TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika eriala

Sven Aller

Dialoogiaktide märgendamine Eesti dialoogikorpuses: ülevaade ressurssidest ja tarkvaraarendus

Magistritöö (30 EAP)

Juhendaja: prof. Mare Koit

Autor: ..... "....." mai 2012
Juhendaja: ..... "....." mai 2012
Lubada kaitsmisele
Professor Mare Koit ..... "....." mai 2012

TARTU 2012

Sisukord

Sissejuhatus ... **3**
1. Dialoogid ... **5**
1.1. Dialoogiaktide piirid ... 6
1.2. Dialoogiaktide klassifitseerimine ... 7
2. Dialoogiaktide märgendamistarkvara ... **11**
2.1. Manuaalne märgendamine ... 11
2.2. Automaatne märgendamine ... 12
3. Poolautomaatne märgendaja DAREC ... **15**
3.1. Algoritm ... 15
3.2. Ülesehitus ... 16
3.3. Testimistulemused ... 17
3.4. Testijate hinnangud ... 19
4. Kasutajaliidese uuendamine ... **20**
4.1. Ühtsus ja standardite järgimine ... 23
4.2. Süsteemi ja reaalse maailma kooskõla ... 24
4.3. Kasutajapoolne kontroll süsteemi üle ja vabadus ... 24
4.4. Ülevaade süsteemi olekust ja vähene koormus kasutaja mälule ... 25
4.5. Paindlikkus ja efektiivsus ... 25
4.6. Veahaldus ... 26
4.7. Abiinfo ja dokumentatsioon ... 27
4.8. Esteetiline ja minimalistlik disain ... 27
5. Tulemused ... **29**
5.1. Kasutajaliidese probleemide lahendamine ... 29
5.2. Testijate hinnangud ... 29
6. Edasiarendusvõimalused ... **31**
Kokkuvõte ... **32**
Summary ... **33**
Kasutatud kirjandus ... **34**
Lisad ... **37**
Lisa 1. Dialoogiaktide tüpoloogia EdiT ... 37
Lisa 2. DARECi testimistulemused ... 41
Lisa 3. DARECi kasutajaliidese hindamise küsimustik ... 44
Lisa 4. DARECi uuendatud kasutajaliidese abiinfo ... 46

Sissejuhatus

Keeletehnoloogias luuakse mitmesugust loomulike keelte automaattöötluse tarkvara, mille kasutajad võivad teatud juhtudel olla küll suurepäraste keelealaste teadmistega, kuid suhteliselt vähese arvutikogemusega. Sellise tarkvara hulka kuuluvad näiteks mitmesugused analüüsiprogrammid (morfoloogiline, süntaktiline, semantiline analüsaator), tekstikorpuste erinevatel keeletöötluse tasemetel märgendamise programmid, märgendatud korpustest info otsimise programmid jne. Tarkvara kavandamisel on seetõttu oluline omistada piisavat tähelepanu ühest küljest selle tõrkekindlusele, teisest küljest aga kindlasti ka kasutus-mugavusele tavatarbija seisukohalt.

Magistritöö käsitleb teatavat liiki keelekorpuse, nimelt dialoogikorpuse märgendamist. Dialoogikorpustesse kogutakse kahe või enama osaleja vahelist suhtlust, kusjuures osalejatest üks on inimene, teine kas inimene või arvuti. Dialoogikorpuste kogumise ja märgendamise üks eesmärk on arvutiprogrammide (näiteks dialoogsüsteemide) loomine, mis suudaksid suhelda kasutajatega loomulikus kirjalikus või suulises keeles, järgides inimestevahelise suhtlemise tavasid ja reegleid. Teiseks eesmärgiks on kommunikatsiooni uurimine: kuidas suhtlevad inimesed omavahel ja kuidas arvutiga, missugused on suhtlejate eesmärgid ja kuidas nad neid saavutavad, kuivõrd sõltub suhtlejate keelekasutus nende omavahelistest rollidest ja suhtluspartnerist jne. Tänu kogutule on võimalik analüüsida dialoogide ülesehitust, tavapäraseid lausungite järgnevusi, parandusmehhanisme vms.

Maailmas on koostatud palju erinevaid dialoogikorpusi, kuhu on kogutud nii inimestevahelisi kui ka inimese ja arvuti vahelisi dialooge erinevates keeltes. Headeks näideteks on *DARPA Communicator*, *HCRC Map Task*, *TRAINS*, *VERBMOBIL* jpt (McTear, 2004). Korpusi on märgendatud sõltuvalt nende kasutamise eesmärgist erinevatel keeletasemetel (süntaktiline, semantiline jne). Enamasti märgendatakse dialoogikorpustes ka dialoogiaktid, selleks on välja töötatud mitmeid erinevaid dialoogiaktide tüpoloogiaid (Koit, 2003).

Magistritöös on vaatluse all Eesti dialoogikorpus ja selles dialoogiaktide märgendamine. Märgendamiseks kasutatakse Tartu Ülikoolis väljatöötatud dialoogiaktide tüpoloogiat, mille teoreetiliseks aluseks on vestlusanalüüs (vt. Hennoste ja Rääbis, 2004). Kuna tüpoloogia on keerukas (sisaldab kokku 127 erinevat akti) ja dialoogiaktid ei ole kõikidel juhtudel täpselt defineeritud, siis on nende määramine dialoogides küllalt raske ülesanne. Dialoogide käsitsi märgendamine on töömahukas, samas on nende struktuur sageli sarnane. Seetõttu on

katsetatud erinevaid automaatseid märgendamismeetodeid, nende hulgas nii reeglitel põhinevaid kui ka statistilisi.

Magistritöö eesmärgid on:

- kirjeldada Eesti dialoogikorpuse ressursside hetkeolukorda;
- anda ülevaade dialoogide märgendamiseks mõeldud vahenditest;
- arendada edasi poolautomaatset dialoogide märgendajat DAREC, esmalt analüüsides ja seejärel kõrvaldades põhjusi, mis on seni takistanud selle praktilist kasutuselevõttu.

Magistritöö koosneb sissejuhatusest, kuuest peatükist, kokkuvõttest eesti ja inglise keeles ning neljast lisast. Esimeses peatükis antakse ülevaade dialoogidest, nende ülesehitusest ja dialoogiaktide klassifitseerimisest. Teine peatükk käsitleb Eesti dialoogikorpuses dialoogiaktide märgendamistarkvara ja kolmas poolautomaatset märgendajat DAREC. Neljandas peatükis kirjeldatakse hea kasutajaliidese omadusi ja neist lähtudes tehtud uuendusi DARECis. Viies peatükk analüüsib tehtud töö tulemusi ja kuues kirjeldab edasiarendusvõimalusi. Lisadena on toodud Eesti dialoogikorpuses kasutatav dialoogiaktide tüpoloogia, DARECi testimisel saadud tulemused, kasutusmugavuse analüüsis kasutatud küsimustik ja DARECi uue kasutajaliidese abiinfo.

Magistritöö raames uuendatud kasutajaliidesega märgendustarkvara DAREC on kättesaadav veebis aadressil http://ats.cs.ut.ee/darec/www1/.

1. Dialoogid

Inimeste suhtlemine toimub harilikult dialoogi vormis, see on suuline või kirjalik loomulikus keeles suhtlus tavaliselt kahe suhtleja vahel. Ka monoloogi (räägib üks inimene) ja polüloogi (võtab osa rohkem kui kaks inimest) võime vaadelda kui vastavalt kärbitud või laiendatud dialoogi.

Dialoogide uurimine võimaldab paremini aru saada dialoogi ülesehitusest üldiselt. See eeldab süstematiseeritud dialoogide kogu e. dialoogikorpust. Dialoogikorpuste moodustusviise on erinevaid: suuliste vestluste salvestamine ja salvestuste häälduse järgi üleskirjutamine e. transkribeerimine (Transkriptsioonimärgid, 2012), nn. Võlur Ozi meetod (salvestatud dialoog kahe inimese vahel arvuti teel, kus üks inimene arvab, et suhtleb dialoogsüsteemiga), inimese ja arvuti vahelise dialoogi salvestamine. Saadud dialoogid erinevad üksteisest oluliselt, näiteks suhtleb inimene inimesega vabamalt, arvuti puhul aga arvestab tehispartneri piiratusega. Samuti erinevad dialoogid ka osalejate motivatsiooni poolest: värvatud kasutajaid iseloomustab võrreldes reaalsete kasutajatega suurem suhtlemisaktiivsus ja -kiirus, sujuvus ning väiksem vajadus abi järele (Ai, 2007). Korpuste uurimisel statistiliste meetoditega on jõutud ka selliste dialoogide genereerimiseni, kus mõlemad osalejad on tehisintellektid ning mis sarnanevad tavalisele inimese ja dialoogsüsteemi omavahelisele suhtlusele (Griol jt, 2009).

Dialoogid võivad korpuses olla nii märgendamata kui märgendatud, viimasel juhul on neile lisatud lingvistilisi andmeid. Märgendid võivad olla lisatud nii sõnadele (näiteks sõnaliik vms) kui lausungitele (näiteks lausungi tüüp). Korpusi võib märgendada erinevatel tasemetel, näiteks prosoodia, morfoloogia, süntaksi, semantika ja pragmaatika tasemel (Treumuth, 2004). Märgendamisviis sõltub eesmärgist, milleks soovitakse dialoogikorpust kasutada.

Eesti dialoogikorpuse (EDiK) loomist alustati 2001. aastal (EDiK, 2012). 2012. aasta märtsi seisuga sisaldas EDiK:

- 1182 inimestevahelist suulist dialoogi, mis on salvestatud ja transkribeeritud, neist 1137 infodialoogi (1026 telefoni teel peetud ja 111 silmast-silma vestlust) ning 45 väitlust, kokku 246 000 sõna;
- 117 kirjalikku arvutisimulatsiooni abil Võlur Ozi meetodil kogutud dialoogi, kokku 14 500 sõna;

- kasutajaga loomulikus eesti keeles suhtlevate dialoogsüsteemide poolt salvestatud arvuti ja inimese vahelist suhtlust.

1.1. Dialoogiaktide piirid

Dialoog koosneb osadest e. nn. dialoogiaktidest, mis väljendavad mingeid kõne abil tehtavaid tegevusi (tervitamine, küsimine jne) (Hennoste jt, 2002). Dialoogiaktide sisu ja nende piiride määratlemine pole enamasti triviaalne ülesanne. Dialoogi puhul saame kandvate üksustena rääkida voorudest (*turn*) ja lausungitest (*utterance*, *utterance unit*). Vooruks nimetatakse ühe kõneleja pidevat häälesolekut (Hennoste jt, 2002). Iga voor võib jaguneda mitmeks lausungiks, mis on kirjakeeles kasutatava mõiste "lause" vaste suulises keeles. Iga lausungiga teeb osaleja teatava keelelise tegevuse: tervitab, tänab, esitab küsimuse, annab infot jne. Dialoogiakti piir ei pruugi ühtida lausungi- või voorupiiriga, sest (keele abil tehtav) tegevus võib hõlmata korraga mitut lausungit, samuti võib ühes lausungis sisalduda mitu tegevust (vt. näide 1) [1].

```
V: aga `kuhu te `sõidate mis=`linn | KYE: AVATUD | | VTE: VASTUSE
  TINGIMUSTE TÄPSUSTAMINE |
```

Näide 1. *Lausung, mis sisaldab mitut tegevust.*

Sageli eristuvad dialoogis selgelt nn. voorupaarid e. naabruspaarid – voorud, mis on sisuliselt üksteisega seotud (näiteks küsimus ja vastus, soov ja info andmine vms). Nende esimest, voorupaari esilekutsuvat osa nimetatakse naabruspaari esiliikmeks, teist, sellele reageerivat osa aga naabruspaari järelliikmeks (näiteks paarid KYE: SULETUD KAS ja KYJ: JAH või KYE: AVATUD ja KYJ: INFO ANDMINE) (vt. näide 2).

```
H: aga `kuidas see algusega `Riiast oleks mis `kella-st `kellaajal ta
  `väljub sis [Riiast.] | KYE: AVATUD |
V: [`Mosk]vasse läheb kolm korda `päevas `buss. | KYJ: INFO ANDMINE |
```

Näide 2. *Naabruspaarid: esiliige ja järelliige.*

Voorupaarid ja keelelised tegevused võivad ka seguneda: ühe partneri poolt algatatud akt võib vastuse saada hoopis hiljem. See teeb dialoogi struktuuri tuvastamise veel keerulisemaks.

