületa praegu 500 kilobaiti. Kui edaspidi peaks

tekkima olukord, et kogu korpus siiski pakituna ületab suuruse 500K, tuleks see saata kahes osas.

Dialoogide valikulist kustutamist ei ole käesolevaks hetkeks realiseeritud ning kustutada saab ainult kogu korpuse korraga (vt. 6. ptk Võimalikud edasiarendused).



Joonis 5. Dialoogide lisamine korpusesse.

#### 4.1.2. Transkriptsioonelementide ja dialoogiaktide loendamine

Korpusesse lisatud dialoogidest on võimalik välja valida üks või mitu dialoogi ning rakendada neile analüüsimoduleid: nähtuste loendamine ja aktide järgnevuste leidmine. Analüüsitavad dialoogid tuleb märkida nimistus valituks, valida soovitud tegevus ning käivitada analüüs.

Nähtuste loendamine seisneb mõningate transkriptsioonelementide ning aktimärgendite loendamises. Väljund kuvatakse uues aknas ning seal dialooginimele klõpsates on võimalik vaadata detailset analüüsi ning dialoogi sisu.

Selle mooduli eesmärk on statistiliste uurimuste võimaldamine. Keeleteadlased uurivad näiteks, millistes vestlustes tehakse rohkem pause või millistes vestlustes on lausungid pikemad, millistes lühemad. Nii on võimalik võrrelda erinevates situatsioonides kogutud dialooge (tänavavestlused, telefonivestlused) ning nende statistilist infot (vt Näide 7).

**Näide 7. Loendatud elemendid (mooduli väljund)**

Dialooge: 4

Aktivmärgendeid: 436

Erinevaid aktivmärgendeid: 81 vt. nimekirja

	Sõnade arv	Lausungite arv	Osa-lausungite arv	Pauside arv	Mikro-pauside arv	Pikkade pauside arv	Möödetud pauside arv	Kokku-hääldamiste arv	Venituste arv	Rõhuliste sõnade arv	Osalejate arv
<a href="#">dialoog_a.txt</a>	37	13	14	2	2	0	0	4	1	7	2
<a href="#">dialoog_b.txt</a>	129	22	24	13	4	1	8	14	2	26	2
<a href="#">dialoog_c.txt</a>	1525	205	271	93	34	2	57	111	7	337	4
<a href="#">dialoog_d.txt</a>	729	109	120	72	39	0	33	51	11	161	3
<b>KOKKU:</b>	<b>2420</b>	<b>349</b>	<b>429</b>	<b>180</b>	<b>79</b>	<b>3</b>	<b>98</b>	<b>180</b>	<b>21</b>	<b>531</b>	<b>11</b>

**Keskmine lausungi pikkus:** 6.93 sõna

**Keskmine osalausungi pikkus:** 5.64 sõna

**Keskmine pauside arv lausungis:** 0.52 pausi

**Keskmine rõhuliste sõnade arv lausungis:** 1.52 rõhulist sõna

**Keskmine rõhuliste sõnade arv osalausungis:** 1.24 rõhulist sõna

Näitest nähtub, et mõõdetakse ka lausungite ja osalausungite pikkust sõnades ning pauside ja rõhuliste sõnade arve lausungis ja osalausungis. Lausungiks loetakse transkriptsioonelementide kohaselt reavahetuse, punkti või küsimärgiga lõppevaid voore. Lausung võib sisaldada ühe või mitu osalausungit. Osalausungite eraldajana kasutatakse koma.

Dialoogi nimetusele klõpsates (näites 6 esimene veerg) on võimalik avada detailne analüüs selle dialoogi kohta, mille väljundis on esmalt näha antud dialoogi transkriptsioonelementide esinemissagedused, seejärel on arvutatud mõned keskmised ning sellele järgneb dialoog, kus on värvidega esile tõstetud pealerääkimised, kokkuhääldamised, kommentaarid ja pausid:

**Fail:** dialoog\_a.txt  
**Sõnade arv:** 37  
**Lausungite arv:** 13  
**Osalausungite arv:** 14  
**Pauside arv:** 2  
    Mikropauside arv: 2  
    Pikkade pauside arv: 0  
    Mõõdetud pauside arv: 0  
**Pealerääkimiste arv:** 0  
**Kokkuhääldamiste arv:** 4  
**Venituste arv:** 1  
**Rõhuliste sõnade arv:** 7  
**Osalejate arv:** 2  
    (Osalejad: H, V)  
**Voorude arv:** 11

**Dialogiaktide sagedused:**

DIREKTIIVI TÄITMINE: INFO ANDMINE	2
ESITLUS	2
HÜVASTIJÄTT	1
JÄTKAJA	1
KUTSUNG	1
KUTSUNGI VASTUVÕTMINE	1
LÕPU PAKKUMISE SIGNAAL	1
MEMOREERIMISKORDUS	1
PALUN	1
PARANDUSE LÄBIVIIMINE	1
SOOV	1
TERVITUS	1
TÄNAN	1
VASTUHÜVASTIJÄTT	1
VASTUTERVITUS	1
ÜLEKÜSIMINE	1

-----  
**Keskmine lausungi pikkus:** 2.85 sõna  
**Keskmine osalausungi pikkus:** 2.64 sõna  
**Keskmine pauside arv lausungis:** 0.15 pausi  
**Keskmine rõhuliste sõnade arv lausungis:** 0.54 rõhulist sõna  
**Keskmine rõhuliste sõnade arv osalausungis:** 0.5 rõhulist sõna

---

Dialog (dialoog\_a.txt)

```
infotelefon
ühtlustas Andriela Rääbis 03.12.2002
((kutsung))      | KUTSUNG |
V:      `Estmar=`info,      | KUTSUNGI VASTUVÕTMINE |      | ESITLUS |
`Leenu=kuuleb      | ESITLUS |
```

```

tere | TERVITUS |
H: ee tervist. | VASTUTERVITUS |
sooviks saada noh ´maaliini ´busside: mingi ´infotelefoni ´numbrit. | SOO
V |
V: neli seitse seitse, | DIREKTIIVI TÄITMINE: INFO ANDMINE |
H: jaa?= | JÄTKAJA |
V: =kaks kaks seitse. | DIREKTIIVI TÄITMINE: INFO ANDMINE |
(.)
H: e neli seitse=seitse? | ÜLEKÜSIMINE |
V: kaks kaks seitse. | PARANDUSE LÄBIVIIMINE |
H: kaks kaks seitse. | MEMOREERIMISKORDUS |
(.) selge | LÕPU PAKKUMISE SIGNAAL |
aitäh? | TÄNAN |
V: palun? | PALUN |
H: nägemist? | HÜVASTIJÄTT |
V: nägemist | VASTUHÜVASTIJÄTT |

```

Mooduli tehniline realisatsioon seisneb loendamisfunktsioonile parameetriga etteantava dialoogi parsimises ning üksuste eraldamises ning loendamises. Dialoogiüksuste eraldamist on kirjeldatud allpool (vt. 4.2 Dialoogiüksuste eraldamine).

### 4.1.3. Dialoogiaktide järgnevuste sagedustabeli leidmine

Korpusesse lisatud dialoogidest on võimalik valida üks või mitu ning rakendada dialoogiaktide bigrammide leidmist.

Moodul väljastab dialoogiaktide järgnevuste sagedustabeli. Samale reale paigutatud dialoogiaktide järjend loetakse bigrammi üheks liikmeks. Aktide järgnevuste leidmisel ei ületata dialoogipiire.

Dialoogiaktide märgendamiseks automatiseeritud abivahend praegu puudub. Aktimärgendite lisamise hõlbustamiseks dialoogi on olemas Evelyn Nurmsalu magistritöös koostatud programm [21], mis võimaldab menüü kaudu valida aktide nimetusi, mida soovitakse lisada märgendatavasse faili.

Eelmainitud (vt 1.2 Korpuste märgendamine) märgendustasanditest on keerukaim dialoogiaktide märgendamine. Semantilist infot kandev dialoogiaktide märgendamine on siiani isegi osaliselt automatiseerimata.

Üheks lahenduseks dialoogiaktide märgendamise osalisel automatiseerimisel oleks kasutada statistilist analüüsi ning N-gramm-loogikat [15].



Dialoogiaktide järgnevuste sagedustabeli abil on võimalik koostada N-gramm-mudel, mis kasutab üht või mitut eelnevat dialoogiakti, et tõenäosuslikult tuvastada järgnev dialoogiakt. Sellisel tõenäosuslikul lähenemisel on üks puudus: teatud järgnevused testhulgal puuduvad täiesti. See aga ei tähenda seda, et neid järgnevusi ei võiks reaalsuses esineda. Selleks, et neid reaalselt võimalikke järgnevusi mitte välistada, tuleb N-gramm mudelit siluda ning omistada ka kõigile neile järgnevustele, mis testhulgas puudusid, väikesed tõenäosused.

Sellise N-gramm-mudeli täpsus sõltub suuresti korpuse suuruselt ja korpuse dialoogide iseärasustest. Samas pakub see kindlasti võimaluse dialoogiaktide tuvastamise osaliseks automatiseerimiseks. Sarnast statistilist meetodit on kasutatud ka eesti keele morfoloogilise ühestaja loomisel ning see lähenemine on ennast õigustanud [20].

Dialoogiaktide järgnevuste sagedustabeli leidmise moodulit võib käsitleda aktide automaatse märgendamise eeltöona. Mooduli töö tulemused väljastatakse tabeli kujul (vt Näide 8).

**Näide 8.** Aktide järgnevuste sagedustabel

Dialooge: 4

Aktimärgendeid: 436

Erinevaid aktimärgendeid: 81

Aktide järgnevusi: 383

Erinevaid järgnevusi: 282

---

#### Dialoogiaktide järgnevuste sagedused:

**Samal real paiknevate aktimärgendite järjendit käsitletakse bigrammi ühe liikmena.**

1.   KYE: AVATUD   --->   KYJ: INFO ANDMINE	10
2.   KYJ: INFO ANDMINE   --->   VR: NEUTRAALNE JÄTKAJA	8
3.   KYJ: INFO ANDMINE   --->   KYJ: INFO ANDMINE	7
4.   KYJ: INFO ANDMINE   --->   VR: NEUTRAALNE INFO OSUTAMINE UUEKS	6
5.   VR: NEUTRAALNE JÄTKAJA   --->   IL: SELETAMINE	4
6.   VR: NEUTRAALNE JÄTKAJA   --->   YA: INFO ANDMINE	4
7.   VR: NEUTRAALNE JÄTKAJA   --->   KYJ: INFO ANDMINE	4
8.   IL: PÕHJENDAMINE   --->   VR: NEUTRAALNE JÄTKAJA	4
9.   KYE: VASTUST PAKKUV   --->   KYJ: JAH	4
10.   VR: NEUTRAALNE INFO OSUTAMINE UUEKS   --->   KYJ: INFO ANDMINE	3
11.   KYE: SULETUD KAS   --->   KYJ: JAH	3
12.   YA: INFO ANDMINE   --->   VR: NEUTRAALNE JÄTKAJA	3

Antud moodul töötab järgneva algoritmi kohaselt:

1. Aktimärgenditest moodustatakse massiiv, milles märgitakse ära samal real paiknevad aktimärgendid ning dialoogipiirid.

```
| AKT 1 |  
| AKT 2 |  
| AKT 3 |
```

2. Massiivist moodustatakse koopia nihkega 1 element ning liidetakse see horisontaalselt esimesele massiivile.

```
| AKT 1 | | AKT 2 |  
| AKT 2 | | AKT 3 |  
| AKT 3 | -----
```

3. Uue massiivi iga element (näiteks | AKT 1 | | AKT 2 |) on nn bigramm e. aktimärgendite paar, mis moodustus kahest teineteisele järgnevast dialoogiaktist. Unikaalsete bigrammide sagedusloend ongi mooduli väljundiks.

Kuna dialoogides samal real paiknevaid aktimärgendite järjendeid soovisid keeleteadlased käsitleda bigrammi ühe poolena, siis kirjeldatud algoritmis on tähistus | AKT 1 | lihtsustatud kuju, tegelikkuses võib see koosneda mitmest aktimärgendist.

#### 4.1.4. Otsing

Korpusesse lisatud dialoogides on võimalik sooritada otsingut (vt Joonis 6).

**Dialogikorpuse tööpink**

**korpuse analüüs**  
**korpuse muutmine**  
**otsing**  
**aruanded**

---

**dialoogi analüüs**  
**dialoogi puhastamine**  
**dialoogi ajateljel**  
**dialoogi XML kuju**

---

**kasutusjuhend**

---

**Reisiagent**  
**salvestatud vestlused**

**Otsing**

tekst:  akt:

Joonis 6. Otsingu aken.

Otsingut saab rakendada üle tunnuste "tekst" (tegelikult sõne) ja "akt" (aktimärgendid). Näiteks on uurijal võimalik leida teatud aktimärgendid või siis ignoreerida aktimärgendeid ja otsida hoopis sõnu või nende järjendeid lausungites. On võimalik siduda mõlemad tasemed ning otsida korraga nii sõnesid kui aktimärgendeid (vt Näide 9).

**Otsitav tekst: kas**

**Otsitav aktimärgend: suletud kas**

**Leiti 3 vastet.**

osaleja	lausung	aktimärgend	dialoog
K:	kas {neil} on garan'tii ka kaasas.	KYE: SULETUD KAS	<a href="#">dialoog_c.txt</a>
L:	[a kas] ned 'ära siis ei 'kulu need plast[mass'osad.]	KYE: TÄPSUSTAV    KYE: SULETUD KAS	<a href="#">dialoog_c.txt</a>
M:	kas sa said oma 'teate 'kätte.	KYE: SULETUD KAS	<a href="#">dialoog_c.txt</a>

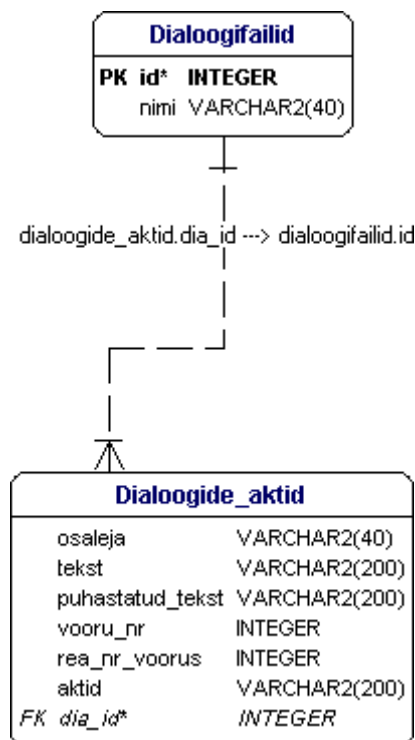
**Näide 9. Otsing teksti ja aktimärgendi järgi**

Otsingutulemustes on näha ka leitud ridade arv ning dialoogi nimele klõpsates on võimalik avada vastav dialoog.

Otsida saab ka transkriptsioonelemente. Näiteks, sisestades ülakoma, otsitakse rõhulisi sõnu; sisestades kooloni, otsitakse venitusi; sisestades võrdusmärgi, otsitakse ning näidatakse kõik kokkuhääldamised. Otsitav märgijada on tunnuse "tekst" puhul väljundis värvitud punaseks. Tunnuse "akt" puhul märgijada ei värvita.

Taolise otsingu realiseerimiseks ei saanud dialoogifaile enam tavateksti kujule jätta. Tuli luua moodul, mis eraldaks tavatekstifailis kõikvõimalikud üksused ning moodustaks nende põhjal struktuurse faili XML formaadis. Ka XML fail ei sobinud otsingu realiseerimiseks, sest selle parsimine ei olnud efektiivne. XML faile on aga võimalik viia relatsioonilisse andmebaasi, misjärel avaneb võimalus läheneda andmetele SQL päringukeele laialdasi vahendeid kasutades.

Relatsiooniline andmemudel, mis realiseeriti MySQL andmebaasina, ei sisalda morfoloogilist märgendustaset (vt Joonis 7). Täielikuma mudeli realiseerimine jääb töö edasiarenduseks. Andmemudeli laiendatud versioon on toodud 6. peatükis "Võimalikud edasiarendused".



**Joonis 7.** Dialoogikorpuse lihtsustatud andmemudel

Edaspidi on plaanis realiseerida ka morfoloogiline otsing (vt. 6. ptk Võimalikud edasiarendused). Morfoloogilise otsing annaks võimaluse sisestatud otsisõna jaoks leida kõik selle sõna esinemised suvalises muitevormis.

#### 4.1.5. Aruanded

Korpusesse lisatud dialoogide põhjal saab genereerida mitmesuguseid aruandeid. Menüüvalik "Aruanded" (vt Joonis 8) sisaldab aruandeid, mille käivitamiseks tuleb klõpsata aruande nimetusele järgnevale viitele "vaata".

Aruande tulemustes on võimalik dialoogi nimele klõpsates avada vastav dialoog.

Dialogikorpuse tööpink	
<p><b>korpuse analüüs</b>  <b>korpuse muutmine</b>  <b>otsing</b>  <b>aruanded</b></p> <hr/> <p><b>dialoogi analüüs</b>  <b>dialoogi puhastamine</b>  <b>dialoog ajateljel</b>  <b>dialoogi XML kuju</b></p> <hr/> <p><b>kasutusjuhend</b></p> <hr/> <p><b>Reisiagent</b>  <b>salvestatud vestlused</b></p>	<p><b>Aruanded</b></p> <ol style="list-style-type: none"> <li>1. Vead transkriptsioonis (akt omaette real) - <b>vaata</b></li> <li>2. Vead transkriptsioonis (aktimärgendite vahel tekst) - <b>vaata</b></li> <li>3. Lausungid, mis on märgendatud enam kui ühe aktimärgendiga - <b>vaata</b></li> <li>4. Aktimärgendita lausungid - <b>vaata</b></li> <li>5. Aktide sagedustabel sorteeritud akti esinemiste järgi kahanevalt - <b>vaata</b></li> <li>6. Aktide sagedustabel sorteeritud akti nimetuse järgi kasvavalt - <b>vaata</b></li> <li>7. Aktide sagedustabel sort. akti nimetuse pikkuse järgi kahanevalt - <b>vaata</b></li> <li>8. Aktide sagedustabel sort. akti nimetuse pikkuse järgi kasvavalt - <b>vaata</b></li> <li>9. Aktid ja nendega seotud lausungid sorteeritud akti nimetuse järgi - <b>vaata</b></li> <li>10. Voorud, mis koosnevad mitmest aktist - <b>vaata</b></li> </ol>

**Joonis 8.** Aruannete aken.

Aruannete abil saab leida vigu transkriptsioonis. Kaks esimest aruannet näitavadki mõningaid vigu, mida transkribeerimisel inimliku eksimuse tõttu ette võib tulla. Teistes aruannetes on võimalik märgata dialoogiaktide kasutusvigu.

Aruanded annavad ka infot korpuses leiduvate dialoogide ning aktimärgendite kohta. Siiani on aruandeid kasutatud dialoogiaktide märgendusjuhendi täiustamisel.

Moodul on tehniliselt realiseeritud funktsioonina, mis saab parameetrina ette SQL päringu, aruande nimetuse ning andmebaasi nimetuse. Funktsiooni väljundiks on HTML tabel. Väljakutselt loob funktsioon ühenduse andmebaasiga ning käivitab parameetriga etteantud päringu (vt Näide 10). Päringu veergude nimetuste põhjal genereeritakse HTML tabeli päis ning päringu vastuse põhjal genereeritakse HTML tabel (vt.Joonis 9).

**Näide 10.** SQL päring, mis antakse HTML tabelit genereerivale funktsioonile sisendparameetrina

```
select
  c.nimi,
  a.vooru_nr,
  a.osaleja,
  a.tekst,
  a.aktid
from
  dialoogide_aktid a,
  (select dia_id, vooru_nr from dialoogide_aktid
   group by dia_id, vooru_nr
   having count(1) > 1) b,
  dialoogifailid c
where
  a.dia_id = b.dia_id and
  a.vooru_nr = b.vooru_nr and
  a.dia_id = c.id
order by
  a.dia_id,
  a.vooru_nr,
  a.rea_nr_voorus;
```

**Aruanne: Voorud, mis koosnevad mitmest aktist**  
Leiti 4525 rida.

nimi	vooru_nr	osaleja	tekst	aktid
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	5	A:	jah Mustamäe	KYJ: JAH    PPJ: LÄBIVIIMINE
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	5	A:	Aka'deemia tee.	IL: TÄPSUSTAMINE    PPJ: LÄBIVIIMINE
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	6	B:	aa,	VR: NEUTRAALNE INFO OSUTAMINE UUEKS    VR: PARANDUSE HINDAMINE
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	6	B:	sis=öö kõige='lihtsam on siit minna 'bussi peale kakskend 'kolm, kakskend=kolm=aa seitse'teist vist läb ka seitseteist='aa, et=ee need lähvad Kris'tiinesse,	KYJ: INFO ANDMINE
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	14	B:	number 'kaks ja 'neli.	KYJ: INFO ANDMINE
<a href="#">so_di_va_386_a3_tanav_valm03.txt</a>	14	B:	Maneezhi 'peatus kesklinnas	KYJ: INFO ANDMINE
<a href="#">in_di_te_97_a12_valm03.txt</a>	2	H:	ee tere.	RIE: TERVITUS
<a href="#">in_di_te_97_a12_valm03.txt</a>	2	H:	[kas] 'puukentsefaliidi vastu 'saab vaksineerida.	KYE: JUTUSTAV KAS

**Joonis 9.** Aruande mooduli poolt genereeritud tabel voorudest, mis sisaldavad mitut akti.

Antud aruandes on näha voorud, mis sisaldasid mitut akti.

Näiteks: dialoogis in\_di\_te\_97\_a12\_valm03.txt, voorus nr. 2 ütles osaleja H kõigepealt

➤ "ee tere", ==> | RIE: TERVITUS |

ning sama osaleja jätkas:

➤ "[kas] 'puukentsefaliidi vastu 'saab vaktsineerida" ==> | KYE: JUTUSTAV KAS |

Seega osaleja sooritas ühes voorus rohkem kui ühe akti.

Aruannete moodulisse on mugav lisada uusi aruandeid. Uue aruande lisamiseks on vaja kirjutada SQL päring ning paigutada kasutajaliidesesse vastava nimetusega aruanne, mille käivitamine annab funktsioonile parameetrina ette vastloodud päringu ja aruande nimetuse.

#### **4.1.6. Dialoogi puhastamine morfoloogilise analüüsi tarvis**

Üksikdialoogidele saab rakendada teisendusmoduleid: dialoogi puhastamine morfoloogilise analüüsi tarvis, dialoogi paigutamine ajateljele ja dialoogi viimine XML kujule.

Dialoogi puhastamine morfoloogilise analüüsi tarvis (vt Joonis 10) eemaldab dialoogist transkriptsioonielemendid ja aktimärgendid, et dialoogi saaks anda sisendiks morfoloogilisele analüsaatorile.



Joonis 10. Dialoogi puhastamine morfoloogilise analüüsi tarvis.