[1] Näited on võetud Eesti dialoogikorpusest, suulise keele üleskirjutamisel on kasutatud vestlusanalüüsi transkriptsiooni (Transkriptsioonimärgid, 2012) ja dialoogiaktid on märgendatud Eesti dialoogiaktide tüpoloogia kohaselt (vt lisa 1).

1.2. Dialoogiaktide klassifitseerimine

Dialoogiaktide märgendamisel kasutatakse erinevaid klassifikatsioone e. tüpoloogiaid (Dybkjær ja Bernsen, 2000). Tüpoloogia valik sõltub märgendamise eesmärgist. Arvesse tuleb võtta, kui suur peaks olema märgendite granuleeritus ning kuivõrd peavad tüübid olema selgelt eristuvad ja ammendavad. Aastatel 1998-2000 läbiviidud projekt MATE (*Multilevel Annotation, Tools Engineering*) keskendus dialoogide märgenduse standardi väljatöötamisele erinevatel tasemetel (prosoodia, morfosüntaks, kaasviitamine, dialoogiaktid, suhtlus-probleemid, tasemetevahelised probleemid) (Dybkjær ja Bernsen, 2000). Projektis tõsteti esile märgendusskeemi DAMSL (*Dialog Act Markup in Several Layers*). DAMSLi puhul on kasutusel neli erinevaid vaatenurki peegeldavat tasandit (Eskor, 2004):

- suhtluslik staatus (*Communicative Status*) (märgendid selle kohta, kas lause on lõpe-tatud, mõistetav vms);
- infotasand (*Information Level*) (lause semantiline tähendus);
- ettepoole vaatav funktsioon (*Forward Looking Function*) (lausungi mõju dialoogi tulevikule, näiteks infopäring, pakkumine jne);
- tahapoole vaatav funktsioon (*Backward Looking Function*) (käesoleva lausungi seos eelnevaga, näiteks vastus varasemale lausungile, nõustumine vms).

Erinevate tasandite märgendid annavad kokkuvõttes infot kõneleja kavatsuste, lausungi sisu, eesmärgi ja rolli kohta. Uuringud sobiva standardi väljatöötamisel jätkuvad, dialoogiaktide märgendamisel kasutatakse ka DiAML (*Dialogue Act Markup Language*) mudelit (Bunt jt, 2010).

Eesti dialoogikorpuse loomise ja analüüsi käigus on välja töötatud eestikeelsetes dialoogides kasutatavate dialoogiaktide tüpoloogia EDiT (Eesti Dialoogiaktide Tüpoloogia), mis lähtub vestlusanalüüsi põhimõtetest (vestlus kui ühistegevus, näiteks naabruspaari esiliikme ja järelliikme seos) (Gerassimenko jt, 2010), kuid samas grupeerib aktid loogilistel alustel. EDiTi loomisel on eeskuju võetud DAMSList, kuid erinevad tasandid puuduvad ilmutatud kujul, tähenduse täpsustamine on läbi viidud märgendite hulga suurendamisega. EDiTi puhul on märgendite loetelus 127 tüüpi (vt. lisa 1), aja jooksul on kasutatavat tüpoloogiat vastavalt vajadustele ka muudetud. Tüüpide eristamine toimub lähtuvalt dialoogiakti funktsioonist, mitte keelelisest vormist. Tüpoloogia väljatöötamisel on silmas peetud eestikeelse suulise

suhtluse uurimise pikaajalisi eesmärke ja samuti loomuliku infodialoogi modelleerimist arvutil (Koit, 2003). Iga märgend koosneb kahest osast, millest esimene näitab akti üldist rühma, teine aga tema täpset funktsiooni (Fišel, 2006). Üldist rühma märkiv osa märgendist koosneb kahest või naabruspaaride puhul kolmest tähest, viimasel juhul näitab kolmas sümbol akti kuulumist kas esi- või järelliikmete hulka (vastavalt E või J). Põhirühmade nimekiri on järgmine (Hennoste ja Rääbis, 2004):

- Naabruspaare moodustavad aktid:
    - Rituaalid (RIE, RIJ) – kutsung, tervitamine, tänamine ja palumine, tutvustamine, vestluse lõpetamine jne;
    - Teemavahetus (TVE, TVJ) – pakkumine, vastuvõtmine, tagasilükkamine;
    - Partneri algatatud parandused (PPE, PPJ) – ümbersõnastamine, üleküsimine, selgitamine, kus üks osaleja algatab nn. paranduse ja partner viib selle läbi jne;
    - Vastuse tingimuste täpsustamine (VTE, VTJ) – vastamiseks vajaliku lisainfo hankimine;
    - Kontakti kontroll (KKE, KKJ) – kontrolli algatus ja selle kinnitamine;
    - Direktiivid (DIE, DIJ) – soov, ettepanek, pakkumine, info andmine, nõustumine, keeldumine jne;
    - Küsimused (KYE, KYJ) – avatud, jutustav “kas”, suletud “kas”, vastust pakkuv, alternatiiv jne;
    - Seisukohavõtud (SEE, SEJ) – väide, arvamus, nõustumine, keeldumine jne.
- Üksikaktid:
    - Infolisad (IL) – täpsustamine, seletamine, põhjendamine, järeldamine, hinnang jne;
    - Vabatahtlikud reaktsioonid (VR) – jätkaja, info vastuvõtuteade jne;
    - Parandused (PA) – eneseparandus;

- Rituaalsed üksikaktid (RY) – tutvustus, äratundmine, kontakteerumine, üleandmine;

- Üksikaktid (YA) – jutustamine, lubadus, retooriline küsimus jne.

Dialoogiaktide tüpoloogia EDiT väljatöötamise esmane eesmärk on olnud inimestevahelise suulise suhtluse uurimine. Seetõttu kuulub sellesse palju erinevaid dialoogiakte, millest igaüks on asjakohane ühe või teise suhtlusnüansi kirjeldamisel, mis võivad aset leida inimestevahelises suhtluses. Samas on paljud neist äärmiselt haruldased inimese ja arvuti vahelises suhtluses (näit. RIE: SOOV RÄÄKIDA, RIJ: VASTUVÕTMINE jne), paljusid neist leidub harva ka Eesti dialoogikorpusesse kogutud inimestevahelistes dialoogides. Mõnede dialoogiaktide määramine on isegi kogemustega lingvistidele keeruline, näiteks suletud ja jutustava kas-küsimuse eristamine. Märgendajate üksmeelt mõõtva kapa-koefitsiendi[2] (Carletta, 1996) väärtus on Eesti dialoogikorpuse märgendamisel olnud parimal juhul vaid 0,8 (Fishel, 2007c).

Oma keerukuse tõttu on dialoogiaktide automaatne tuvastamine vastavalt tüpoloogiale EDiT tõsine väljakutse. Ülesande teeb veel raskemaks asjaolu, et Eesti dialoogikorpuse arendamisel on esikohale seatud inimestevahelise suhtluse uurimine, mistõttu on korpusesse kogutud võimalikult eriliigilist suhtlust: erinevaid telefonikõnesid (infotelefon, reisibüroo, bussijaam, polikliiniku registratuur, kauplused, taksodispetšer jm) ja silmast-silma vestlusi (kaubandus, teenindus, reisibüroo, teejuhatamine jm). Muidugi peab arvesse võtma ka asjaolu, et inimestevahelise suhtluse salvestamisel tuleb kinni pidada mitmesugustest andmekaitselistest piirangutest, mis teeb materjali kogumise keerulisemaks.

Dialoogiaktide märgendamist selgitab Eesti dialoogikorpusest valitud näide 3. Selles toodud dialoogi osalejateks on infotelefoni operaator (V) ja infoliinile helistaja (H), dialoogiaktide märgendid asuvad pärast lausungit püstkriipsude vahel.

---

[2] Kapa-koefitsient arvutatakse valemist $\kappa = \frac{P(A) - P(E)}{1 - P(E)}$, kus *P(E)* on võimalike juhuslike nõustumiste osakaal ja *P(A)* tegelike nõustumiste osakaal.

```
((kutsung)) | RIE: KUTSUNG |
V: infotelefon | RIJ: KUTSUNGI VASTUVÕTMINE |
= > Kersti < | RY: TUTVUSTUS |
=tere | RIE: TERVITUS |
H: tere. | RIJ: VASTUTERVITUS |
sooviksin ´Maarjamõisa polikliinikus stomato´loogia mm mis on seal
  info[numbrit] | DIE: SOOV |
V: [kas] ´hambaproteesi või: | KYE: VASTUST PAKKUV | | KYE: VASTUSE
  TINGIMUSTE TÄPSUSTAMINE |
H. jah. jah. | KYJ: JAH |
V: jah. | VR: NEUTRAALNE VASTUVÕTUTEADE |
(7.0)
V: registra´tuur [seitse] kolm üks? | DIJ: INFO ANDMINE |
H: [jah] | VR: NEUTRAALNE JÄTKAJA |
H: jah | VR: NEUTRAALNE JÄTKAJA |
V: üheksa kaks? (0.5) kaheksa kaks. | DIJ: INFO ANDMINE |
H: kaheksa kaks. | VR: NEUTRAALNE VASTUVÕTUTEADE |
{---} ee seitse? | KYE: AVATUD | | PPE: MITTEMÕISTMINE |
V: kolm üks? | KYJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE |
H: jah? | VR: NEUTRAALNE JÄTKAJA |
V: üheksa kaks? | DIJ: INFO ANDMINE |
H: jah, | VR: NEUTRAALNE JÄTKAJA |
V: kaheksa kaks. | DIJ: INFO ANDMINE |
H: aitäh teile? | RIE: TÄNAN |
V: palun. | RIJ: PALUN |
```

Näide 3. *Märgendatud dialoog Eesti dialoogikorpusest.*

2. Dialoogiaktide märgendamistarkvara

Dialoogikorpuse hilisemaks uurimiseks ja kasutamiseks on vajalik dialoogiaktide märgendamine. Seda saab teha küll käsitsi, kuid ühest küljest on manuaalne märgendamine väga aeganõudev, teisest küljest sageli rutiinne: dialoogide ülesehitus on tihti sarnane, seda eriti konkreetset tüüpi dialoogide puhul. Kõnet infotelefonile alustatakse enamasti tervitusest, seejärel esitatakse küsimusi, vastatakse, täpsustatakse, jäetakse hüvasti. Samuti pole manuaalne märgendamine reaalseid vajadusi silmas pidades alati sobiv lahendus: näiteks dialoogsüsteemiga reaalajas toimuva dialoogi puhul peab analüüs olema kohene, et süsteem oskaks kasutaja lausungile adekvaatselt reageerida.

2.1. Manuaalne märgendamine

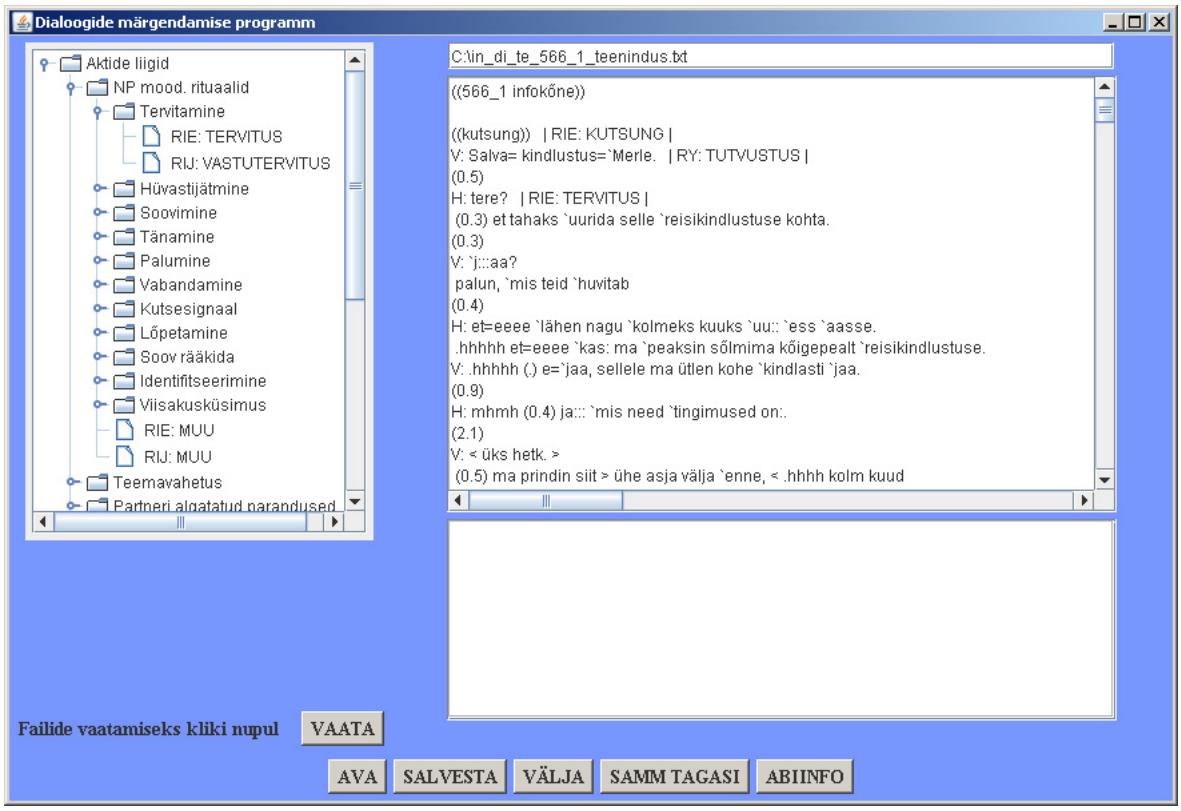
Manuaalse märgendamise korral määrab lingvist dialoogiakti märgendi käsitsi. Arvestades analüüsi puhul tekkida võivaid erinevusi, teeb seda tööd üksteisest sõltumatult mitu inimest, saadud tulemused ühtlustatakse. Eesti dialoogikorpuse märgendamisel tegi tööd paralleelselt kaks lingvisti, kelle märgendeid ühtlustas kolmas. Igas dialoogivoorus eraldati lausungid üksteisest ja igaüks neist pidi saama vähemalt ühe märgendi.