Enne dialoogi suunamist morfoloogilisse analüüsi on dialoog vaja puhastada märgenditest, mis segavad morfoloogilist analüsaatorit. Näiteks tuleb eemaldada võrdusmärgid, mis tähistavad kokkuhääldamisi ning ühendavad omavahel kokkuhääldatud sõnu (näiteks Narva=maanteed). Lisaks tuleb eemaldada pealerääkimised ja rõhuliste sõnade tunnused, mis lõhestavad sõnu (näiteks: käigu['kastist]). Lõhestatud sõnad ning kommentaarides ja aktimärgendites sisalduvad sõnad on morfoloogilise analüsaatori sisendiks sobimatud. Morfoloogiline analüsaator peab sisendiks saama lause, kus sõnad on eraldatud tühikute või kirjavahemärkidega. Kommentaarid (mis vastavalt transkriptsioonikeelele paiknevad topeltsulgude vahel, nt. ((taustal oli müra))) ning aktimärgendid (nt. | KYE:AVATUD|) ei vaja morfoloogilist analüüsi.

Puhastamise käigus eemaldatakse transkriptsioonelemendid, kommentaarid ja dialoogiaktide märgendid. Puhastatud dialoog on sobivaks sisendiks morfoloogilisele analüsaatorile

### Mooduli sisendiks on puhastamata dialoog:

```
V: tere=päevast.      | RIE: TERVITUS |
H: tere?             | RIJ: VASTUTERVITUS |
(1.0) sooviksin=    | DIE: SOOV |
V: =kuulen?        | DIJ: MUU |
```



(1.0)

V: hallo? | KKE: ALGATUS |  
H: hallo? | KKJ: KINNITAMINE |

### Mooduli väljundiks on puhastatud dialoog:

V: <s> tere päevast. </s>  
H: <s> tere sooviksin </s>  
V: <s> kuulen </s>  
V: <s> hallo </s>

Puhastusreeglid, mida programm kasutab, on defineeritud järgnevalt (rakendamise järjekorras):

- asenda aktimärgend tühikuga
- asenda paus tühikuga
- asenda kommentaar tühikuga
- asenda tabulaator tühikuga
- kustuta rõhulisust tähistav ülakoma
- kustuta väljahingamine sõna lõpus =h näiteks: tule=h
- asenda võrdusmärk tühikuga
- kustuta nn. ebaselgused {---}, {-}, {-}:, kuid ära kustuta ebaselget juttu {jutt}
- kustuta naer sõna seest (h) näiteks: s(h)õna
- kustuta ülearused märgid { , } , [ , ] , ? , £ , @ , ¤ , & , % , \$ , # , \* , < , >
- kustuta koolon, kui ta tähistab venitust, kuid jäta koolon alles osalejatunnusele
- eemalda järjestikused korduvad tühikud
- eemalda kõik, mis jääb ettepoole kõige esimest osalejatunnust
- eemalda aktide märgendamisest tekkinud liigsed voorusisesed reavahetused
- pane iga kõneleja jutt märkide <s> ja </s> vahele näiteks: A: <s> kõne </s>

Peale puhastusreeglite rakendamist saab dialoogi suunata morfoloogilisse analüsaatorisse ESTMORF [17] ning tulemuseks saadakse morfoloogiliselt analüüsitud dialoog (vt Näide 6 jaotises 3.1.3).

#### 4.1.7. Dialoogi paigutamine ajateljele

Ajateljele paigutatud dialoogi on mugavam käsitsi morfoloogiliselt ja süntaktiliselt märgendada või automaatset märgendust üle kontrollida, sest ühe osaleja lausungid asuvad

ühel real. Paigutus ajateljele annab ka visuaalset tagasidet pealerääkimiste paiknevuse osas. Dialoogi esitus ajateljel sobib edaspidi ka helifaili lisamiseks dialoogi transkriptsioonile.



Joonis 11. Dialoogi paigutamine ajateljele.

### Mooduli sisendiks on dialoog tavapärasel ülalt-alla kujul:

```
V:      ´Estmar=´info,      | KUTSUNGI VASTUVÕTMINE | | ESITLUS |
´Leenu=kuuleb | ESITLUS |
[tere] | TERVITUS |
H:      [tere] | VASTUTERVITUS |
(0.5) halloo? | KONTAKTI KONTROLL |
V:      tere? | KONTAKTI KINNITAMINE | | TERVITUS |
```

### Mooduli väljundiks on ajateljele paigutatud dialoog:

	1	2	3
V:	´Estmar=´info, ´Leenu=kuuleb [tere]		tere?
H:		[tere] (0.5) halloo?	

Joonis 12. Dialoog ajateljel.

Ajateljele paigutatud dialoogi on kujutatud tabelina (vt Joonis 12). Iga osaleja jaoks on tabelis üks rida. Esimesele reale paigutatakse selle osaleja voor, kes alustas dialoogi. Osalejate poolt

öeldud lausungid on kollase taustaga lahtrites. Lisaks on tabelil päiserida, mis sisaldab vooru numbrit.

Kollastes lahtrites olevas tekstis on säilitatud ka transkriptsioonimärgendid, millest huvitavamad on pealerääkimismärgid. Antud näites on näha, et kasutajad ütlesid sõna "tere" samaaegselt. Pealerääkimiste eripaigutust ajateljel esialgu ei ole käsitletud. Pealerääkimised on värvidega esile toodud. Edasises plaanis on ajateljele paigutust täiendada nõnda, et pealerääkimised paikneksid üksteise all.

### **Algoritm dialoogi paigutamiseks ajateljele:**

- asenda aktimärgend tühikuga
- asenda tabulaator tühikuga
- asenda järjestikused korduvad tühikud ühe tühikuga
- eemalda kõik, mis jääb ettepoole esimest osalejatunnust
- eemalda aktide märgendamisest tekkinud voorusisesed reavahetused
- moodusta maatriks dialoog[osaleja, kõnevoor] = "osaleja tekst vastavas kõnevoorus"
- väljasta maatriks ridahaaval, värvides mitte-tühjad elemendid kollaseks, tühjad halliks
- lisaks ajateljele viidud dialoogile, väljastatakse ka dialoog esialgsel kujul

#### 4.1.8. Dialoogi viimine XML kujule

Praegu on dialoogikorpuses iga dialoog lihtsalt tekstifail. Mitmete märgendustasandite olemasolu muudab ühetasandilisse tekstifaili üleskirjutatud dialoogid arvutil raskesti töödeldavaks. Rahvusvaheline standardiseerimine TEI (Text Encoding Initiative) on korpuste loomisel ja märgendamisel võtnud kasutusele SGML keele, mille üks populaarseim alaliik on XML. Sellepärast on ka käesolevas tööpingis võetud kasutusele XML formaat ning loodud eksperimentaalne moodul (vt Joonis 13), mis annab võimaluse mitme märgendustaseme salvestamiseks ühte struktuursesse faili.

Dialoogikorpuse tööpink	
<p><a href="#">korpuse analüüs</a> <a href="#">korpuse muutmine</a> <a href="#">otsing</a> <a href="#">aruanded</a></p> <hr/> <p><a href="#">dialoogi analüüs</a> <a href="#">dialoogi puhastamine</a> <a href="#">dialoog ajateljel</a> <a href="#">dialoogi XML kujule</a></p> <hr/> <p><a href="#">kasutusjuhend</a></p> <hr/> <p><a href="#">Reisiagent</a> <a href="#">salvestatud vestlused</a></p>	<p><b>Dialoogi viimine XML kujule</b></p> <p>Dialoog: <input type="text"/> <input type="button" value="Browse..."/></p> <p><input type="button" value="Genereeri dialoogi XML kujule"/></p>

Joonis 13. Dialoogi viimine XML kujule

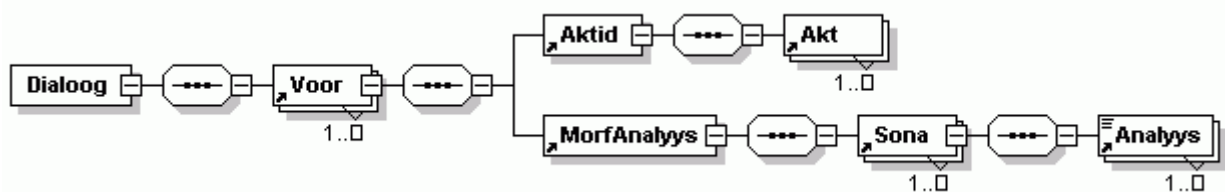
XML dokumentide vaatamiseks ning hierarhias liikumiseks sobib suvaline veebilehitseja. XML dokumentide haldamiseks on olemas hulgaliselt tarkvara. Samas ei ole korpusega töötaval inimesel mugav dialooge koheselt XML kujul tekitada. Selleks ongi loodud moodul, mis teisendab tavalised ühetasandilisse tekstifaili üleskirjutatud dialoogid XML formaati.

Hiljem on XML parserite abil võimalik automatiseerida XML dialoogide viimine näiteks relatsioonilisse andmebaasi, mis pakub mugavaid võimalusi korpuse automaattöötamiseks (otsing, mitmesugused analüütilised vaated).

Koostatud XML dokumentide loomiseks ning valideerimiseks kasutatav DTD skeem ja selle esitus diagrammina on järgmine (vt Joonis 14).

(Vooru all mõeldakse ühe osaleja poolt lausunud teksti kuni järgmise osaleja tekstini. Tegelikult on kasutusel ka nn pikad voorud, mis ei "katke", kui teine osaleja sooritab vaid pealerääkimisi, kuid ülesande lihtsustamise huvides me neid ei vaatle.)

```
<?xml version="1.0" encoding="utf-8"?>
<!ELEMENT Dialoog (Voor+)>
<!ELEMENT Voor (Aktid, MorfAnalyys)>
<!ATTLIST Voor
  osaleja CDATA #REQUIRED
  tekst CDATA #REQUIRED
>
<!ELEMENT Aktid (Akt+)>
<!ELEMENT Akt EMPTY>
<!ATTLIST Akt
  tekst CDATA #REQUIRED
  aktid CDATA #REQUIRED
>
<!ELEMENT MorfAnalyys (Sona+)>
<!ELEMENT Sona (Analyys+)>
<!ATTLIST Sona
  originaal CDATA #REQUIRED
>
<!ELEMENT Analyys (#PCDATA)>
```



**Joonis 14.** XML dokumendi DTD skeemi diagramm

XML dokumendi DTD skeemis kirjeldatakse, et dialoog võib koosneda mitmest voorust, ning iga vooru jaoks luuakse kaks alajaotist **Aktid** ja **MorfAnalyys**. Jaotises **Aktid** on element **Akt**, mis moodustatakse lausungist ja seda märgendavast aktimärgendist. Jaotises **MorfAnalyys** on element **Sona**, millel võib olla mitu alamelementi **Analyys**. Kuigi morfoloogilise analüüsi järel rakendatakse ka morfoloogilist ühestamist, jäävad tavaliselt ikkagi sisse mõningad mitmesused ning ühele sõnale jääb mitu analüüsi.

## Näide 8. Dialoogi XML kujul

Dialoog tavapärasel kujul:

V: Estmar='info, | KUTSUNGI VASTUVÕTMINE | | ESITLUS  
'Leenu=kuuleb | ESITLUS |  
tere | TERVITUS |

Dialoog XML kujul:

```
<?xml version="1.0" encoding="utf-8" ?>
<Dialog>
<Voor osaleja="V" tekst="Estmar='info, 'Leenu=kuuleb tere">
<Aktid>
  <Akt tekst="Estmar='info," aktid="| KUTSUNGI VASTUVÕTMINE | | ESITLUS |" />
  <Akt tekst="Leenu=kuuleb" aktid="| ESITLUS |" />
  <Akt tekst="tere" aktid="| TERVITUS |" />
</Aktid>
<MorfAnalyys>
<Sona originaal="Estmar">
  <Analyys>####</Analyys>
</Sona>
<Sona originaal="info">
  <Analyys>info+0 //_S_ sg n, //</Analyys>
</Sona>
<Sona originaal="Leenu">
  <Analyys>Leenu+0 //_H_ sg n, //</Analyys>
</Sona>
<Sona originaal="kuuleb">
  <Analyys>kuul+b //_V_ b, //</Analyys>
</Sona>
<Sona originaal="tere">
  <Analyys>tere+0 //_S_ sg n, //</Analyys>
</Sona>
</MorfAnalyys>
</Voor>
</Dialog>
```

Ülaltoodud näites on näha ka morfoloogiline analüüs, mis lisandus XML kujule teisendusel automaatselt. Selleks on kasutatud Tartu Ülikooli ja Filosoofi poolt loodud morfoloogilist analüsaatorit ja ühestajat [17]. Enne morfoloogilise analüüsi rakendamist on tekst ka puhastatud aktimärgenditest (vt. 4.1.6 Dialoogi puhastamine morfoloogilise analüüsi tarvis).

Ühestamine on tinglik, st. mõnele sõnale jääb ka peale automaatset ühestamist mitu analüüsi. Neid saab käsitsi ühestada.

XML kujule viidud dialooge suulise kõne korpuse haldajad ise otseselt ei kasuta. Moodul on praegu pigem eksperimentaalne ning leiab tegelikku kasutust hetkel vaid dialoogi viimisel relatsioonilisse andmebaasi.

XML kujul olevasse dialoogi saab edaspidi vajaduse korral mugavalt lisada uusi märgendustasemeid: süntaktiline, pragmaatiline.

Kui XML kujule viidud dialooge peaks olema vaja muuta, siis sobib selleks kõige paremini mõni XML tekstide toimetamiseks mõeldud programm (vt Joonis 15), näiteks XMLSPY.<sup>1</sup>

The image shows three instances of the XMLSpy interface, each displaying a different XML node for morphological analysis. Each instance has a 'Sona' list and an 'Analüüs' table.

**Instance 1 (Top):** Sona (6)

	= originaal	Analüüs
1	e	e+0 //_Y_?, //
2	sees	sees+0 //_K_//
3	ei	ei+0 //_V_ neg, //
4	ole	ole+0 //_V_ o, //
5	hh	####
6	.	._Z_//

**Instance 2 (Middle):** Sona (5)

	= originaal	Analüüs						
1	on	Analüüs (2) <table border="1"> <thead> <tr> <th></th> <th>Rbc Text</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>ole+0 //_V_ b, //</td> </tr> <tr> <td>2</td> <td>ole+0 //_V_ vad, //</td> </tr> </tbody> </table>		Rbc Text	1	ole+0 //_V_ b, //	2	ole+0 //_V_ vad, //
	Rbc Text							
1	ole+0 //_V_ b, //							
2	ole+0 //_V_ vad, //							
2	võimalik	Analüüs (1)						
3	tellida	Analüüs (1)						
4	sis	Analüüs (1)						
5	.	Analüüs (1)						

**Instance 3 (Bottom):** Sona (2)

	= originaal	Analüüs
1	jah	jah+0 //_D_//
2	.	._Z_//

Joonis 15. XMLSPY aken XML dokumendi töötlemiseks.

Joonisel on näha, kuidas saab avada sobiva koha XML dokumendist ning muuta seal näiteks morfoloogilise analüüsi tulemust sõnal, mis sai esialgselt kaks analüüsi. Sõnale "on" on morfoloogiline ühestaja jätnud kaks analüüsi. XMLSPY abil saab avada haru Analüüs ning teha aktiivseks lahtri selle analüüsiga, mida soovitakse parandada või kustutada.

<sup>1</sup> <http://www.xmlspy.com>

## 4.2. Dialoogiüksuste eraldamine

Hääduspäraselt üleskirjutatud dialoogid, mis on ka dialoogiaktidega märgendatud, sisaldavad mitmeid elemente e. dialoogiüksusi, mida peab automaattöötlemises suutma eristada. Näiteks tuleb eristada kommentaarid, aktimärgendid, pealerääkimised, pausid.

Üksuste eraldamiseks dialoogides tuli alustada grammatika defineerimisest. Grammatika on defineeritud kui regulaaravaldiste hulk, mis kirjeldavad dialoogides esinevaid üksusi.

Kui transkriptsioonikeele grammatika oli regulaaravaldiste abil defineeritud, siis võis alustada parsimisega. Peale mõningaid katseid selgus, et parsimiseks on mõistlik viia dialoog ühte tekstimuutujasse ning loobuda dialoogide töötlemisest üksikutel ridadel paikneva tekstina, mida tuleb ridahaaval failist lugeda. Ühtses tekstimuutujas paiknev dialoog on lihtsamini töödeldav, sest regulaaravaldiste süntaktilised laiendused võimaldavad käsitleda ka ühes muutujas paiknevat nõ. mitmerealist teksti. Regulaaravaldisi seetõttu palju muutma ei pidanud, tuli käsitleda nende rakenduspiire arvestades reavahetusi.

### Regulaaravaldiste abil defineeritud grammatika

\* aktimärgendid:

$$\backslash | \cdot * ? \backslash |$$

Antud regulaaravaldis leiab minimaalselt kõik, mis jääb kahe püstkriipsu vahele, püstkriipsud kaasaarvatud. "Minimaalselt" tähendab, et kahe püstkriipsu vahel teisi püstkriipse ei esine.

\* kõik aktimärgendid ühel real:

$$\backslash [ \wedge \backslash n ] * \backslash |$$

Antud regulaaravaldis leiab kõik, mis jääb kahe püstkriipsu vahele (püstkriipsud kaasa arvatud) ning ei sisalda reavahetusi.

Ühel real võib olla mitu aktimärgendit. Neid on teatud juhtudel vaja käsitleda ühise märgendina. Näiteks aktide järgnevuste sagedustabeli leidmisel käsitletakse samal real paiknevate märgendite järjendit bigrammi ühe liikmena.

\* kommentaarid:

$$(\backslash ( \{ 2 \} \cdot * ? \backslash ) \{ 2 \} | \backslash \{ \{ 2 \} \cdot * ? \backslash \} \{ 2 \} )$$



Kommentaariid on üldjuhul topeltsulgude vahel. Kohati on kasutatud ka loogelisi sulge.

Antud regulaaravaldis leiab kõik, mis on kahekordsete sulgude/ kahekordsete loogeliste sulgude vahel, näiteks ((kommentaar)).

\* osalejatunnused:

$^{\wedge}[\text{ABCDEFGHIJKLMNÖPQRSTUVÖÄÖÜWXYZ}]{1}\cdot*?:$

Osalejatunnuseks on kõik, kuni esimese koolonini, välja arvatud tekst, mis jääb ettepoole esimesest osalejatunnusest. Osalejatunnuse algustäht on alati suurtäht. Osalejatunnuses ei ole lubatud tühikud.

\* tekst enne esimest osalejatunnust (esimene sulgavaldis) ja ülejäänud dialoog (teine sulgavaldis):

$(^{\wedge}[^{\wedge}\backslash:]^{\wedge}n)([\text{ABCDEFGHIJKLMNÖPQRSTUVÖÄÖÜWXYZ}]{1}[^{\wedge}\backslashn]^{\wedge})*?:)$

Enne esimest osalejatunnust on dialoogides taustakirjeldusele viitavaid märke või transkribeerija märkmeid. Antud avaldises on näha, et esimesele osalejatunnusele ei tohi eelneeda kooloneid.

Reavahetuste käsitluseks tuleb antud avaldise korral kasutada vastavat regulaaravaldiste modifikaatorit 's'.

\* pausid:

$(\backslash(\backslash\backslash)|\backslash(\backslash\backslash\backslash\backslash)|\backslash([0-9]^{\wedge}+\backslash.[0-9]^{\wedge}+\backslash)|\backslash([0-9]^{\wedge}+\backslash))$

\* mikropaus:

$\backslash(\backslash\backslash)$

Punkt sulgude vahel (.).

\* pikk paus:

$(\backslash\backslash\backslash\backslash)$

Kolm punkti sulgude vahel (...).

\* mõõdetud paus:

$(\backslash([0-9]^{\wedge}+\backslash.[0-9]^{\wedge}+\backslash)|\backslash([0-9]^{\wedge}+\backslash))$

Pausi pikkus sekundites sulgude vahel. Võib sisaldada komakohti.

\* osalejad (esimene sulgavaldis) ja lausungid (teine sulgavaldis):

$(^{\wedge}[\text{ABCDEFGHIJKLMNÖPQRSTUVÖÄÖÜWXYZ}]{1}[^{\wedge}]^{\wedge})*?:)(\cdot*\$)$

Esimeses pooles korratakse osalejatunnuse definitsiooni ning avaldise teine sulgudega ümbritsetud osa hõlmab kõik kuni reavahetuseni. Kasutada tuleb regulaaravaldiste modifikaatorit 'm', et käsitlus oleks mitmerealisust toetav. Eeltööna

- tuleb eemaldada tekst, mis jääb ettepoole esimest osalejatunnust
- aktide märgendamisele tekkinud liigsed voorusisesed reavahetused tuleb asendada abisümboliga

Grammatika koostamisel ilmnes ka puudujääke suulise kõne rühma poolt kasutatavas transkriptsioonikeeles. Nimelt on mõned formalismid kaheselt mõistetavad.

Vaadeldgem juhtu, kus dialoogi transkriptsioon näeb välja järgnevalt:

A: kas teie olete doktor Õunapuu | AKTIMÄRGEND |

B: tere päevast | AKTIMÄRGEND |

Õunapuu: on tõesti mu nimi | AKTIMÄRGEND |

Viimases reas ei ole aru saada, kas tegemist on venitusega sõna Õunapuu lõpus või on tegemist osalejatunnusega Õunapuu.

Selgub, et pärisnimed tuleks transkribeerida väiketähtedega.

Lisaks võis osalejatunnus sisaldada tühikuid, ning siis tekiks viga juhul, kui venitatakse suvalist sõna peale Õunapuud.

A: kas teie olete doktor Õunapuu | AKTIMÄRGEND |

B: tere päevast | AKTIMÄRGEND |

Õunapuu on tõesti mu: nimi | AKTIMÄRGEND |

Sellisel juhul loetakse osalejatunnuseks "Õunapuu on tõesti mu".

Osalejatunnustes tühikute kasutamine on seetõttu lubamatu.

Grammatika koostamine aitas taolisi probleeme märgata. Osalejatunnustes loobuti tühikute kasutamisest ning reavahetuste lisamisel jälgiti, et pärisnimi, mis lõpeb venitusega, ei satuks esimese sõnana uuele reale.

Grammatika reeglite rakendamisel tuli jälgida ka konteksti. Reavahetus ei eelnenud mitte ainult osalejatunnusele, vaid seda kasutati ka aktide märgenduse järel. Neid kahte reavahetust tuli eristada konteksti alusel ning üks nendest asendada töötluse hõlbustamiseks ajutise abisümboliga.

Venituste lugemisel, mida tähistatakse kooloniga, tuli eelnevalt veenduda, et kontekstiks on puhas lausung, sest osalejatunnuse koolonid ja kommentaarides esinevad koolonid ei tule venituste lugemisel arvesse.

### **4.3. Dialoogikorpuse tööpingi kasutusjuhend**

Alljärgnevalt on toodud dialoogikorpuse tööpingi kasutusjuhend, mida saab avada ka tööpingi menüüst.

#### **Juhend**

Antud tööpink võimaldab dialooge/dialoogikorpust analüüsida, aitab leida transkriptsiooni- ja märgendusvigu, sooritada otsingut üle korpuse ning rakendada dialoogidele mõningaid teisendusi.

#### *Korpuse analüüs*

Korpusesse lisatud dialoogidest on võimalik valida üks või mitu dialoogi ning rakendada neile analüüsimooduleid: nähtuste loendamine ja aktide järgnevuste leidmine. Analüüsitavad dialoogid tuleb märkida nimistus valituks, valida nimistu alt soovitud tegevus ning käivitada analüüs.

Nähtuste loendamine seisneb mõningate transkriptsioonelementide ning aktimärgendite loendamises. Väljund kuvatakse uues aknas ning seal dialooginimele klõpsates on võimalik vaadata detailset analüüsi ning dialoogi sisu.