Töö kergendamiseks kasutati tavalise tekstiredaktori asemel Evely Nurmsalu-Vuti poolt 2001. aastal loodud ning 2003. ja 2005. aastal täiendatud dialoogiaktide märgendusprogrammi (Nurmsalu, 2001). Vahend on kirjutatud programmeerimiskeeles *Java* ning võimaldab dialoogifaile (txt-laiendiga tekstifaile) avada, voore lausungiteks jagada, valida dialoogiaktide struktureeritud nimekirjast sobivaid märgendeid ja tulemust salvestada (vt. joonis 1). Programmi kasutajaliides on lihtne, intuitiivne ja kiiresti omandatav, kasutusmugavusele ja töökiirusele pole aga pandud erilist rõhku. Näiteks esinevad järgmised puudujäägid:

- märgendite nimekiri nõuab alajaotustes valikute tegemiseks ülearuseid operatsioone (näiteks märgendi RIE: TERVITUS valimisel tuleb avada menüü “NP mood. rituaalid”, sealt “Tervitamine” ning alles siis jõutakse vajaliku märgendini);

- märgendite nimekirja lahter on väike, nõudes isegi ainult veidi avatud kujul vertikaalset kerimist;

- puuduvad klaviatuurikombinatsioonidega antavad kiirkäsud, mis tõstaksid ekspertkasutajate puhul oluliselt töö kiirust.



Joonis 1. *Dialoogiaktide manuaalne märgendaja.*

2.2. Automaatne märgendamine

Dialoogiaktide automaatseks märgendamiseks on erinevates dialoogikorpustes kasutatud nii reeglipõhiseid kui ka statistilisi meetodeid (Klüwer jt, 2010; Stolcke jt, 2000). Reeglipõhiste meetodite puhul antakse süsteemile ette eeskirjad, mille järgi otsuseid teha, programm valib neile tuginedes lahenduse. Probleemiks on siinkohal asjaolu, et inimesel on raske reeglite koostamisel ning tunnuste valimisel ette näha kõiki võimalusi ja nendest tulenevaid mõjusid. Keerulise seostevõrgu puhul on korrektsete, konflikte mittetekitavate muudatuste tegemine programmis seotud suure riskiga. Eestikeelsetes dialoogides dialoogiaktide automaatseks tuvastamiseks reeglipõhiseid meetodeid seni veel katsetatud ei ole.

Statistilisi meetodeid hakati dialoogitöötluses kasutama palju hiljem kui näiteks masintõlkes, nimelt 1990. aastate lõpul, siis, kui olid juba olemas piisavalt mahukad dialoogikorpused. Need on masinõppe tehnikad, mis vajavad tunnuste leidmiseks varem töödeldud (praegusel juhul dialoogiaktidega märgendatud) andmeid e. nn. treeningandmeid ning nõuavad rohkem riistvaralisi ressursse. Statistiliste meetodite hulgast on dialoogiaktide märgendamiseks kasutatud otsustuspuid, tehisnärvivõrke, Markovi peitmudelit, Bayesi klassifitseerijat jne (Jurafsky ja Martin, 2009).

Statistiliste meetodite mõõtmiseks kasutatakse mõisteid täpsus (*precision*) ja saagis (*recall*) (Witten, 2011). Nende edaspidiseks selgitamiseks tähistame mingisse klassi õigesti määratud elementide arvu *TP* (*true positives*) ja valesti määratute arvu *FP* (*false positives*), klassi mittemääratud, kuid sinna kuuluvate elementide arvu *FN* (*false negatives*), sinna mitte-määratud ja mittekuuluvate elementide arvu *TN* (*true negatives*) (vt. tabel 1).

Tabel 1. *Täpsuse ja saagise arvutamisel kasutatud tähised.*

	Klassi kuuluvad	**Klassi mittekuuluvad**
Klassi määratud	TP	FP
Klassi mittemääratud	FN	TN

Täpsus on korrektsete ja kõikide sellesse klasside määratud märgendite arvude omavaheline suhe e. korrektselt klassifitseeritud dialoogiaktide arvu suhe kõigisse samasse klassi määratute arvu:

$$täpsus = \frac{TP}{TP + FP}.$$

Saagis on korrektselt määratud ja kõikide sellesse klassi kuuluvate märgendite arvude suhe e. korrektselt klassifitseeritud dialoogiaktide arvu suhe kõigisse samasse klassi kuuluvate arvu:

$$saagis = \frac{TP}{TP + FN}.$$

Eestis on dialoogiaktide automaatseks märgendamiseks tüpoloogia EDiT kohaselt katsetatud erinevaid statistilisi meetodeid: tehisnärvivõrke, otsustuspuid, tõenäosuslikke sufiksipuid. Tehisnärvivõrkude (*artificial neural networks*) meetod on nn. musta kasti meetod: treenimisel töödeldakse sisendit kaalutud seostega ühendatud nn. tehisneuronite võrgu abil ja valitakse seostele sobivad kaalud, et tulemus vastaks näidiskorpusele. Hiljem tarvitatakse sel viisil saadud võrku uute sisendite analüüsil. Dialoogiaktide puhul kasutati ühe ja kahe peitkihiga tajureid ning rekurrentseid võrke. Tehisnärvivõrkudes oli dialoogiaktide keskmine tuvastamistäpsus alla 20%. Dialoogiaktide klasse tuvastati küll mõnevõrra paremini, näiteks edukaimaks osutunud laiendatud rekurrentsed võrgud tuvastasid direktiivide esiliikmeid (DIE) täpsusega 27,4% ja direktiivide järelliikmeid (DIJ) täpsusega 48,1% (Fišel ja Kikas, 2006).

Otsustuspuude (*decision tree*) puhul on tegemist hierarhiliste otsuste struktuuriga, kus puu servad kujutavad endast võimalikke valikuid, tippudeks on aga valiku tegemisel saadud tulemused. Puu läbimisel kontrollitakse tipuga seotud keelelisi tunnuseid (näiteks märksõnad, intonatsioonimärgid vms) ja vastavalt sellele otsustatakse, millist serva järgides edasi liigutakse. Sisetipud tähistavad sel viisil vahetulemusi, lehed lõplikke järeldusi. Puu lahti-mõtestamine on inimese jaoks lihtne, konstrueerimine aga mitte. Eksperimendi tulemusel leiti, et parimaid tulemusi andsid mõnel juhul morfoloogilisi atribuute, mõnel puhul aga märksõnu ja intonatsioonimärke kasutavad puud (Fišel ja Kikas, 2006). Otsustuspuude tuvastamistäpsus ületas eksperimendi käigus tehisnärvivõrkude oma, kuid hoolimata harva-esinevate dialoogiaktide väljajätmisest ja märgendamisel lisaks ka naaberaktidega arvesta-misest oli see vaid 44,7%.

Tõenäosuslike sufiksipuude (*probabilistic suffix tree*) meetod põhineb tähelepanekul, et dialoogiakte eristavad neis esinevad alamsõned ja -sekventsid (Kikas, 2007). Puukujulises struktuuris on tippudega seotud alamsõnede (näiteks lausungite) esinemise ja nende klassidesse kuulumise tõenäosused. Dialoogiaktide tuvastamisel on klassideks erinevad dialoogiaktid ja iga klassi elementideks tekstiüksuste järjendid (lausungite osad), millega üht või teist dialoogiakti väljendatakse. Puu servad on kaalutud, nende kaalud tähistavad sisetippude puhul üleminekutõenäosuseid, lehtede puhul aga väljunditõenäosuseid. Klassifitseerimisel läbitakse puud, otsides uuritavat sõne. Dialoogiaktide tuvastamine alamsekventside põhjal andis keskmiseks täpsuseks 46,8%. Eelneva ja järgneva dialoogiakti arvestamine tõstis saagise ja täpsuse 56,5%-ni. Samas leidus dialoogiakte, mida tuvastati halvasti, näiteks alternatiivküsimuse tuvastamisel oli saagis vaid 9,4% ja täpsus 19,1%.

3. Poolautomaatne märgendaja DAREC

Nagu eelnevalt mainitud, on nii manuaalsel kui ka automaatsel märgendamisel omad puudused: esimene neist nõuab palju aega ja tööjõudu, teine aga teeb vähemalt esialgu veel palju vigu. Loomulikult ei tekita teatud praktiliste ülesannete puhul väike hulk eksimusi suuremaid probleeme, kuid sageli on täpsus siiski väga oluline. Eeltoodut arvesse võttes on osutunud otstarbekaks keskenduda dialoogiaktide poolautomaatse märgendamise tarkvara loomisele ja edasiarendamisele. Mark Fišel koostas 2007. aastal poolautomaatse märgendaja DAREC (*Dialogue Act RECognizer*), mis teeb ära dialoogiteksti esialgse lausungiteks jagamise ja dialoogiaktide märgendamise, seejärel lubab kasutajal tehtut muuta ning vajadusel ka automaatset märgendamist valikuliselt uuesti korrata (Fishel, 2007b). Märgendaja sisendiks on tekstifail (*.txt*), kus voorud (aga mitte lausungid) on eraldi ridades, väljundiks tekstifail, kus voorud on lausungiteks jagatud, igale lausungile on omistatud märgend ja faili esialgsele nimele on lisatud laiend *.darec*.

3.1. Algoritm

DAREC kasutab Bayesi klassifitseerijat (Manning ja Schütze, 2003). Tegemist on meetodiga, mille puhul leitakse mingitele tunnustele tuginedes suurima tõenäosusega märgend või märgendid:

$$\hat{t} = \arg\max_{t} P(t \mid f_1, ..., f_n),$$

kus t on märgend ning f_1, f_2... f_n märgendavat lausungit iseloomustavad tunnused. Lausungile vastavaks valitakse suurima tõenäosusega märgend. Probleemiks on info suur kogus, kõigi võimalike tunnustega arvestamine on seetõttu reaalselt võimatu. Samas annab suur hulk neist praktikas liiga vähe informatsiooni ega paranda seepärast oluliselt väljundit. Seega valiti eksperimenteerimise käigus välja vaid mõned tunnused, mida kasutatakse: trigrammide tõenäosus, lausungi pikkus ning sõna ja märgendi tõenäosuste geomeetriline keskmine:

$$\hat{t} = \arg\max_{t} \left[P(t \mid t_{n-1}, t_{n-2}) \cdot P(t \mid k) \cdot \left(\prod_{i=1}^{k} P(t \mid \omega_i) \right)^{\frac{1}{k}} \right].$$

Valemis tähistavad t_{n-1} ja t_{n-2} märgendile t eelnevat kahte märgendit, k lausungi pikkust ning ω_i lausungis sisalduvat sõna.

Trigrammide kõrval katsetati ka bigrammide ja tetragrammidega, võrreldes esimestega parandasid trigrammid tulemusi oluliselt, tetragrammide puhul aga oli paremus marginaalne, samal ajal suurenes oluliselt mäluvajadus ja saadud mudeli suurus. Parimate tulemuste jaoks piisas ainult puhtal tekstil põhinevatest tunnustest, polnud vaja isegi morfoloogilise taseme tunnuste kasutamist: katsete käigus selgus, et näiteks lemmatiseerimise kasutamine polnud eriti informatiivne (Fishel, 2007c).