Aktide järgnevuste leidmine väljastab dialoogiaktide järgnevuste sagedustabeli. Samale reale paigutatud dialoogiaktide järjend loetakse bigrammi üheks liikmeks.

#### *Korpuse muutmise*

Dialoogide lisamine/eemaldamine: Dialoogide analüüsimiseks on võimalik lisada dialoogid korpusesse. Mitmekaupade lisamiseks saab kasutada ZIP faili. Dialoogi lisamisel kirjutatakse samanimeline dialoog korpuses üle. Lisatava dialoogifaili või ZIP faili suurim lubatud suurus on 500K. Dialoogide valikulist kustutamist ei ole käesolevaks hetkeks realiseeritud ning kustutada saab vaid kogu korpuse korraga.

### *Otsing*

Korpusesse lisatud dialoogides on võimalik sooritada otsingut. Otsingut saab rakendada üle tunnuste tekst (kõne) ja akt (aktimärgendid). Otsida saab ka transkriptsioonelemente. Näiteks sisestades ülakoma, otsitakse rõhulisi sõnu; sisestades kooloni, otsitakse venitusi; sisestades võrdusmärgi, otsitakse ning näidatakse kõik kokkuhääldamised. Otsitav märgijada on väljundis värvitud punaseks tunnuse tekst jaoks, tunnuse akt puhul märgijada ei värvita.

Otsingutulemustes on võimalik dialoogi nimele klõpsates avada vastav dialoog.

### *Aruanded*

Korpusesse lisatud dialoogide põhjal on võimalik genereerida mitmesuguseid aruandeid. Menüüvalik "Aruanded" sisaldab aruandeid, mille käivitamiseks tuleb klõpsata aruande nimetusele järgnevale viitele "vaata".

Aruande tulemustes on võimalik dialoogi nimele klõpsates avada vastav dialoog.

Menüüvalikud:

*dialoogi analüüs*

*dialoogi puhastamine*

*dialoog ajateljel*

*dialoogi XML kujule*

Üksikutele dialoogidele saab rakendada analüüsi e. nähtuste loendamist ning teisendusmooduleid: puhastamine, ajateljele paigutamine ja XML kujule viimine.

Dialoogi analüüs seisneb mõningate transkriptsioonelementide ning aktimärgendite loendamises. Väljund kuvatakse uues aknas.

Dialoogi puhastamine morfoloogilise analüüsi tarvis eemaldab dialoogist transkriptsioonelemendid ja aktimärgendid, et dialoogi saaks anda sisendiks morfoloogilisele analüsaatorile.

Dialoogi paigutamine ajateljele on eksperimentaalne moodul ning pakub võimaluse viia dialoog ajateljele, mis võiks olla vajalik tulevikus, kui dialoogid seotakse ning sünkroniseeritakse helifailidega.

Dialoogi viimine XML kujule on katse alustada dialoogide märgendustasemete kasvades dialoogide hoidmist struktuurses XML failis. See annab võimaluse mitme märgendustaseme salvestamises ühte faili. Hetkel on realiseeritud aktimärgendite ja morfoloogiliste märgendite struktuurne kujutamine, kusjuures morfoloogiline märgendus lisatakse mooduli rakendamisel automaatselt koos statistilise ühestamisega.

Tööpingi menüüst leiab ka viite dialoogsüsteemi Reisiagent viimasele veebipõhisele versioonile ning selle dialoogsüsteemiga kogutud dialoogidele.

#### **4.4. Konverteerimismoodulite automaatne loomine**

Tartu Ülikooli Arvutiteaduse instituudis on loodud ka tarkvaratoode AMADEUS, mis võimaldab automatiseerida konverteerimismoodulite loomist [16]. Loodud abivahendit saaks edukalt kasutada ka dialoogidele muu sobiva visuaalse või struktuuraalse kuju andmiseks.

Konverteerimismoodulite automaatne loomine süsteemis AMADEUS koosneb järgnevatest etappidest:

- koostatakse lähtekeele grammatika,
- grammatika põhjal genereeritakse lähtekeele parsere, mis tuvastab lähtekeele üksused,
- koostatakse üleminekumudel lähtekeelest sihtkeelele,
- rakendatakse üleminekut.

Saadud tulemust on sihtkeeles on võimalik esitada mitmel moel:

- skeemesitusena,
- puuna,

- struktuurse tekstifailina (XML).

See süsteem on väga mugava kasutajaliidesega ning skeemesitus võimaldab tulemust mugavalt parandada ja kujundada.

Seda programmi saab võrrelda antud magistritöös loodud mooduliga, mis viib dialoogi XML kujule. Selle programmi eeliseks on see, et seal saab kasutaja parsitud dialoogile mugavalt täiendusi teha ning siis salvestada selle oma äranägemise järgi defineeritud XML struktuuri.

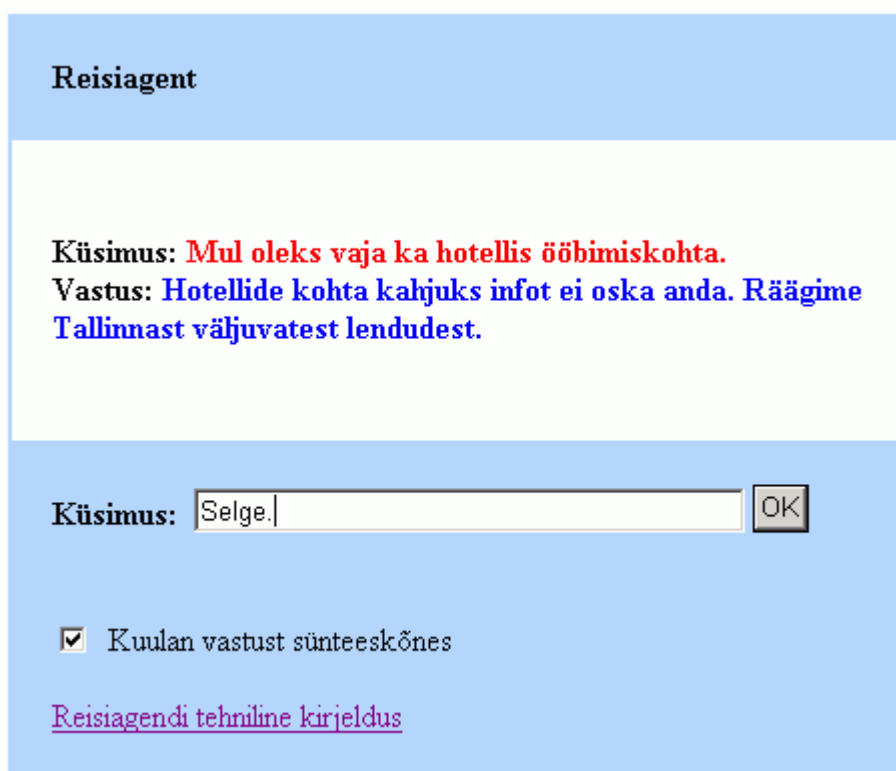
Magistritöös loodud teisendusmoodul ei võimalda kasutajal valida, milline peaks loodav XML fail välja nägema. Küll aga lisab magistritöös loodud programm automaatselt ka morfoloogilise märgenduse, mis võib mõnel juhul päris mugavaks osutuda.

Tarkvaraarendussüsteemi AMADEUS on edaspidi plaanis kasutada dialoogiaktide grammatika formaliseerimiseks ning grammatika abil parsitud dialoogide märgendite järeltötluseks ja sobivate XML kujude genereerimiseks.

## 5. Dialoogsüsteem Reisiagent

Dialoogikorpuse tööpingi menüüs leidub ka viide dialoogsüsteemi Reisiagent viimasele veebipõhisele versioonile (vrd bakalaureusetöö [1]) ning selle dialoogsüsteemiga kogutud dialoogidele. Samas on Reisiagent siiski tööpingist eraldiseisev osa, sest ta ei võimalda rakendada teisenusi ega tee ka iseseisvalt analüüsi. Seetõttu kirjeldame Reisiagenti eraldi peatükis, väljaspool tööpingi moodulite kirjeldusi.

Dialoogsüsteem Reisiagent on realiseeritud veebipõhisena (vt Joonis 16) ning töötab kõikide uuemate veebilehitsetajatega. Valitud platvorm on vabavaraline. Veebiserveriks on Apache ning programmeerimiskeeleks on PHP.



Joonis 16. Dialoogsüsteemi Reisiagent tööaken.

Süsteem suhtleb üle veebi kolme välise süsteemiga:

- eesti keele morfoloogiline analüsaator (<http://kiisu.eki.ee/>),
- eesti keele morfoloogiline süntesaator (<http://www.filosoft.ee/>),
- eestikeelse kõne süntesaator (<http://kiisu.eki.ee/>).

Veebipõhiste pöördumiste eeliseks on see, et alati on kasutada tarkvara viimane versioon: näiteks parandatakse kõnesünteesi kvaliteeti või täiustatakse morfoloogilist analüsaatorit ning Reisiagent saab kohe sellest osa. Puuduseks on see, et veebipõhised teenused ei ole hallatavad antud süsteemi autori poolt ning võivad muutuda ülesehituselt mitteühilduvaks, minna üle uuele aadressile või olla serveri rikete tõttu ligipääsmatud. Selliste olukordadega saab Reisiagent hakkama ning jätab lihtsalt need teenused kasutama. Sellisel juhul ei leia Reisiagent sõnadest algvorme ega nimisõnu ning leidmata sobivat vastust nende abita, vastatakse lihtsalt "Kahjuks ei saanud teist aru".

Lisaks on Reisiagenti integreeritud Filosoofi morfoloogiline ühestaja, mis paikneb dialoogsüsteemiga samas serveris.

## 5.1. Dialoogsüsteemi Reisiagent moodulite kirjeldused

### 1. Morfoloogiline analüüs

Morfoloogilise analüüsi mooduli abil leitakse kasutaja poolt sisestatud lausest sõnade algvormid. Eesti Keele Instituudis loodud veebipõhisele morfoloogilisele analüsaatorile (<http://kiisu.eki.ee/>) saadetakse kasutaja poolt sisestatud lause. Vastuseks saadakse morfoloogiliselt analüüsitud lause, millest dialoogsüsteem moodustab unikaalsete algvormide loendi. Seda loendit kasutab dialoogsüsteem hilisemas võtmesõnade otsingus, et leida kasutaja küsimusele vastus (vt Näide 11).

**Näide 11.** *Morfoloogilise analüsaatori poolt leitud algvormi kasutamine*  
Küsimus: Kust lennuk väljub?

Vastus: Lennud väljuvad Tallinna lennujaamast.

#### **Kommentaar näitele:**

Sihtkohta, kohanime ega tuttavat võtmesõna ei leidunud. Küsimus lemmatiseeriti morfoloogilise analüsaatori abil. Lemmatiseeritud sõnade hulgas oli 'väljuma' ning selle vormi kohta oli juba võtmesõnade andmebaasis vastus olemas.

### 2. Morfoloogiline süntees

Morfoloogilise sünteesi moodulit kasutatakse sõnavormide genereerimiseks, sest teatud juhtudel soovitakse dialoogsüsteemis kasutajale vastamisel kasutada nimisõna mitmuse



osastavas käändes. Filosofti veebipõhisele morfoloogilisele süntesaatorile (<http://www.filosoft.ee/>) saadetakse nimisõna, mida soovitakse käänata. Vastuseks saadakse sõna soovitud käändes. Saadud sõna ning sobiva lausemalli abil moodustatakse vastus kasutajale (vt Näide 12).

**Näide 12.** *Morfoloogiliselt sünteesitud vormi kasutamine*

Küsimus: Tahtsin pannkoogijahu osta.

Vastus: Pannkoogijahude kohta kahjuks infot ei anna. Räägime Tallinnast väljuvatest lendudest.