Algoritmi treenimisel ja testimisel kasutati Eesti dialoogikorpuses olevaid transkribeeritud kõnedialooge infotelefonile. Dialoogide hulk oli 835, kokku sisaldasid need 34319 lausungit, nende hulgas olid ka nn. tühjad lausungid nagu pausid, kommentaarid jms.

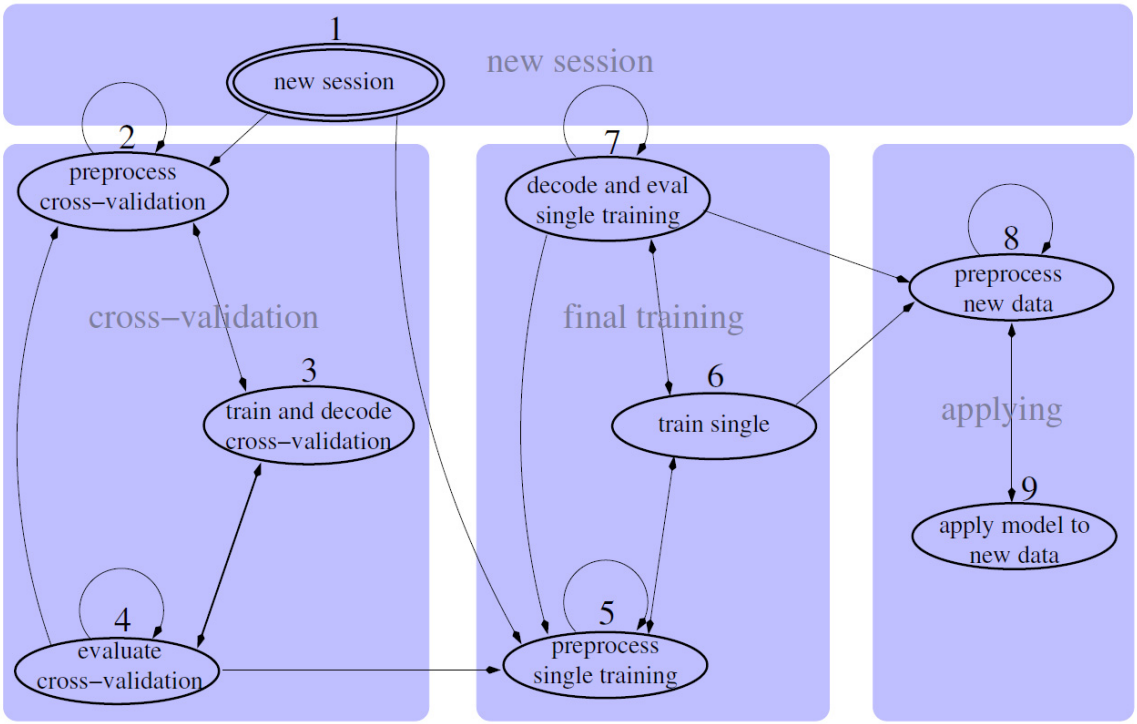
3.2. Ülesehitus

DAREC on statistilisel meetodil töötav märgendaja, mis kasutab Bayesi klassifitseerimis-meetodit ja eelnevalt märgendatud dialooge kui treeningandmeid (Fishel, 2007c). Märgendaja raamistik koosneb treening- ja märgendusosast (Fishel, 2007a). Treeningule võib vajadusel eelneda ristvalideerimine, mille puhul ei jagata andmeid lihtsalt treening- ja testandmeteks, vaid kasutatakse erinevaid sõltumatuid korpuse osi kord treening-, kord testandmete rollis (Manning ja Schütze, 2003). Ristvalideerimine on vajalik väiksema korpuse puhul, kuna lubab kasutada samu andmeid korduvalt (vt. joonis 2).

DARECi treenimisel luuakse uus sessioon, vajadusel tehakse ristvalideerimine koos eel-töötlusega, saadakse treenimise abil mudel ja kasutatakse seda uute andmete märgendamisel.

Klassifitseerija töötab programmeerimiskeeles *Perl* kirjutatud skriptidena, samuti käivitatakse uue dialoogi puhul nii lausungiteks jagamisel kui märgendamisel vastavad skriptid.

DARECi veebiliides on loodud programmeerimiskeeles *PHP*, mis suhtleb kasutajaga ja käivitab vastavalt vajadusele serveris olevaid *Perli* skripte. Kasutajamugavuse ja töökiiruse tõstmiseks on lisatud skripte *JavaScriptis*. Serverisse peab olema installeeritud vähemalt *PHP* 5. versioon koos failide avamiskäskude *file_get_contents* ja *fopen* lubamisega. Kasutajalt nõutakse graafilist veebibrauserit koos *JavaScripti* toega, muu tööks vajalik on lahendatud serveripoolsete vahenditega.

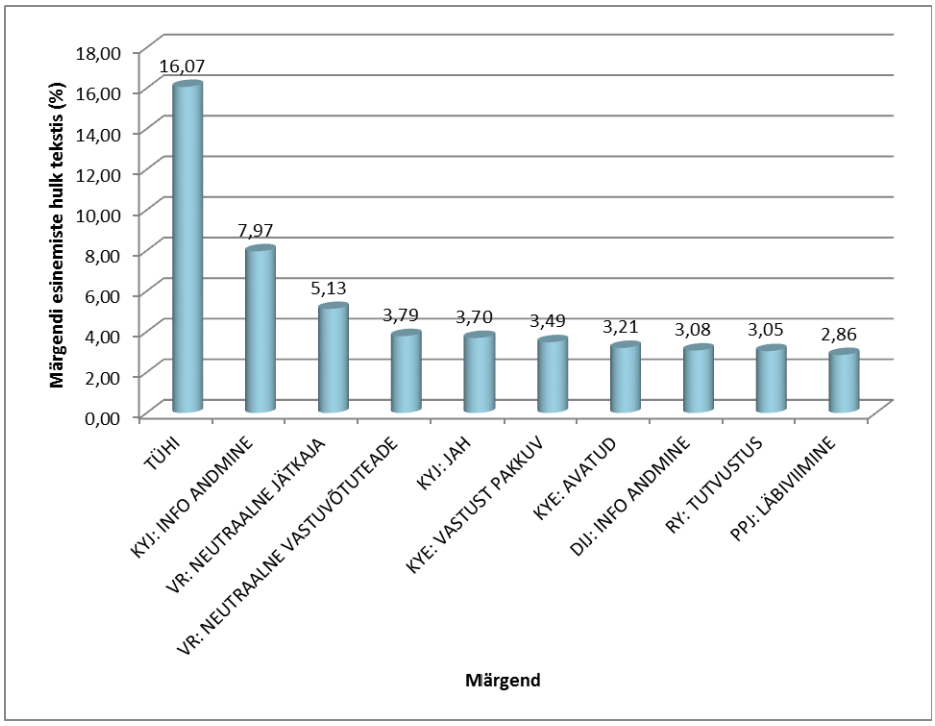


Joonis 2. *DARECi raamistiku skeem (Fishel, 2007a). Numbrid tähistavad algoritmi töö seisundit antud hetkel.*

3.3. Testimistulemused

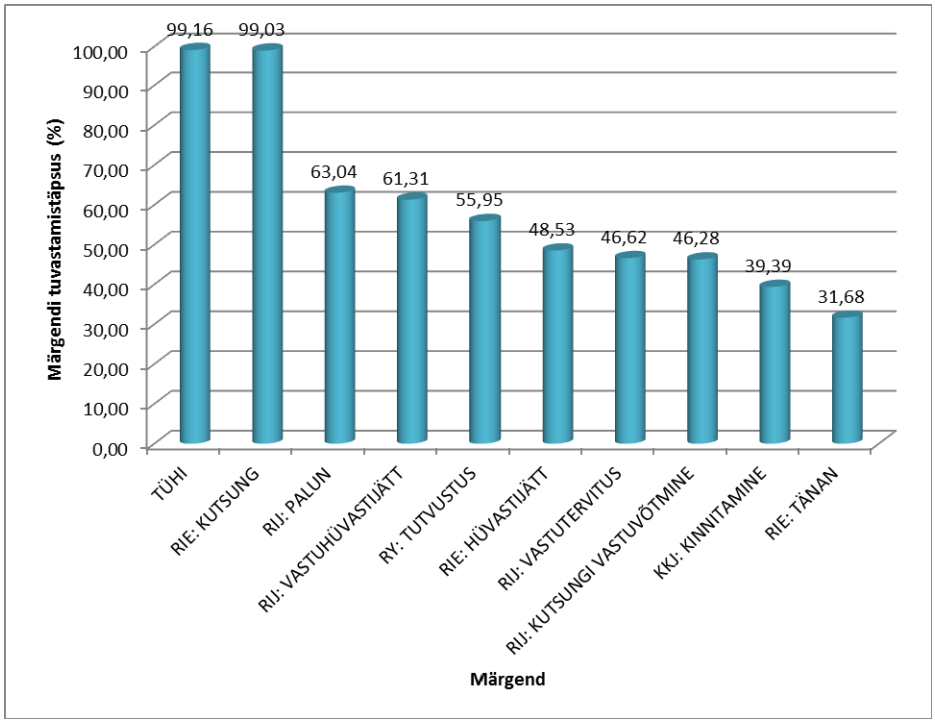
DARECit treeniti, kasutades dialoogikorpuse ristvalideerimist ja jagades selle jaoks andmed sajaks alagrupiks. Saadud mudelit hinnati, testimine andis keskmiseks saagiseks 64,7% ja täpsuseks 33,0% (vt. lisa 2). Ühest küljest on need madalad näitajad, põhjuseks muuhulgas ka treeningkorpuse väike maht, teisest küljest on tegemist siiski poolautomaatse märgendajaga, mis tähendab, et lingvisti hilisemad parandused on nagunii vajalikud. Samuti on arvutamise aluseks iga lausungi puhul kõige tõenäolisem märgend, kasutajaliideses aga esitatakse esmaseks valikuks viis kõige tõenäolisemat varianti. See tähendab, et parandamisel ei pea enamasti otsima sobivat märgendit mitte kõigi, vaid arvatavasti ainult pakutud viie hulgast.

Märgendite sagedus dialoogides oli erinev (vt. joonis 3). Kõige rohkem esines tühje lausungeid (16,07%), kahtteist lausungitüüpi dialoogidest ei leitud. Tühjadeks klassifitseeriti näiteks pause (tähistatud sulgudes oleva kümnendarvuga, mis näitab pausi pikkust sekundites) ja üldiselt ka kommentaare (tähistatud topeltsulgudes oleva tekstiga), samas kommentaar "((kutsung))" märgendatakse kui dialoogiakt RIE: KUTSUNG.



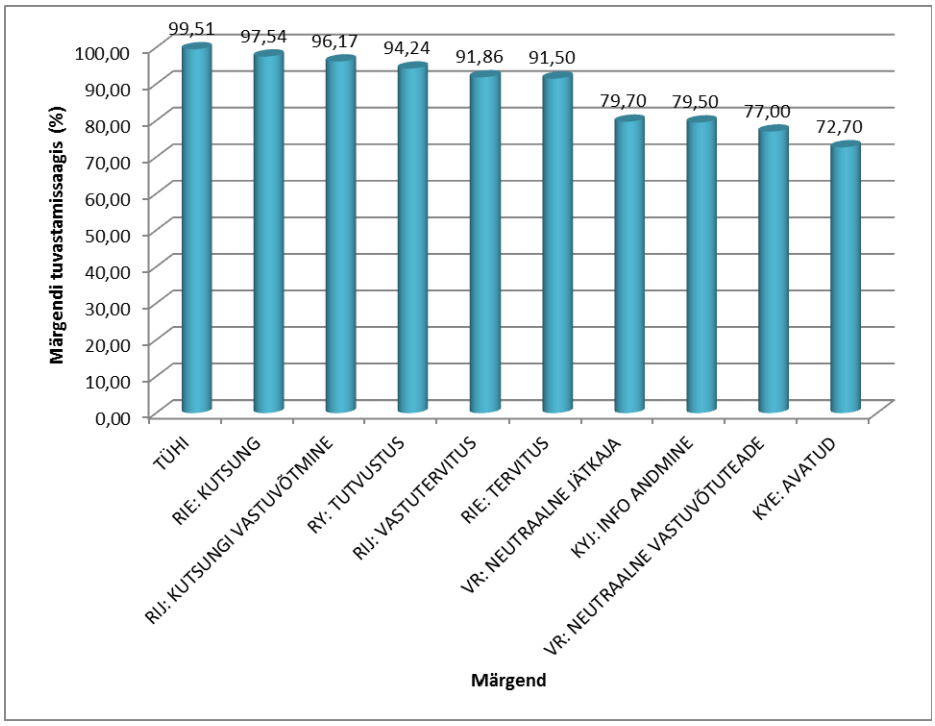
Joonis 3. *Kümme sagedasemat märgendit dialoogikorpuses.*

Kõige suurema täpsusega tuvastati peale tühja lausungi dialoogiakte RIE: KUTSUNG, RIJ: PALUN ja RIJ: VASTUHÜVASTIJÄTT (vt. joonis 4).



Joonis 4. *Kümme suurima täpsusega tuvastatud märgendit dialoogikorpuses.*

Suurima saagisega leiti lisaks tühjale lausungile dialoogiaktid RIE: KUTSUNG, RIJ: KUTSUNGI VASTUVÕTMINE ja RY: TUTVUSTUS (vt. joonis 5).



Joonis 5. *Kümme suurima saagisega tuvastatud märgendit dialoogikorpuses.*

3.4. Testijate hinnangud

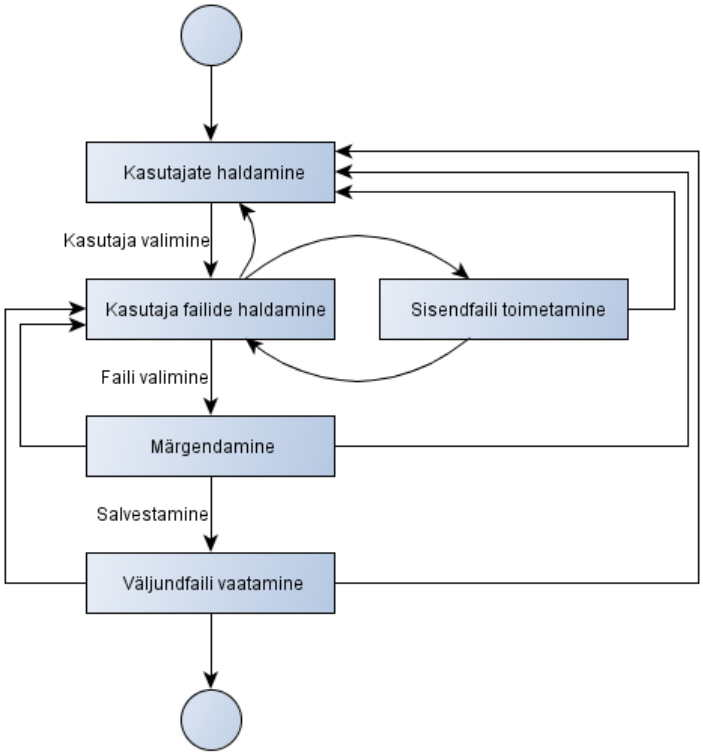
Detsembris 2007 ja aprillis 2008 paluti testijatel hinnata DARECit reaalse töö käigus. Testimisel tehtud märkuste põhjal selgus, et valesti märgendatud dialoogiaktide osakaal oli subjektiivsete hinnangute kohaselt 30-35%, neist omakorda 70% juhtudest oli õige valik esimese viie kõige tõenäolisema märgendi hulgas. Arvestades seda, et dialoogiaktide klassifitseerimine pole alati üheselt määratletav ning tegemist on poolautomaatse märgendajaga (inimese hilisemad parandused on kindlasti vajalikud), võib tulemusega rahul olla. Põhiliste negatiivsete külgedena tõid kasutajad välja probleemid kasutajaliidesega:

- ebaintuitiivsed kontrollelemendid;
- keeruline navigeeritavus;
- võimaluse puudumine vajadusel rida märgendamata jätta ja tühjast reast vabaneda;
- sama lausungi puhul märgendite omavahelise järjekorra muutumatus;
- võimaluse puudumine valitud märgendeid lukustada;
- ebaselge juhend.

4. Kasutajaliidese uuendamine

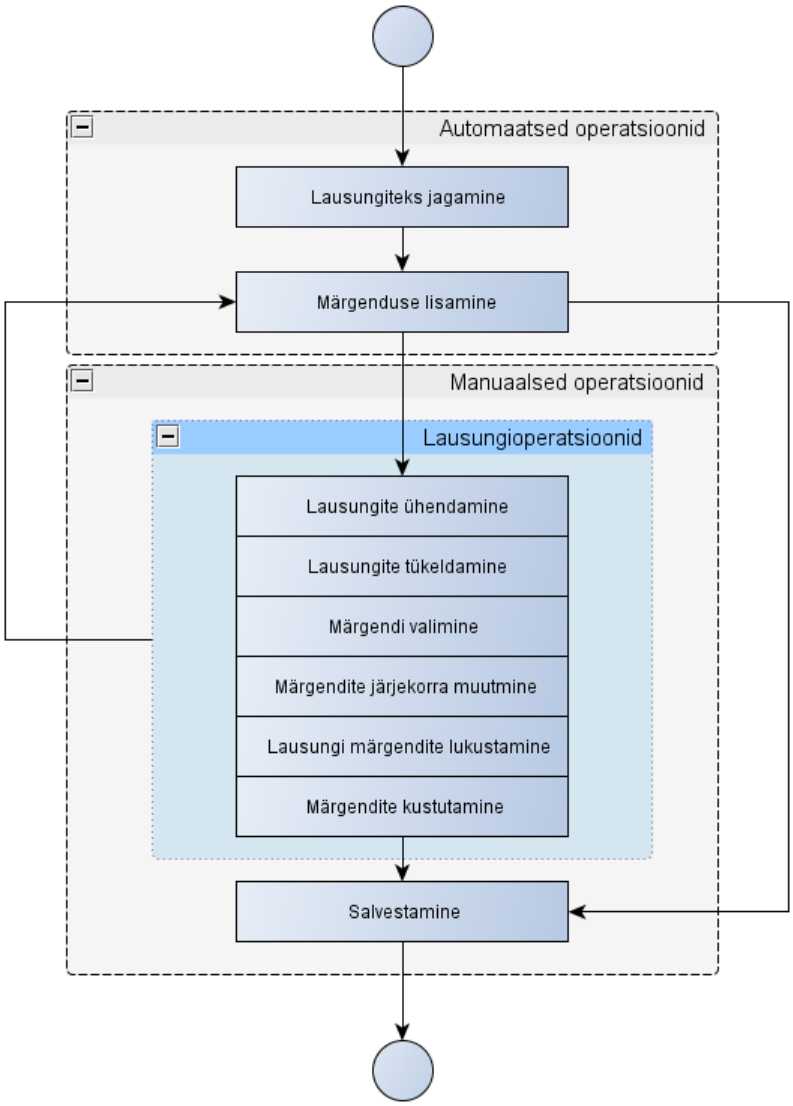
Testijate tagasiside DARECi kasutajaliidesele oli kokkuvõttes negatiivne, selle kohmakus tingis ka selle, et tarkvara ei hakatudki tegelikult kasutama. Seetõttu pöörati magistritöö praktilises osas esmajoones tähelepanu kasutajaliidese uuendamisele.

DARECi uus kasutajaliides (http://ats.cs.ut.ee/darec/www1/) koosneb viiest alalehest (vt. joonis 6): kasutajate haldamine, kasutaja failide haldamine, sisendfaili toimetamine, märgendamine ja väljundfailide vaatamine. Võrreldes eelmise kasutajaliidesega (http://ats.cs.ut.ee/darec/) on lehtede arv ja nende üldine ülesanne jäänud samaks, kuid suur osa loetletud funktsioonidest on uued.



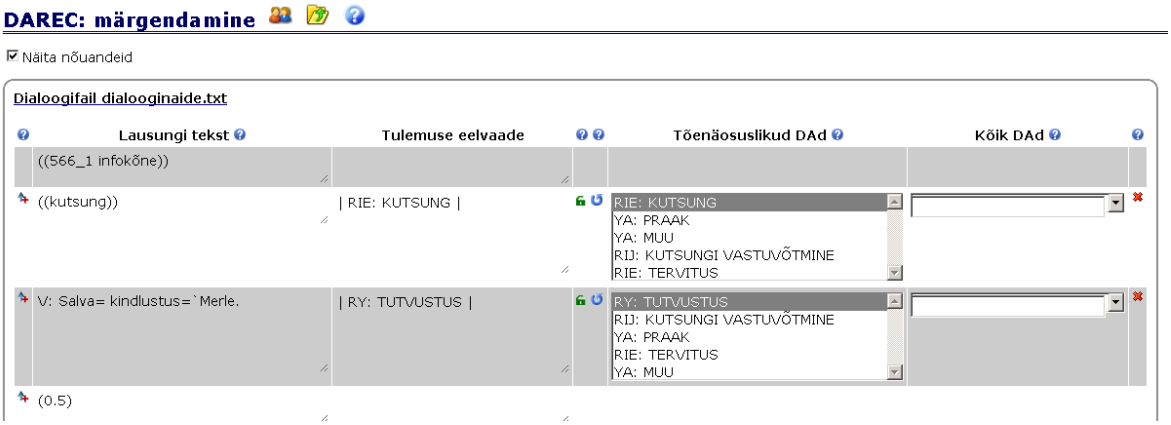
Joonis 6. *DARECi kasutajaliidese struktuur.*

Kasutajate haldamise puhul on võimalik neid lisada ja eemaldada. Kasutaja valimisel liigutakse failioperatsioonide lehele. Failide haldamine on kasutajakeskne: korraga näidatakse ainult konkreetse kasutaja sisend- ja väljundfaile. Faile saab vaadata ja kustutada, sisendfaile ka üles laadida, muuta ja märgendamisele saata (vt. joonis 7).



Joonis 7. *Märgendamise etapid.*

Märgendamislehel näidatakse esialgse automaatse märgendamise tulemust: dialoog on jagatud lausungiteks ning lisatud on tõenäolisim märgend (vt. joonis 8).



Joonis 8. *Lõik märgendamisele saadetud dialoogist märgendamislehel.*

Kasutaja saab parandada lausungiteks jaotust, valida uusi märgendeid, lukustada valikuid ja käivitada uuesti taasmärgendamist. Valmistööd saab salvestada ja seejärel vaadata eraldi aknas.

DARECi sihtrühmaks on reeglina suhteliselt hea arvutikasutusoskusega lingvistid, kes märgendavad dialoogiakte sageli ning peaksid olema seega süsteemi seisukohalt ekspert-kasutajateks. See loob ühest küljest võimaluse, teiselt poolt aga ka kohustuse lisada kasutaja-liidesesse elemente, mis pole esmakasutamisel intuitiivsed, kuid mis kiirendavad ja hõlbustavad tööd programmiga. Samas peab kasutajaliides toetama ka nende tööd, kes tarvitavad programmi harva.

Hea kasutajaliidese omadusteks on (Shneiderman ja Plaisant, 2009):

- kerge õpitavus (intuitiivsus, võimalikul kiire omandamine);
- efektiivsus (võimalus teha tööd kiiresti, erivõimalused ekspertkasutajale);
- meeldejäävus (pärast kasutamispausi oskuste kiire meeldetuletamine);
- eksimiskindlus (eksimisvõimaluste vähesus, tehtud vigade kerge parandamine);
- esteetilisus (disaini meeldivus).

Jakob Nielsen (Nielsen, 2012) toob välja kümme heuristikut veebilehtede hindamiseks:

- ülevaade süsteemi olekust (pidev reaalajas tagasiside momendi tegevuste kohta);
- süsteemi ja reaalse maailma kooskõla (loogiline tegevuste järjekord, kasutajale tuttava keele tarvitamine);
- kasutajapoolne kontroll süsteemi üle ja vabadus (võimalus alati süsteemist väljuda);
- ühtsus ja standardite järgimine (platvormile omased käitumismallid, ühtsus kasutata-vates sõnades, graafikas jms);
- vigade ennetamine (vähe võimalusi vigu teha);
- vähene koormus kasutaja mälule (vajaliku info pidev kättesaadavus);

- paindlikkus ja efektiivsus (ekspertkasutajale vajalikud kiirkäsklused sagedasemate operatsioonide tegemiseks);
- esteetiline ja minimalistlik disain (värvilahendus, selged dialoogid, ülearuse info puudumine);
- vigade äratundmine, diagnoosimine ja nende lihtne parandamine (selged inimsõbralikud veateated, soovitused edasiseks tegevuseks);
- abiinfo ja dokumentatsioon (lihtne otsida, asjakohane).

Nendele põhimõtetele tuginedes parandati käesoleva töö raames DARECi kasutajaliidest. Järgnevalt on kirjeldatud tehtud muutusi ja nende vastavust Jakob Nielseni poolt pakutud heuristikutele.