### **Kommentaar näitele:**

Sihtkohti, kohanimesisid, võtmesõnu, lemmatiseeritud võtmesõnu ei leidunud. Morfoloogiline analüüs koos ühestamisega eraldas nimisõnad. Nimisõnade hulka jäi "pannkoogijahu". Morfoloogiline süntees genereeris nimisõnast 'pannkoogijahu' mitmuse osastava. Vastati, et 'Pannkoogijahude kohta info puudub'.

### **3. Morfoloogiline ühestamine**

Morfoloogilist ühestajat kasutatakse dialoogsüsteemis selleks, et eemaldada mitmesus morfoloogilisest analüüsist. Dialoogsüsteem kasutab ühestajat, et saada selliste nimisõnade loend, mis oleksid kindlasti nimisõnad ning mitte omadussõnad ega verbivormid (näiteks "naist peeti targaks" – siin tuleb ühestajal tuvastada, et tegemist pole nimisõnaga *peet*, vaid verbiga *pidama*). Leitud nimisõnu saab hiljem vastuse moodustamisel kasutada. (vt Näide 13)

**Näide 13.** *Leitud nimisõna kasutamine vastuse genereerimisel*

Küsimus: Nagu seinaga räägiks!

Vastus: Seinte kohta kahjuks infot ei anna. Räägime Tallinnast väljuvatest lendudest.

### **4. Kõnesüntees**

Kõnesünteesi kasutatakse selleks, et kasutajale vastus kõnena kõlarite kaudu ette mängida. TTÜ küberneetika instituudi ja EKI koostöös loodud veebipõhisele kõnesünteesimoodulile (<http://kiisu.eki.ee/>) saadetakse kasutajale vastamiseks moodustatud lause. Kõnesüntesaator moodustab sellest helifaili, mis kasutaja arvutis audiovahendite olemasolul ette mängitakse. Kõnesünteesi mooduli kasutuse saab kasutaja sisse/välja lülitada.

## 5. Kasutajalt saadud küsimusele vastuse leidmine

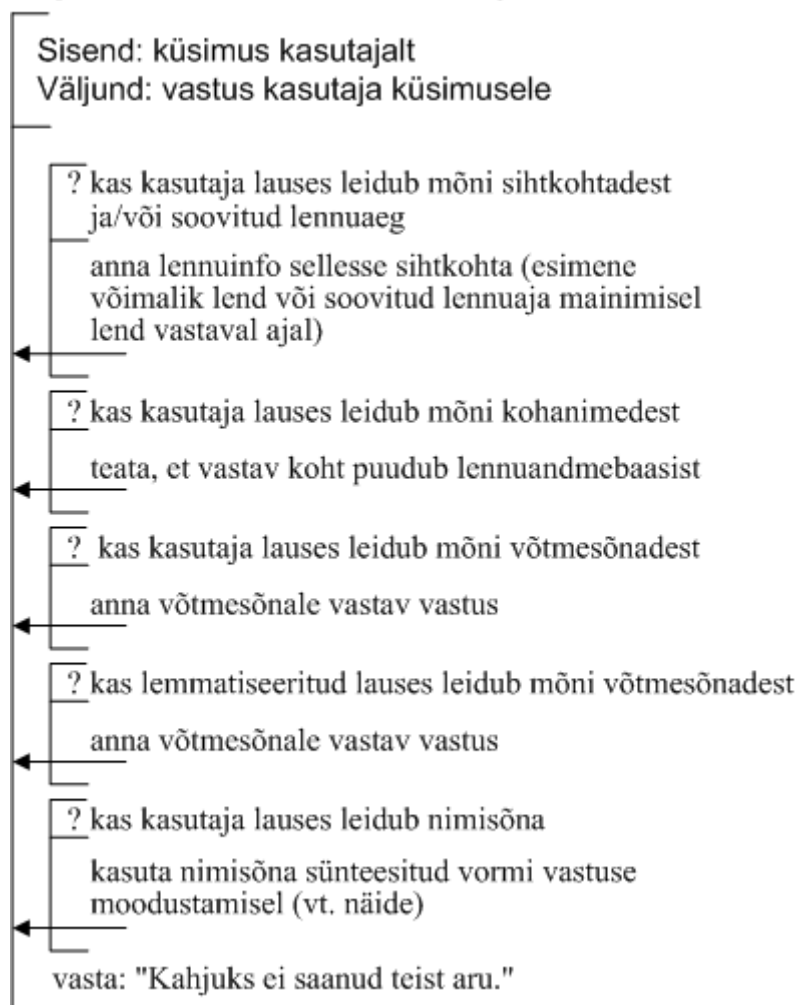
Kasutaja poolt esitatud küsimusele vastuse leidmiseks rakendatav algoritm kasutab eelkirjeldatud mooduleid ning toimub joonisel (vt. Joonis 17) kirjeldatud algoritmi kohaselt:

**Sisend:** tekst, lendude ja vastuste massiivid

(Lendude massiiv on staatiline sihtkohtade ja lennuaegade andmebaas; vastuste massiiv on staatiline vastuste andmebaas, kus igale võtmesõnale on defineeritud teatud hulk vastuseid, vrd [1].)

**Väljund:** vastus kasutajale, mis võib sisaldada infot soovitud lennu kohta või võtmesõnade läbivaatusel leitud vastust või leitud nimisõnade alusel genereeritud vastust.

Algoritm: vastuse leidmine kasutaja küsimusele



Joonis 17. Dialoogsüsteemi Reisiagent vastuse leidmise algoritm kasutaja küsimusele.

Reisiagent saab aru ka kuupäevadest, nädalapäevadest ja ajamäärsõnadest (täna, homme, ülehomme), suudab meeles pidada kõnealuse sihtkoha kuni järgmise sihtkoha mainimiseni ning kasutada seda, kui kasutajalt saadakse teada soovitud väljalennu aeg. Reisiagent suudab riigi mainimisel leida linna sellest riigist ning otsida väljumisi sellesse linna. Suudab kasutajaga ka muud reisijuttu ajada, seda küll piiratud vastustega, ning suunata kasutaja edasi näiteks kas helistama reisibüroosse või minema piletikassasse.

**Näide 14.** *Reisiagendi rakendusnäide koos kommentaaridega.*

Vestlus	Kommentaar
<p>Küsimus: Tere!</p> <p>Vastus: Tere. Kuhu soovite lennata?</p>	<p>Sihtkohta ega kohanime küsimuses e. lauses ei leidunud. Vastuse leidmisel vaadati läbi võtmesõnade andmebaas. Sealt leiti võtmesõna 'Tere', millele võimalike vastuste hulgast valiti juhuslikult 'Tere. Kuhu soovite lennata?'</p>
<p>Küsimus: Tädi juurde Nigeeriasse!</p> <p>Vastus: Nigeeria puudub lennuandmebaasist.</p>	<p>Sihtkohtade andmebaasist vastust ei leitud. Seejärel vaadati läbi kohanimede andmebaas. Leiti, et Nigeeria on kohanimi. Vastati, et selline kohanimi puudub sihtkohtade seast e. lennuandmebaasist.</p>
<p>Küsimus: Tahtsin sealt pannkoogijahu osta.</p> <p>Vastus: Pannkoogijahude kohta kahjuks infot ei anna. Räägime Tallinnast väljuvatest lendudest.</p>	<p>Sihtkohti, kohanimed, võtmesõnu, lemmatiseeritud võtmesõnu ei leidunud. Morfoloogiline analüüs koos ühestamisega eraldas nimisõnad. Morfoloogiline süntees genereeris nimisõnast 'pannkoogijahu' mitmuse osastava. Vastati, et 'Pannkoogijahude kohta info puudub'.</p>
<p>Küsimus: Lendaks siis Helsingisse.</p> <p>Vastus: Helsingisse saab lennata reedel kell 06:20, kell 09:10, kell 11:30, kell 14:50, kell 17:45 ja kell 20:50.</p>	<p>Leiti sihtkoht. Vastuseks väljastati lennuajad sellesse sihtkohta küsimuse esitamise ajale järgneva nädalapäeva kohta.</p>
<p>Küsimus: Kust see lennuk väljub?</p> <p>Vastus: Lennud väljuvad Tallinna lennujaamast.</p>	<p>Sihtkohta, kohanime ega tuttavat võtmesõna ei leidunud. Küsimus lemmatiseeriti morfoloogilise analüsaatori abil. Lemmatiseeritud sõnade hulgas oli 'väljuma' ning selle vormi kohta oli juba võtmesõnade andmebaasis vastus olemas.</p>

**Kui veebiliideses märkida, et soovitakse kõnesünteesi, siis suunatakse kõik vastused kõnesünteesi.**

Reisiagendiga peetavad dialoogid salvestatakse ning on nähtavad Dialoogikorpuse tööpingi vastava menüüvaliku kaudu. Nende dialoogide arvel täieneb eesti dialoogikorpuse. Praegu on need dialoogid väga lihtsa ülesehitusega ega paku dialoogikorpuse uurijatele erilist huvi, kuid programmi edasiarendamise tulemusel saab edaspidi genereerida ka keerulisemaid dialooge.

Reisiagendis on mõnel määral arvestatud ka eestikeelsete suuliste dialoogide uurimise tulemusi, nt tervitused, tänamised, hüvastijätud, vestluse lõpetamised. Näiteks on vaadeldud, milliste sõnadega inimesed enamasti tänu avaldavad, ning need sõnad on lisatud Reisiagendi võtmesõnade andmebaasi.

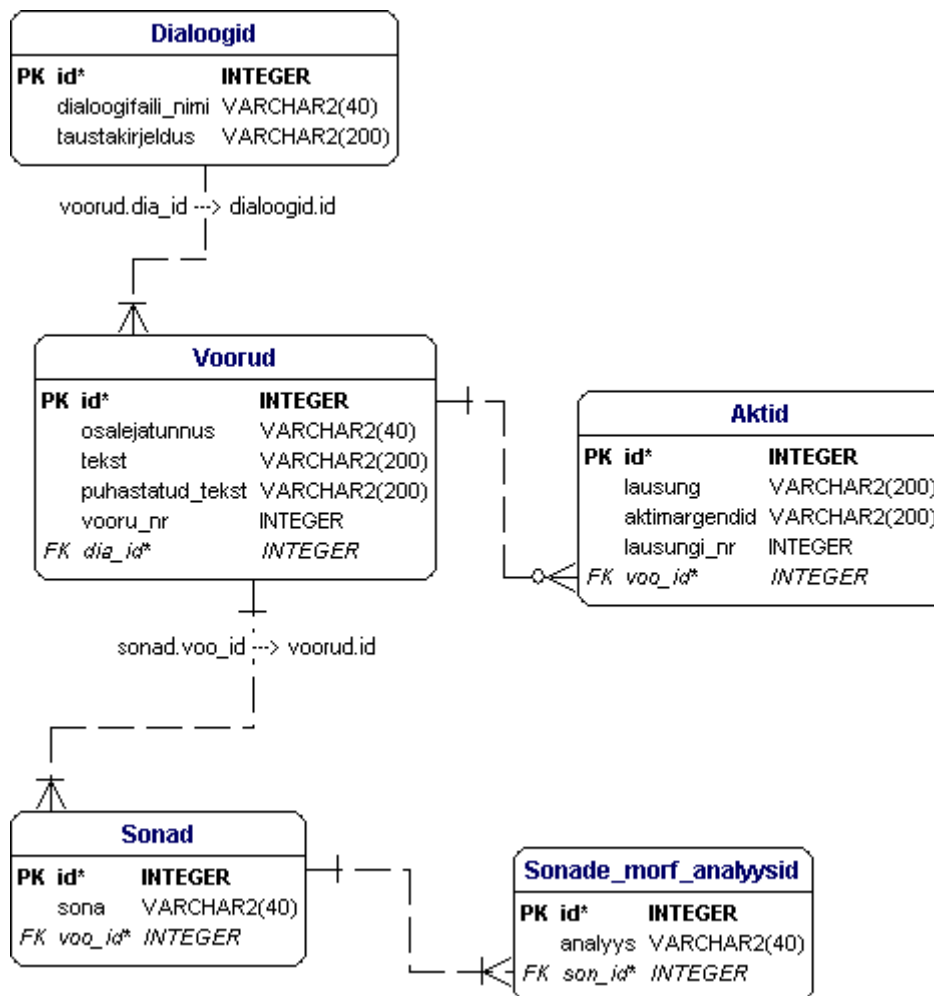
Dialogsüsteemi Reisiagent edasiarendusplaanid on kirjeldatud 6. peatükis "Võimalikud edasiarendused".