4.1. Ühtsus ja standardite järgimine

Kuigi paljud DARECi kasutajad on igapäevased arvutikasutajad, on iga programmi kasutajaliides veidi erinev ja nõuab seetõttu aega harjumiseks. Tänu kasutajaliidese intuitiivsusele on kohanemine kergem nii esimesel kokkupuutel kui ka hilisemal pika pausi järel uuesti kasutama hakkamisel.

DARECi kasutajaliidese uuendamisel pöörati tähelepanu selgete ja konkreetsete ikoonide valikule, mis oleksid semantiliselt informatiivsed ja harjumuspärased (vt. joonis 9). Samas kohas, kuid erinevat olekut kajastavad ikoonid on eristatavad värvi järgi (vt. joonis 9 (b) ja (c)).

(a) (b) (c) (d) (e) (f) (g) (h)

Joonis 9. *Märgendite kustutamine (a), märgendi avatud (b) ja suletud (c) olek, taasmärgendamine (d), sisendfaili vaatamine (e) ja muutmine (f), failide avamine (g) ja abiinfo (h).*

Sarnaste operatsioonide jaoks (näiteks kasutaja või faili kustutamine, samuti kasutaja või faili lisamine) kasutati vastavale alusikoonile (vt. joonis 10 (a) ja (d)) ühesuguste elementide lisamist (vt. joonis 10).

(a) (b) (c) (d) (e) (f) (g)

Joonis 10. *Kasutajatega ja failidega seotud lehtede ikoone: kasutajate nimekiri (a), kasutajate lisamine (b) ja kustutamine (c), sisendfailide nimekiri (d), nende lisamine (e) ja kustutamine (f), darec-laiendiga väljundfailide nimekiri (g).*

Ühtlane kujundus lubab parandada töökiirust: kasutaja, olles harjunud leidma objekte ühel lehel teatud kohas ja teatud kujundusega, suudab kergemalt leida sarnaseid objekte analoogsetes kohtades teistel lehtedel. Ühtlase struktuuriga pealkirjad nii lehel kui selle tiitelribal (algavad programmi nimega DAREC ja jätkuvad konkreetset alalehte iseloomustava tekstiga) annavad kiiresti ülevaate sisust.

4.2. Süsteemi ja reaalse maailma kooskõla

DARECi operatsioonide järjekord (vt. joonis 6) imiteerib dialoogiaktide käsitsi märgendamise skeemi: pärast õige kasutaja valimist jõutakse failioperatsioonide juurde, neist ühe või mitme sisendfaili valimisel tehakse algne automaatne märgendamine ja pakutakse võimalust tulemust parandada.

Märgenduslehel on lausungid esitatud ridade kaupa, töövahendite järjekord vastab tavalisele märgendamisele: lausungite jagamine ja ühendamine, märgendi eelvaade koos lukustus- ja taasmärgendusvõimalusega, vajadusel uue märgendi valik või kõigi kustutamine.

DARECis on kasutatud sama terminoloogiat, mida tarvitatakse keeleteaduses dialoogiaktide märgendamisel, see teeb süsteemi kasutamise lingvistidele mugavamaks ja kiiremini õpitavaks.

4.3. Kasutajapoolne kontroll süsteemi üle ja vabadus

Süsteem lubab kasutajal liikuda igalt leheküljelt hierarhilises mõttes üldisema lehe poole. See võimaldab igal ajal mugavalt katkestada momendil käimasoleva tegevuse või lõpetada töö. Samade operatsioonide puhul on kontrollelemente ka dubleeritud (näiteks failide eelvaade ikooni ja failinime abil või lausungite taasmärgendamine nii ridade sees kui leheküljelõpus), et kasutajal oleks vabadus leida enda jaoks mugavaim variant.

4.4. Ülevaade süsteemi olekust ja vähene koormus kasutaja mälule

DARECi puhul on tegemist mitut alalehte kasutava süsteemiga, seega on oluline anda infot momendi asukohast, seda eesmärki täidavad pealkirjad. Samuti on märgendamislehel näha märgendamise tulemus, mis uue märgendi valimisel või olemasolevate kustutamisel reaalajas muutub.

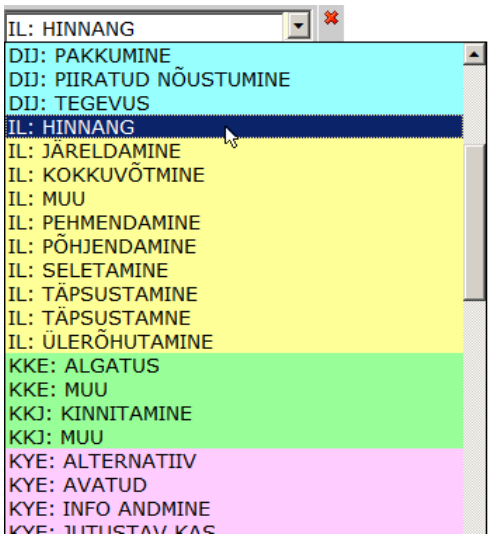
4.5. Paindlikkus ja efektiivsus

Efektiivseks tööks programmiga on vajalik mugav navigeerimine erinevate komponentide vahel. Sel eesmärgil on lisatud lehtedele otseteed liikumiseks kõigile üldisematele tasemetele (näiteks faili toimetamislehelt kasutaja failide juurde ja samuti otse kasutajate haldamisliidesesse) (vt. joonis 6). Samuti on kõikjal kättesaadav abiinfo. Reeglina toimub töö ühes programmiaknas, kuid operatsioonid, mille puhul on oluline algse lehe säilimine (abiinfo vaatamine, faili salvestamine), suunatakse uude aknasse.

Teisest küljest on oluline navigeerimine ka lehesiseselt: pikkade failide puhul taasmärgendamise käivitamiseks lehekülje kerimine on ebaefektiivne. Seetõttu on taasmärgendusikoon toodud iga lausungi juurde, mis lubab ühest küljest skripti kiiresti käivitada, teisest küljest aga pärast mugavalt tagasi liikuda viimati muudetud reale.

DARECi potentsiaalsed kasutajad on reeglina need, kes märgendavad dialoogiakte regulaarselt. Sellisel juhul on lisaks kergesti omandatavusele oluline tõsta kasutuskiirust lühikäsklustega. DARECis on sel põhjusel loobutud eraldi nuppudest valitud märgendite järjekorra muutmiseks, vaid asendatud märgendi loetelus ülespoole tõstmisega topeltklõpsu abil, vähendades nii hiire liikumise trajektoori.

Samuti pole tõenäosuslike märgendite nimekirja liigendatud alarühmadesse. Selline lähenemine tuleks kasuks küll ülevaatlikkusele, kuid nõuaks lisaoperatsioone alammenüüde avamisel või rohkem ekraaniruumi. Õige märgendi kiiremaks leidmiseks on alagruppidel ühine taustavärv, samuti saab nimekirja kerimise kiirendamiseks valida klaviatuuri abil märgendi esitähte (vt. joonis 10).



Joonis 10. *Kõigi märgendite hulgast sobiva valimine.*

Märgenduslehel on sageli järjest kasutatavad ikoonid paigutatud lähestikku: näiteks on pärast märgendi lukustamist hea teiste lausungite puhul parema tulemuse saamiseks korrata automaatset märgendust.

4.6. Veahaldus

Nii DARECi raamistik kui ka kasutajaliides on varustatud selgete veateadetega. Kasutajat teavitatakse analoogsete vigade puhul sarnases stiilis, kohas ja kujunduses infoga (vt. näide 4).

```
Fail nimega uus.txt on juba olemas.
Kasutaja nimega kasutaja on juba olemas.
```
Näide 4. *Veateated.*

Nii kasutaja- kui failinimed peavad koosnema ainult inglise keele tähestiku tähtedest, numbritest ja alakriipsudest. Reeglitele mittevastavad nimed teisendatakse sobivale kujule (ebasobivale sümbolile sarnaseim täht või alakriips) ja teavitatakse sellest kasutajat (vt. näide 5).

```
Uus kasutaja Karl on lisatud.
Uus kasutaja Ülo (Ulo) on lisatud.
```
Näide 5. *Kinnitusteated.*

Ohtlikumate tegevuste puhul (näiteks kasutajate ja failide kustutamine, lausungite ühendamine ja jagamine) küsitakse kasutajalt enne operatsiooni sooritamist kinnitust (vt. joonis 11).

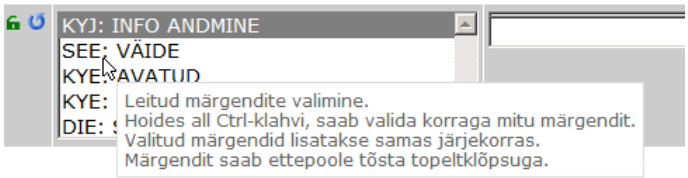
Joonis 11. *Kinnituse küsimine enne lausungi jagamist.*

Kuigi teised tegevuste ikoonid paiknevad reas üksteise lähedal, on valitud märgendite kustutamise nupp viidud rea lõppu, et vähendada juhuslikke vigu.

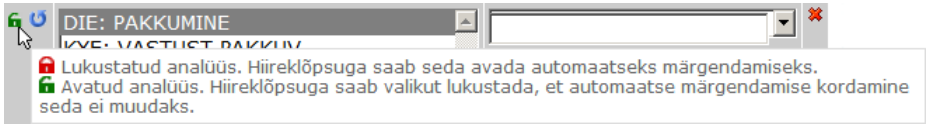
4.7. Abiinfo ja dokumentatsioon

DAREC pakub abiinfot mitmel erineval viisil. Tavapärane manuaal on kättesaadav kõigilt lehtedelt, enamasti on viidatud leheküljega haakuvale alalõigule. Info on manuaalis jagatud ülesannete järgi alagruppidesse ja järjestatud vastavalt tavalisele tööjärjekorrale. Eraldi on lisatud ka ikoonide loetelu (vt. lisa 4).

Lisaks traditsioonilisele manuaalile on märgendamislehel kasutusel ka sisse- ja väljalülitatav kontekstitundlik abiinfo. Sõltuvalt hiirekursori asukohast antakse raami sees infot valitud elemendiga seonduvatest operatsioonidest (vt. joonis 12).



(a)



(b)
Joonis 12. *Kontekstitundlik abiinfo dialoogiaktide valikul (a) ja valiku lukustamisel (b).*

4.8. Esteetiline ja minimalistlik disain

DARECi puhul on tegemist töövahendiga, mis peab esmajärjekorras olema efektiivne, meeldivus ja atraktiivsus on teisejärgulised. Seetõttu on valitud värvid, mis ei häiri silma ka

pikemaajalisel tööl (halltoonid, sinise erinevad varjundid). Samas on erksamaid värve lisatud harvemini kasutatavatel lehtedel (näiteks kasutajate haldamine) ja sagedamini kasutatavate ikoonide puhul. Olulise olekuindikaatorina kasutatav lukustusikoon eristub samuti lisaks kontuurile ka värvi poolest (vt. joonis 9 (b) ja (c)).

5. Tulemused

5.1. Kasutajaliidese probleemide lahendamine

Eelmise kasutajaliidese probleemid (vt. peatükk *3.4. Testijate hinnangud*) on lahendatud. Lausungit saab jätta märgendamata, tühje ridu saab kustutada kas sisendfaili redigeerimisega või (juhul, kui märgendamine on osaliselt juba toimunud) nende ühendamisega eelmise lausungiga. Lisatud on ühele lausungile mitme märgendi puhul nende omavahelise järjekorra muutmine. Ikoonid (sh. märgendite lukustamisikoonid) on selgelt eristuvad. Abiinfot on oluliselt täiendatud: struktureeritud ja illustreeritud manuaali kõrval on ka kontekstitundlik abiinfo ning märgendamislehel tulpade seletused.

5.2. Testijate hinnangud

DARECi kasutajaliidest paluti hinnata süsteemiga tulevikus reaalselt tööle hakkavatel lingvistidel, kes vastasid pärast kasutamist liidest puudutavale küsimustikule (vt. lisa 3). Korraldati heuristiline hindamine (Matera, 2006), ankeedi eeskujuks oli Gary Perlmani küsimustik (Perlman, 1997). Küsimused hõlmasid Jakob Nielseni kasutatavuse heuristikuid (Nielsen, 2012), et saada ülevaadet, kui hästi süsteem neile vastab.