## 6. Võimalikud edasiarendused

### 6.1. Dialoogikorpuse tööpingi edasiarendus

Edasiarendusena on kavas realiseerida joonisel (vt. Joonis 18) toodud täielikum relatsiooniline mudel, mis võimaldaks laialdasema otsingu ja aruandluse loomist dialoogikorpusele. Antud andmemudelil on lisandunud tasand morfoloogilise märgenduse tarvis. Selle realiseerimisel on võimalik teostada otsingut ning aruandlust morfoloogiliste tunnuste järgi.

Olemile **dialoogid** on lisandunud ka atribuut **taustakirjeldus**. Hiljem, kui realiseerub taustakirjelduste struktuurne kogumine, tuleb ka taustakirjelduse atribuut viia omaette olemiks. Taustakirjelduste struktuurne kogumine seisneb taustakirjelduses leiduvatele atribuutidele väärtuste omistamises. Enne transkribeerimist peaks transkribeerija täitma vastava taustakirjeldusankeedi, mis katab lisa 2 toodud atribuudid (vt Lisa 2)



**Joonis 18.** Dialoogikorpuse laiendatud andmemudel.

Lisaks on dialoogikorpuse tarkvarale vaja teha järgnevad täiendused:

- dialoogide kustutamine (st. korpusest eemaldamine) ühekaupa,
- dialoogi XML kuju genereerimine mitmele dialoogile korraga,
- dialoogide valimine korpusest ka ajateljele paigutuse, puhastuse ja XML kujule viimise osas
- samanimeliste dialoogide lisamisel ülekirjutamisest hoiatuse andmine kasutajale

Dialoogiaktide järgnevuste sagedusi teades tuleks püüda realiseerida ka aktide automaatne märgendaja, mis põhineks N-gramm mudelil.

Samuti on vaja luua võimalused igale kasutajale oma testkorpuse rajamiseks. Hetkel tuleb seda teha käsitsi, st. luua koopia kogu tarkvarast teisele URLile.

## 6.2. Dialoogsüsteemi Reisiagent edasiarendus

Dialoogsüsteemis kasutatav lendude andmebaas on hetkel staatiline ning sisaldab vaid Tallinnast väljuvaid lende. Eesmärk on siduda Reisiagent reaalse online lennuandmebaasiga (näiteks <http://www.estravel.ee> kaudu). Reisibüroo Estravel on huvitatud veebipõhise dialoogsüsteemi paigutamisest oma veebi ning selle sidumisest reaalse lennuinfo andmebaasiga. Töö ühe edasiarendusena integreeritaksegi Reisiagent lähitulevikus Estraveli lennuandmebaasiga ning paigutatakse Estraveli veebi.

Reisiagendi vastusemallides tuleb teha täiendusi. Näiteks antakse praegusel juhul kasutajale, kes ei täpsustanud oma lennuaega, vastus nõnda, et pakutakse talle esimest ettejuhtuvat päeva. Tegelikult peaks sellisel juhul lugema ette kõik nädalapäevad, millal on võimalik soovitud sihtkohta lennata.

Reisiagent võiks salvestada vestlused nõnda, et oleks näha vestluse kellaeg ning erinevates sessioonides toimunud vestlused hoitaks üksteisest eraldi.

Reisiagendi tööaknas võiks olla lisaks viimasele öeldud lausele näha ka kõik eelnevad laused. Reisiagent võiks suuta väljuda kordustest. Kui kasutaja hakkab ühte ja sama asja kordama, siis peaks Reisiagent näiteks kolmandal korral püüdma teemat vahetada või kasutajale mainima, et toimus kordamine. Kaugemas tulevikus on plaanis lisada Reisiagendile piiratud kõnetuvastuse moodul, sõltuvalt kõnetuvastuse mooduli valmimisest. Kõnetuvastuse arendamisega tegeldakse Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris koostöös Eesti Keele Instituudiga.

## Kokkuvõte

Töö annab ülevaate eesti dialoogikorpusest ja selle tarbeks loodud tarkvarast.

Töö on jätkuks autori bakalaureusetööle, mille käigus loodi dialoogsüsteem Reisiagent. Magistritöö raames realiseeriti see dialoogsüsteem veebipõhisena ning täiendati selle funktsionaalsust morfoloogilise sünteesi ja morfoloogilise ühestamise kasutuselevõtuga.

Dialoogsüsteemi Reisiagent abil kogutakse uutmoodi sisendit dialoogikorpusele.

Loodud tarkvara kasutatakse juba dialoogikorpuse analüüsimisel ning dialoogidele teisenduste rakendamisel. Sellest johtuvalt on paranenud transkriptsiooni kvaliteet ning täienenud dialoogiaktide märgendusjuhend. Antud tööpink võimaldab dialooge/dialoogikorpust analüüsida, aitab leida transkriptsiooni- ja märgendusvigu, sooritada otsingut üle korpuse ning rakendada dialoogidele mõningaid teisendusi nagu paigutus ajateljele, dialoogi puhastamine morfoloogilise analüüsi tarvis ning dialoog viimine XML kujule.

Töö jätkub kavatsuse kohaselt doktoriõppes tarkvara täiendamisega ning dialoogsüsteemi Reisiagent viimisega laiemate kasutajaskondade ette.



## **Abstract**

### **Estonian dialogue corpus and a software tool for the corpus**

Margus Treumuth

Spoken Language Dialogue Systems are becoming more important daily. Dialogues must be gathered in electronic corpuses or archives for language studies and dialogue modeling.

There was a need for software to enable automatic acquisition of knowledge from Estonian dialogue corpora, provide some statistical information for creating statistical models and transfer the corpus to a structured form that is easily adaptable by other systems.

The thesis gives an overview of Estonian dialogue corpus and describes the software that was created for the corpus.

The thesis also completed some plans that were stated in the Bachelor's Paper, which includes building an experimental web-based dialogue system that can be used to collect dialogues. The system incorporates the Estonian morphological analyzer and Estonian speech synthesizer.

The results of the thesis will benefit spoken language dialogue systems and scientists working with Estonian dialogue corpus.

The software for the corpus and the web-based dialogue system are open source database-driven web sites using PHP, JavaScript and MySQL.

## Kasutatud kirjandus

1. Treumuth, Margus 2002. Eestikeelne dialoogsüsteem lennuinfo andmebaasi liidesena. Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut. Tartu
2. Eestikeelse dialoogi modelleerimine arvutil. ETFi projekti nr 4555 (2001-2003) lõpparuanne. Tartu, 2004 (käsikiri).
3. Korpuslingvistika kursus  
[http://www.cl.ut.ee/ee/kursused/www\\_kmuis/korp\\_ling01.html](http://www.cl.ut.ee/ee/kursused/www_kmuis/korp_ling01.html) (kasutatud 19.04.04)
4. SIGdial Resources  
<http://www.sigdial.org/resources.php?tjek=2#Language%20resources>  
(kasutatud 19.05.04)
5. Tiit Hennoste, Liina Lindström, Andriela Rääbis, Piret Toomet, Riina Vellerind. Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse. Arvutuslingvistikalt inimesele. Lk. 245-283. Tartu 2000 Lk. 245-283.
6. Fred Karlsson. Üldkeeleteadus. Eesti Keele Sihtasutus, Tallinn 2002
7. Tartu Ülikooli suulise kõne uurimisrühma kodulehekülg  
<http://sys130.psych.ut.ee/~linds/> (kasutatud 19.04.04)
8. Kullasaar, Maret 2001. Eestikeelse dialoogikorpuse arendamine "võlur Ozi" tehnikaga. Magistritöö. Tartu Ülikool, arvutiteaduse instituut. Tartu
9. MATE Multilevel Annotation, Tools Engineering  
<http://mate.nis.sdu.dk/> (kasutatud 19.04.04)
10. CLAN - Computerized Language Analysis  
<http://childes.psy.cmu.edu/> (kasutatud 19.04.04)
11. PHP: Documentation  
<http://www.php.net/docs.php> (kasutatud 19.04.04)
12. MySQL Documentation  
<http://dev.mysql.com/doc/> (kasutatud 19.04.04)
13. Apache HTTP Server Documentation  
<http://httpd.apache.org/docs-project/> (kasutatud 19.04.04)
14. Cascading Style Sheets Documentation  
<http://www.w3.org/Style/CSS/> (kasutatud 19.04.04)
15. Language and computers, Markov models  
<http://www-rohan.sdsu.edu/~gawron/stat/hmm354.htm> (kasutatud 19.04.04)
16. Jüri Kiho. Tarkvara grammatika-põhine visuaalne modelleerimine ja konverteerimine  
<http://www.cs.ut.ee/~kiho/fRED/Conversion/> (kasutatud 19.04.04)
17. ESTMORF, eesti keele morfoloogiline analüsaator ja süntesaator  
<http://www.eki.ee/keeletehnoloogia/projektid/estmorf/estmorf.html> (kasutatud 19.04.04)
18. Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Doktoritöö. Tartu, TÜ arvutiteaduse instituut, 2000.
19. Müürisep, Kaili. Eesti keele kitsenduste grammatika süntaksianalüsaatori märgendid.  
<http://www.cs.ut.ee/~kaili/syntax.html> (kasutatud 19.04.04)
20. OÜ Filosoft. Morfoloogilise analüsaatori ESTMORF kasutamine.  
[http://www.filosoft.ee/html\\_morf\\_et/morfoutinfo.html](http://www.filosoft.ee/html_morf_et/morfoutinfo.html) (kasutatud 19.04.04)
21. Nurmsalu, Evely 2001. Eestikeelse dialoogikorpuse märgendamistarkvara. Magistritöö. Tartu Ülikool, arvutiteaduse instituut. Tartu

## Lisa 1. Transkriptsioonimärgid

.	langev intonatsioon
?	tõusev intonatsioon
,	poollangev intonatsioon
´sõna, <u>sõna</u>	(ülakoma või allajoonimine) rõhutamine
[ ]	pealerääkimine
(.)	mikropaus (0.2 sek või lühem)
(...)	pikem paus
(0.5)	pausi pikkus sekundites
=	kaks üksust liituvad ilma pausita
> <	kiirendatud lõik
< >	aeglustatud lõik
* *	muust kõnest vaiksem lõik
AHA	(suurtähed) hääle kõvendamine
e::i	(kaksikpunkt) häälikuvenitus
si-	(sidekriips) sõna jääb pooleli
{ või }	ebaselgelt kuulnud lõik
{ - }	ebaselgeks jäänud sõna
{ --- }	pikem ebaselgeks jäänud lõik
(( ))	literaeria kommentaarid ja seletused situatsiooni kohta
hehe	naer
s(h)õna	(sulgudes olev h sõna sees) sõna on lausutud naerdes
\$ \$	naerev häääl, kuid mitte tõeline naer
@ @	hääletooni või hääle kvaliteedi muutumine
.hh	sissehingamine
hh	väljahingamine
.jaa	(punkt sõna ees) sõna on lausutud sisse hingates
mt	matsutamine

## Lisa 2. Taustakirjelduse näide

97 AR7A6

telefonikõne hambakliinikusse (vt näide 1).

NB

- sulgudes pakutud valikutest tuleb valida üks
- kooloni järele tuleb kirjutada ise vajalik materjal kooloni ees toodud nõudeid ja soovitusi arvestades

### 0. Tehniline info

0.2. Lindistaja: **Andriela Rääbis**

Litereerija: **Andriela Rääbis, üle kuulatud suulise kõne tööruhma poolt**

Mitmes litereering (1, 2, 3, lõplik):

Milline osa lindist on litereeritud: **kogu situatsioon**

### 1. Situatsioon ja olukord

1.1. Aeg ja koht

-päev, kuu, aasta: **23. jaan 1998 kell 18.15**

vahetu suhtlus

-koht (linn, maakond, vald, küla, talu):

-kodu (eramaja, korter) / ametiasutus (kontor, kauplus jne):

#### telefonisuhtlus

Helistaja:

-koht (**linn**, maakond, vald, küla, talu): **Tartu**

-**kodu** (eramaja, **korter**) / ametiasutus (kontor, kauplus jne):

Vastaja:

-koht (**linn**, maakond, vald, küla, talu): **Tartu**

-kodu (eramaja, korter) / **ametiasutus** (kontor, kauplus jne): **hambakliinik**

**kohalik** / kaugekõne, **tavaliselt telefonilt** / mobiililt **tavalisele telefonile** / mobiilile

1.4. Osalejate asetus ruumis (istuvad / seisavad; laua, toolide jm esemete kasutamine - kirjeldada):

suhtlusdistant (alla poole meetri, pool kuni poolteist, pikem):

1.6. Situatsiooni kultuuriline määratlus:

-vestluse põhitüüp (**argisituatsioon (helistaja)** / **avalik situatsioon (vastaja)**), eravestlus / **ametialane vestlus**):

Näiteks: ametialane vestlus argisituatsioonis – koduses sünnipäevalauas räägivad õpetajad omavahel koolitööst; eravestlus avalikus situatsioonis – töökaaslased räägivad tööruumis omavahel eraelujuttu.

-nimetus võimalikult täpselt: **helistaja soovib infot hambaravi kohta**

1.11. Situatsioonis suhtlust häirivad või seda positiivselt mõjutavad situatsioonivälised faktorid (telefonikõne, vööra tulek, toidu toomine, kohvi pakkumine, teadmine, et tuleb lindistada jne):

## 2. Suhtlejad, nende omadused ja omavahelised suhted

2.1. Konkreetsete suhtlejad

-nimed: **Andriela Rääbis (helistaja) ja hambakliiniku registraator (vastaja)**

-suhtlejate hulk (**kaks inimest** / väike rühm kuni kümnekond inimest / suur rühm):

-rollid: suhtlejad:

juuresolijad:

-konkreetsed sotsiaalsed rollid: **potentsiaalne patsient ja registraator**

Suhtlejate sotsiaalbioloogilised omadused:

Helistaja

-sugu: **naine**

-vanus või sünniaeg: **31-a**

-haridus (alg / põhi / kesk / **kõrg**):

-rahvus / hõim: **eestlane**

-kodukant / lapseõlvekodu: **sündinud Tartus, keskkooli lõpuni elanud Kohtla-Järvel, edasi Tartus**

-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**

Vastaja: naine

2.5. Suhtlejate selle hetke omadused (väsimus, emotsionaalne seisund, kaine / purjus jne):

2.7. Suhtlejate omavahelised suhted üldse ja konkreetsetes situatsioonides

-**võõras** / tuttav / lähedane (kirjeldada) :

-staatussuhted (võrdne / alluv / kõrgem; - lisada, kumb on kumb):

## 3. Ainestik ja teema

3.1. Konkreetne teema või teemad: **hambaravi hinnad ja soodustused**

3.5. Teemasündmuse seotus suhtlemisolukorraga ja ümbrusega:

-**seos olukorraga puudub** (jutt on varasematest sündmustest / väljamõeldud sündmustest / abstraktne jutt):

-seos olukorraga on: sündmused või asjad samas ruumis või samal ajal: vaatavad koos paberit, telekat jms ja kõnelevad seal nähtavast; sündmused toimuvad reaalselt (nt spordi otsereportaaz, blanketi täitmine):

## 4. Tekst ja suhtlus

4.1. Teksti ja suhtluse liik:

-**dialoog** / monoloog / polüloog

4.2. Tekstiosa / teemaosa retooriline tüüp (jutustus, kirjeldus, loogiline arutus, veenmine, assotsiatiivne vestlus, instruktsioon, referaat, imitatsioon, romaan, lehe uudis muu):

4.3. Teksti planeerituse aste

-**varem** / samal ajal: (**helistaja poolt**)

-planeerimissügavus (**teema** / **eesmärgid** / struktuur / märksõnad / sõnavara / süntaks / intonatsioon):

4.4. Teksti fikseeritus : **ei**

-paberil / peas:

-kogu tekst / osa / mitte midagi:

## 5. Keel ja keelekasutus

5.5. Võti:

-energilisus (käsin / **palun**):

-toon (mäng / **tõsine**, ironiline, vinguv, käskiv, pehme, laiav etc):

## 6. Lisa

Vabas vormis ülevaade tekstis ja situatsioonis silma torganud huvitavate joonte kohta.

Osalejate tähistus tekstis:

H - helistaja

**V – vastaja**

## **Lisa 3. Eesti dialoogikorpuse märgendamisel kasutatavate dialoogiaktide tüpoloogia**

Iga dialoogiakti nimetus (nt RIE: TERVITUS või IL: TÄPSUSTAMINE) koosneb kahest osast:

(1) akronüüm, mis sisaldab kas kaks või kolm tähte, millest esimesed kaks näitavad aktirühma (nt RI = rituaalid, IL = infolisa) ning kolmas, mida kasutatakse ainult naaberpaare moodustavate aktide puhul, näitab, kas on tegu naaberpaari esi- või järelliikmega (vastavalt E või J)

(2) akti nimi (nt TERVITUS, TÄPSUSTAMINE).

### **I. Naaberpaare moodustavad aktid**

#### **1. Rituaalid**

Tervitamine

-esiliige

RIE: TERVITUS

-järelliige

RIJ: VASTUTERVITUS

Hüvastijätmine

-esiliige

RIE: HÜVASTIJÄTT

-järelliige

RIJ: VASTUHÜVASTIJÄTT

Soovimine

-esiliige

RIE: SOOVIMINE

-järelliikmed

RIJ: TÄNAMINE

RIJ: VASTUSOOVIMINE

Viisakusküsimus

-esiliige

RIE: VIISAKUSKÜSIMUS

-järelliige

RIJ: VIISAKUSVASTUS

Tänamine

-esiliige

RIE: TÄNAN

-järelliige

RIJ: PALUN

Palumine

-esiliige

RIE: PALUN

-järelliige

RIJ: TÄNAN

Vabandamine

-esiliige

RIE: VABANDUS

-järelliige  
RIJ: VABANDUSE VASTUVÕTMINE

Esitlemine  
-esiliige  
RIE: ESITLUS  
-järelliikmed  
RIJ: VASTUESITLUS  
RIJ: HINNANG

Kutsesignaal  
-esiliige  
RIE: KUTSUNG  
-järelliige  
RIJ: KUTSUNGI VASTUVÕTMINE

Lõpetamine  
-esiliige  
RIE: LÕPUSIGNAAL  
-järelliikmed  
RIJ: LÕPETAMISE VASTUVÕTMINE  
RIJ: LÕPETAMISE TAGASILÜKKAMINE

Rituaalid: muu  
RIE: MUU  
RIJ: MUU

## 2. Teemavahetus

-esiliikmed  
TVE: PAKKUMINE  
TVE: MUU  
-järelliikmed  
TVJ: VASTUVÕTMINE  
TVJ: TAGASILÜKKAMINE  
TVJ: MUU

## 3. Partneri algatatud parandused

-esiliikmed  
PPE: ÜMBERSÕNASTAMINE  
PPE: ÜLEKÜSIMINE  
PPE: MITTEMÕISTMINE  
PPE: MUU

-järelliige  
PPJ: LÄBIVIIMINE  
PPJ: MUU

## 4. Kontakti kontroll

-esiliikmed  
KKE: ALGATUS  
KKE: MUU



-järelliikmed  
KKJ: KINNITAMINE  
KKJ: MUU

## 5. Direktiivid

-esiliikmed  
DIE: SOOV  
DIE: ETTEPANEK  
DIE: PAKKUMINE  
DIE: PALVE OODATA  
DIE: MUU

-järelliikmed  
DIJ: INFO ANDMINE  
DIJ: INFO PUUDUMINE  
DIJ: KEELDUMINE  
DIJ: NÕUSTUMINE  
DIJ: MITTENÕUSTUMINE  
DIJ: PIIRATUD NÕUSTUMINE  
DIJ: TEGEVUS  
DIJ: EDASILÜKKAMINE  
DIJ: MUU

## 6. Küsimused

-esiliikmed  
KYE: SULETUD KAS  
KYE: JUTUSTAV KAS  
KYE: ALTERNATIIV  
KYE: AVATUD  
KYE: VASTUST PAKKUV  
KYE: VASTUSE TINGIMUSTE TÄPSUSTAMINE  
KYE: MUU

-järelliikmed  
KYJ: JAH  
KYJ: EI  
KYJ: MITTENÕUSTUV JAH  
KYJ: NÕUSTUV EI  
KYJ: ALTERNATIIV: ÜKS  
KYJ: ALTERNATIIV: MÕLEMAD  
KYJ: ALTERNATIIV: KOLMAS VALIK  
KYJ: ALTERNATIIV: EITAV  
KYJ: TEGEVUS  
KYJ: INFO ANDMINE  
KYJ: INFO PUUDUMINE  
KYJ: KEELDUMINE  
KYJ: EDASILÜKKAMINE  
KYJ: VASTUS ALTERNATIIVINA  
KYJ: MUU

## 7. Seisukohavõtud

-esiliikmed  
SEE: VÄIDE  
SEE: ARVAMUS  
SEE: MUU

-järelliikmed  
SEJ: NÕUSTUMINE  
SEJ: MITTENÕUSTUMINE  
SEJ: PIIRATUD NÕUSTUMINE  
SEJ: KEELDUMINE  
SEJ: MUU

## II. Üksikaktid

### 1. Infolisad

IL: TÄPSUSTAMINE  
IL: SELETAMINE  
IL: PÕHJENDAMINE  
IL: JÄRELDAMINE  
IL: KOKKUVÕTMINE  
IL: ÜLERÕHUTAMINE  
IL: PEHMENDAMINE  
IL: HINNANG  
IL: MUU

### 2. Vabatahtlikud reaktsioonid

VR: HINNANGULINE JÄTKAJA  
VR: NEUTRAALNE JÄTKAJA  
VR: HINNANGULINE VASTUVÕTUTEADE  
VR: NEUTRAALNE VASTUVÕTUTEADE  
VR: HINNANGULINE INFO OSUTAMINE UUEKS  
VR: NEUTRAALNE INFO OSUTAMINE UUEKS  
VR: HINNANGULINE PIIRITLEJA  
VR: NEUTRAALNE PIIRITLEJA  
VR: PARANDUSE HINDAMINE  
VR: MUU

### 3. Parandused

PA: ENESEPARANDUS  
PA: PARTNERI PARANDUS  
PA: MUU

### 4. Rituaalsed üksikaktid

RY: ÜLEANDMINE  
RY: TUTVUSTUS  
RY: ÄRATUNDMINE  
RY: KONTAKTEERUMINE

RY: KUTSUMINE  
RY: MUU

## 5. Üksikaktid

YA: EELTEADE  
YA: JUTUSTAMINE  
YA: LUBADUS  
YA: INFO ANDMINE  
YA: JUTU PIIRIDE OSUTAMINE  
YA: RETOORILINE KÜSIMUS  
YA: RETOORILINE VASTUS  
YA: REFERAAT  
YA: MUU  
YA: MITTEVERBAALNE AKT  
YA: PRAAK