Küsimustik koosnes kolmest osast, need hindasid õpitavust, kasutajasõbralikkust ning tagasisidet ja veakäsitlust. Osade igale alapunktile sai anda hinnangu täisarvulisel skaalal 1 kuni 5, kus 1 tähistas väga halba ja 5 väga head hinnangut. Lisaks sellele oli kasutajal võimalus anda iga osa puhul vabas vormis kommentaare ning lõpuks lisada kogu süsteemi puudutavaid positiivseid ja negatiivseid märkusi.

Küsimustik saadeti kuuele inimesele, neist viis vastas sellele. Üldiselt olid hinnangud positiivsed, alapunktidele antud hinded olid enamasti 4 ja 5 Kolmest üldisest punktist oli kõige kõrgemalt hinnatud kasutajasõbralikkust (keskmine 4,8), sellele järgnesid õpitavus (4,6) ning tagasiside ja veakäsitlus (4,3). Konkreetsetest alapunktidest said parima keskmise hinnangu kasutajapärasus ja disaini meeldivus (mõlemad 5), liidese selgus ning tegevuste ja juhtelementide kokkusobivus said keskmiseks hindeks 4,8. Nõrgima keskmise tulemuse sai vigade ennetamine (4).

Lisaks hinnangutele andsid palju infot kasutajate kommentaarid. Tänu sellele selgusid süsteemi edasised parandus- ja täiendusvõimalused. Raskusi tekitas kasutajate arvates ühele

lausungile mitme märgendi lisamise süsteem, mõnel puhul polnud seda võimalust üldse üles leitud. Ka kordusmärgendamise järel viimasena vaadeldud reale tagasiminekut tähistava noole leidmine oli ebaintuitiivne. Ülearust tööd tekitas kasutajate arvates ka ohtlike operatsioonide (lausungite jagamine ja ühendamine) puhul kinnituse küsimine, hea ideena pakuti välja, et hoiatusi võiks saada soovi korral välja lülitada.

Positiivsetest külgedest toodi välja liidese loogilisus ja intuitiivsus, ikoonide semantiline selgus ja juurde lisatud selgitused. Mitmel korral mainiti tunnustavalt kontekstitundlikku abiinfot, mis andis põhjalikumat teavet iga objekti kohta kasutamise ajal.

6. Edasiarendusvõimalused

Dialoogide poolautomaatse märgendaja DAREC edasiarendusvõimalusi on mitmesuguseid. Arvestades sellega, et tegemist on statistilist meetodit kasutava süsteemiga, tuleb suurendada treeningandmete hulka, mis kokkuvõttes peaks sama algoritmi juures andma paremaid tulemusi. Samuti peab treeningkorpust laiendama ka sisulises mõttes: suuliste inimestevaheliste dialoogide kõrvale tuleb lisada ka kirjalikke inimese ja arvuti vahelisi (tegelikke ja simuleeritud) dialooge.

Algoritmist lähtudes tuleb eksperimenteerida ka teiste erinevate tunnustekomplektidega. Võimalik, et uuendatud korpusel annavad muud tunnused parema tulemuse kui eelmiste testimiste puhul, näiteks trigrammide asendamine tetragrammidega võib parandada tulemusi olulisel määral rohkem. Edaspidi tuleb kõne alla ka algoritmi modifitseerimine, sh. statistilise meetodi kombineerimine reeglitega, mille sõnastamiseni on lingvistid jõudnud dialoogikorpuse analüüsi tulemusel.

Kasutajaliidese seisukohalt on vaja lisada juurde ekspertkasutajale sobivaid võtteid, et töökiirust parandada. Kasutusmugavust tõstaks ka töölauaprogrammi loomine, mis märgendamisel kasutaks võrgu kaudu DARECi raamistikku, kuid lubaks brauseris töötavast veebiversioonist paindlikumat klahvikombinatsioonide ja otseteede tarvitamist ning kasutajatele isikliku töökeskkonna kohandamist. Testijad tundsid puudust märgendatud faili taasredigeerimise võimalusest: juba salvestatud märgendatud faili peaks saama DARECis uuesti avada ja märgendamist jätkata.

Praktilisest vaatenurgast saab DARECit katsetada reaalsetes süsteemides (näiteks mõnes dialoogsüsteemis) mitte kui poolautomaatset, vaid kui automaatset dialoogiaktide tuvastajat. See nõuab ühest küljest küll mitmesuguste veaparandusvõtete rakendamist, kuid selle kasutamise jälgimine annab infot DARECi parandamiseks.

Kokkuvõte

Magistritöö eesmärgiks on kirjeldada Eesti dialoogikorpuse ressursside hetkeolukorda ja dialoogide märgendamiseks kasutatavaid vahendeid ning arendada edasi poolautomaatset märgendajat DAREC.

Töös on kirjeldatud dialoogide ülesehitust, Eestis kasutatavat dialoogiaktide märgendamistüpoloogiat EDiT, samuti nii manuaalse kui ka automaatse märgendamistarkvara positiivseid ja negatiivseid külgi.

2007. aastal Mark Fišeli poolt loodud dialoogiaktide poolautomaatne märgendaja DAREC põhineb statistilisel meetodil. Esimeste testijate hinnangud olid küllaltki positiivsed seoses DARECi töö sisuliste tulemustega, kuna see kergendas oluliselt isegi väiksema täpsusega tuvastamise puhul õigete märgendite leidmist kuid negatiivsed seoses kasutajaliidesega. Viimasele heideti ette ebamugavust, ebapiisavat abiinfot, mõnede vajalike operatsioonide puudumist jms. Nende arvamuste põhjal kõrvaldati või leevendati käesoleva töö raames nimetatud puudusi, võttes aluseks heade kasutajaliideste loomise põhimõtted. Seejärel paluti dialoogide märgendajatel testida uut kasutajaliidest ning hinnangutest selgus, et süsteemi kasutajamugavus on olulisel määral kasvanud. Kõrgeimalt hinnati kasutajapärasust ja disaini ning kontekstitundlikku abiinfot, kuid samuti esitati erinevaid ideid süsteemi efektiivsemaks muutmiseks.

Töös tuuakse ka võimalusi DARECi edasiarendamiseks: tuvastamistäpsuse ja -saagise tõstmine algoritmi parandamise ja dialoogikorpuse suurendamise läbi, ekspertvõimaluste lisamine jne.

Recognition of Dialogue Acts in the Estonian Dialogue Corpus: Overview of Resources and Software Development

Sven Aller

Summary

The aim of the thesis was to describe the present situation of the resources of the Estonian Dialogue Corpus and markup tools for dialogue acts as well as to develop the semi-automatic dialogue act markup tool DAREC.

The thesis describes the structure of dialogue acts, the markup typology EdiT used in the Estonian Dialogue Corpus as well as the positive and negative sides of manual and automatic markup tools.

The semi-automatic markup tool DAREC created by Mark Fishel in 2007 is based on a statistical method. Linguists' first opinions were quite positive in terms of markup results. On the other hand, testers were critical about some features of the user interface, such as not beeing user-friendly, a poor manual, the absence of some important functions. Based on the users' opinions and principles of creating good user interfaces most of the weaknesses were eliminated. The heuristic tests revealed that the usability of DAREC had remarkable improved. The most highly scored features included its user-friendliness, design and contextual help. At the same time various ideas for making the system more effective were suggested.

The thesis also suggests several possibilities for developing DAREC, for example, increasing precision and recall of recognition by improving algorithm as well as the size of the dialogue corpus and adding more expert features.

Kasutatud kirjandus

1. Ai, Hua; Raux, Antoine; Bohus, Dan; Eskenazi, Maxine; Litman, Diane (2007). Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 124-131. Association for Computational Linguistics Stroudsburg, USA.
2. Bunt, Harry; Alexandersson, Jan; Carletta, Jean; Choe, Jae-Woong; Fang, Alex Chengyu; Hasida, Koiti; Lee, Kiyong; Petukhova, Volha; Popescu-Belis, Andrei; Romary, Laurent; Soria, Claudia; Traum, David (2010). Towards an ISO standard for dialogue act annotation. *Seventh conference on International Language Resources and Evaluation*, pp. 2548-2555. Tilburg Centre for Cognition and Communication, Tilburg University.
3. Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics 22 (2)*, pp. 249-254.
4. Dybkjær, Laila; Bernsen, Niels Ole (2000). The MATE markup framework. *SIGDIAL '00 Proceedings of the 1st SIGdial workshop on Discourse and dialogue 10*, pp. 19-28. Association for Computational Linguistics Stroudsburg, USA.
5. *Estonian Dialogue Corpus (EDiC)* (2012). http://www.cs.ut.ee/~koit/Dialoog/EDiC. Kontrollitud 23.04.2012.
6. Eskor, Liina (2004). *Dialoogiaktid ja suhtlusstrateegiad: eesti dialoogikorpuse analüüs.* Magistritöö. Tartu Ülikool.
7. Fishel, Mark (2007a). *DARec framework manual* (darec.manual.pdf).
8. Fishel, Mark (2007b). Machine Learning Techniques in Dialogue Act Recognition. *Estonian Papers in Applied Linguistics 3*, pp. 117-134.
9. Fishel, Mark (2007c). Complex Taxonomy Dialogue Act Recognition with a Bayesian Classifier. *Proceedings: DECALOG'2007 Workshop on the Semantics and Pragmatics of Dialogue*, pp. 161-162. Rovereto, Italy.
10. Fišel, Mark; Kikas, Taavet (2006). Dialoogiaktide automaatne tuvastamine. *Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6*, lk. 233-245.
11. Gerassimenko, Olga; Kasterpalu, Riina; Koit, Mare; Rääbis, Andriela, Strandson, Krista (2010). Direktiivsed aktipaarid eestikeelsetes infodialoogides ja nende automaatne tuvastamine. *Eesti Rakenduslingvistika Ühingu aastaraamat 6*, lk. 67-86. Tallinn: Eesti Keele Sihtasutus.

12. Griol, David; Callejas, Zoraida; López-Cózar, Ramón (2009). A comparison between dialog corpora acquired with real and simulated users. *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, pp. 326–332. Association for Computational Linguistics Stroudsburg, USA.
13. Hennoste, Tiit; Koit, Mare; Kullasaar, Maret; Rääbis, Andriela; Vutt, Evely (2002). Eesti dialoogikorpuse loomise probleemid. *Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3*, lk. 143-160.
14. Hennoste, Tiit; Rääbis, Andriela (2004). *Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs.* Tartu Ülikooli Kirjastus.
15. Jurafsky, Daniel; Martin, James H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 2nd edition. Prentice-Hall.
16. Kikas, Taavet (2007). *Dialoogiaktide tuvastamine eestikeelsetes dialoogides sufiksipuude abil.* Magistritöö. Tartu Ülikool.
17. Klüwer, Tina; Uszkoreit, Hans; Xu, Feiyu (2010). Using Syntactic and Semantic based Relations for Dialogue Act Recognition. *Coling 2010: Poster Volume*, pp. 570-578. Beijing.
18. Koit, Mare (2003). Märgendatud dialoogikorpus kui keeleressurss. *Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi Toimetised, 12*, Tallinn, lk. 119-136.
19. Manning, Christopher D.; Schütze, Hinrich (2003). *Foundations of Statistical Natural Language Proccessing.* MIT Press, Massachusetts, USA.
20. Matera, Maristella; Rizzo, Francesca; Carughi, Giovanni Toffetti (2006). Web Usability: Principles and Evaluation Methods. *Web Engineering.* Springer Berlin Heidelberg.
21. McTear, Michael F. (2004). *Spoken dialogue technology: toward the conversational user interface.* London: Springer Verlag.
22. Nielson, Jakob (2012). *Ten Usability Heuristics.* http://www.useit.com/papers/heuristic/heuristic_list.html. Kontrollitud 02.05.2012.
23. Nurmsalu, Evely (2001). *Eestikeelse dialoogikorpuse märgendamistarkvara.* Magistritöö. Tartu Ülikool.
24. Perlman, Gary (1997). *Tutorial on Practical Usability Evaluation.* Conference in human factors in computing systems. Atlanta, USA.

25. Shneiderman, Ben; Plaisant, Catherine; Cohen, Maxine; Jacobs, Steven (2009). *Designing the User Interface. Strategies for effective human-computer interaction*. 5th edition. Pearson Addison-Wesley.
26. Stolcke, Andreas; Ries, Klaus; Coccaro, Noah; Shriberg, Elizabeth; Bates, Rebecca; Jurafsky, Daniel; Taylor, Paul; Martin, Rachel; Van Ess-Dykema, Carol; Meteer, Marie (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics.* September 2000, Vol. 26, No. 3, pp. 339-373.
27. *Transkriptsioonimärgid* (2012). http://www.cl.ut.ee/suuline/Transk.php. Kontrollitud 30.04.2012.
28. Treumuth, Margus (2004). *Eesti dialoogikorpus ja selle töötlemise tarkvara.* Magistritöö. Tartu Ülikool.
29. Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2011). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, Amsterdam.

Lisad

Lisa 1. Dialoogiaktide tüpoloogia EdiT

EDiT tüüpide loend seisuga mai 2012.

I. Naabruspaare 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
- Kutsesignaal
  - esiliige
    RIE: KUTSUNG
  - järelliige
    RIJ: KUTSUNGI VASTUVÕTMINE
- Lõpetamine
  - esiliige
    RIE: LÕPUSIGNAAL
  - järelliikmed
    RIJ: LÕPETAMISE VASTUVÕTMINE
    RIJ: LÕPETAMISE TAGASILÜKKAMINE
- Soov rääkida

    - esiliige
      RIE: SOOV RÄÄKIDA
    - järelliikmed
      RIJ: NÕUSTUMINE
      RIJ: MITTENÕUSTUMINE
      RIJ: SUUNAMINE
      RIJ: MITTESUUNAMINE
- Identifitseerimine
    - esiliige
      RIE: IDENTIFITSEERIMINE
    - järelliige
      RIJ: IDENTIFITSEERIMINE
- Muu
    - esiliige
      RIE: MUU
    - järelliige
      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
    - järelliikmed
      PPJ: LÄBIVIIMINE
      PPJ: MUU
4. Vastuse tingimuste täpsustamine
    - esiliikmed
      VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE
      VTE: MUU
    - järelliikmed
      VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE
      VTJ: MUU
5. Kontakti kontroll
    - esiliikmed
      KKE: ALGATUS
      KKE: MUU
    - järelliikmed
      KKJ: KINNITAMINE
      KKJ: MUU
6. 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

7. Küsimused
   - esiliikmed
     KYE: AVATUD
     KYE: JUTUSTAV KAS
     KYE: SULETUD KAS
     KYE: VASTUST PAKKUV
     KYE: ALTERNATIIV
     KYE: MUU
   - järelliikmed
     KYJ: INFO ANDMINE
     KYJ: JAH
     KYJ: EI
     KYJ: NÕUSTUV EI
     KYJ: MITTENÕUSTUV JAH
     KYJ: ALTERNATIIV: ÜKS
     KYJ: ALTERNATIIV: MÕLEMAD
     KYJ: ALTERNATIIV: KOLMAS VALIK
     KYJ: ALTERNATIIV: EITAV
     KYJ: INFO PUUDUMINE
     KYJ: KEELDUMINE
     KYJ: EDASILÜKKAMINE
     KYJ: VASTUS ALTERNATIIVINA
     KYJ: TEGEVUS
     KYJ: MUU
8. 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: MUU
4. Rituaalsed üksikaktid
   - RY: ÜLEANDMINE
   - RY: TUTVUSTUS
   - RY: ÄRATUNDMINE
   - RY: KONTAKTEERUMINE
   - 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: PRAAK

Lisa 2. DARECi testimistulemused

Lausungite arv: 34319
Keskmine täpsus: 32,95%
Keskmine saagis: 64,68%

Märgend	Sagedus	Täpsus	Saagis
EMPTY	16,07%	99,16%	99,51%
KYJ: INFO ANDMINE	7,97%	19,77%	79,50%
VR: NEUTRAALNE JÄTKAJA	5,13%	26,56%	79,70%
VR: NEUTRAALNE VASTUVÕTUTEADE	3,79%	20,47%	77,00%
KYJ: JAH	3,70%	18,58%	38,77%
KYE: VASTUST PAKKUV	3,49%	8,52%	68,66%
KYE: AVATUD	3,21%	9,87%	72,70%
DIJ: INFO ANDMINE	3,08%	20,95%	62,51%
RY: TUTVUSTUS	3,05%	55,95%	94,24%
PPJ: LÄBIVIIMINE	2,86%	11,44%	45,14%
IL: TÄPSUSTAMINE	2,54%	9,87%	62,53%
VR: NEUTRAALNE INFO OSUTAMINE UUEKS	2,46%	21,81%	72,29%
RIE: TÄNAN	2,06%	31,68%	66,29%
RIE: TERVITUS	2,04%	25,77%	91,50%
DIE: SOOV	2,00%	14,66%	66,88%
RIE: KUTSUNG	1,86%	99,03%	97,54%
RIJ: KUTSUNGI VASTUVÕTMINE	1,86%	46,28%	96,17%
RIJ: VASTUTERVITUS	1,72%	46,62%	91,86%
KYE: JUTUSTAV KAS	1,58%	9,67%	51,61%
PPE: ÜLEKÜSIMINE	1,58%	8,12%	45,25%
RIJ: PALUN	1,53%	63,04%	70,38%
YA: INFO ANDMINE	1,38%	6,19%	40,19%
YA: MUU	1,27%	5,85%	42,69%
KYE: SULETUD KAS	1,24%	6,88%	39,84%
YA: PRAAK	1,11%	5,37%	44,72%
VR: NEUTRAALNE PIIRITLEJA	1,06%	9,95%	52,77%
VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE	1,06%	3,61%	22,89%
KYE: VASTUSE TINGIMUSTE TÄPSUSTAMINE	0,95%	8,88%	36,90%
VR: MUU	0,95%	3,87%	39,68%
VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE	0,92%	9,05%	30,28%
RIE: HÜVASTIJÄTT	0,83%	48,53%	55,83%
KYJ: EDASILÜKKAMINE	0,76%	12,55%	42,42%
IL: ÜLERÕHUTAMINE	0,74%	4,35%	18,84%
DIJ: NÕUSTUMINE	0,69%	3,52%	11,40%
DIJ: EDASILÜKKAMINE	0,69%	27,68%	60,89%
IL: PÕHJENDAMINE	0,68%	7,44%	25,75%
VR: PARANDUSE HINDAMINE	0,65%	5,97%	33,07%
KYJ: EI	0,65%	10,33%	26,77%
PPE: ÜMBERSÕNASTAMINE	0,59%	2,98%	10,00%
DIE: PAKKUMINE	0,58%	8,02%	22,37%
KYJ: INFO PUUDUMINE	0,46%	5,62%	14,84%
PPE: MITTEMÕISTMINE	0,46%	10,49%	31,28%
KYE: ALTERNATIIV	0,43%	4,94%	10,06%
KYJ: MUU	0,43%	2,34%	13,02%
DIE: ETTEPANEK	0,42%	5,79%	21,82%
RIJ: VASTUHÜVASTIJÄTT	0,41%	61,31%	51,85%
KYE: TÄPSUSTAV	0,39%	3,70%	7,84%
RIE: LÕPUSIGNAAL	0,33%	5,39%	28,24%
RIJ: LÕPETAMISE VASTUVÕTMINE	0,31%	2,14%	2,48%
SEJ: NÕUSTUMINE	0,31%	0,00%	0,00%
IL: KOKKUVÕTMINE	0,30%	5,04%	10,34%

Märgend	Sagedus	Täpsus	Saagis
KYJ: NÕUSTUV EI	0,29%	6,26%	23,89%
KYJ: ALTERNATIIV: ÜKS	0,27%	5,26%	1,89%
SEE: ARVAMUS	0,27%	0,93%	1,89%
SEE: VÄIDE	0,26%	1,71%	1,92%
KYE: MUU	0,23%	1,55%	2,25%
YA: JUTU PIIRIDE OSUTAMINE	0,22%	23,31%	44,71%
YA: EELTEADE	0,18%	18,18%	40,00%
VR: HINNANGULINE VASTUVÕTUTEADE	0,17%	2,55%	7,69%
DIJ: INFO PUUDUMINE	0,16%	6,25%	4,69%
DIJ: MUU	0,15%	0,94%	3,39%
IL: SELETAMINE	0,15%	0,78%	1,69%
IL: HINNANG	0,14%	1,22%	1,82%
KKE: ALGATUS	0,13%	1,55%	26,92%
DIE: PALVE OODATA	0,13%	7,50%	23,53%
TVE: PAKKUMINE	0,13%	11,43%	7,84%
RIE: SOOVIMINE	0,12%	3,68%	25,53%
KYJ: KAHTLEV	0,11%	0,00%	0,00%
DIJ: MITTENÕUSTUMINE	0,11%	0,00%	0,00%
KKJ: KINNITAMINE	0,11%	39,39%	30,23%
TVJ: VASTUVÕTMINE	0,11%	3,85%	4,65%
DIJ: PIIRATUD NÕUSTUMINE	0,11%	0,00%	0,00%
PA: ENESEPARANDUS	0,10%	0,00%	0,00%
RIJ: TÄNAN	0,09%	3,87%	16,22%
IL: MUU	0,09%	4,88%	5,88%
RIE: PALUN	0,08%	0,75%	3,03%
VR: HINNANGULINE INFO OSUTAMINE UUEKS	0,08%	1,90%	6,06%
YA: LUBADUS	0,08%	0,00%	0,00%
RY: KONTAKTEERUMINE	0,07%	22,54%	55,17%
DIJ: TEGEVUS	0,07%	0,00%	0,00%
SEJ: PIIRATUD NÕUSTUMINE	0,07%	0,00%	0,00%
RIJ: MUU	0,07%	2,21%	15,38%
SEJ: MUU	0,07%	0,00%	0,00%
SEJ: MITTENÕUSTUMINE	0,06%	0,00%	0,00%
KYJ: KEELDUMINE	0,06%	0,00%	0,00%
IL: JÄRELDAMINE	0,06%	0,00%	0,00%
PA: PARTNERI PARANDUS	0,06%	0,00%	0,00%
YA: RETOORILINE KÜSIMUS	0,06%	0,00%	0,00%
KYJ: MUU KAS-VASTUS	0,05%	0,00%	0,00%
IL: PEHMENDAMINE	0,05%	0,00%	0,00%
KYJ: ALTERNATIIV: MÕLEMAD	0,04%	0,00%	0,00%
DIJ: KEELDUMINE	0,04%	0,00%	0,00%
RIE: VABANDUS	0,04%	0,00%	0,00%
RY: ÜLEANDMINE	0,04%	12,50%	7,14%
KYJ: ALTERNATIIV: KOLMAS VALIK	0,03%	0,00%	0,00%
YA: REFERAAT	0,03%	0,00%	0,00%
DIE: MUU	0,03%	0,00%	0,00%
VR: HINNANGULINE PIIRITLEJA	0,03%	0,00%	0,00%
PPE: MUU	0,02%	0,00%	0,00%
VR: HINNANGULINE JÄTKAJA	0,02%	0,00%	0,00%
RIJ: VASTUSOOVIMINE	0,02%	0,00%	0,00%
KYJ: ALTERNATIIV: EITAV	0,02%	0,00%	0,00%
KYJ: ALTERNATIIV: MUU	0,02%	0,00%	0,00%
YA: JUTUSTAMINE	0,02%	0,00%	0,00%
KYJ: TEGEVUS	0,01%	0,00%	0,00%
RIJ: LÕPETAMISE TAGASILÜKKAMINE	0,01%	0,00%	0,00%
RY: ÄRATUNDMINE	0,01%	0,00%	0,00%
YA: RETOORILINE VASTUS	0,01%	0,00%	0,00%
KYJ: MITTENÕUSTUV JAH	0,01%	0,00%	0,00%
RIE: IDENTIFITSEERIMINE	0,01%	0,00%	0,00%

Märgend	Sagedus	Täpsus	Saagis
RIE: MUU	0,01%	0,00%	0,00%
RIJ: IDENTIFITSEERIMINE	0,01%	0,00%	0,00%
DIJ: ETTEPANEK	0,01%	0,00%	0,00%
DIJ: KAHTLEV	0,01%	0,00%	0,00%
RIE: VABANDUSE VASTUVÕTMINE	0,01%	0,00%	0,00%
SEJ: KEELDUMINE	0,01%	0,00%	0,00%
RIJ: TÄNAMINE	0,01%	0,00%	0,00%
RIJ: VABANDUSE VASTUVÕTMINE	0,01%	0,00%	0,00%
SEE: MUU	0,01%	0,00%	0,00%
VR: JÄTKAJA	0,01%	0,00%	0,00%
DIJ: PAKKUMINE	0,00%	0,00%	0,00%
KKE: MUU	0,00%	0,00%	0,00%
KKJ: MUU	0,00%	0,00%	0,00%
KYE: INFO ANDMINE	0,00%	0,00%	0,00%
KYE: TÄPSUSTAV KÜSIMUS	0,00%	0,00%	0,00%
KYJ: VASTUS ALTERNATIIVINA	0,00%	0,00%	0,00%
KYJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE	0,00%	0,00%	0,00%
PA: MUU	0,00%	0,00%	0,00%
SEE: NÕUSTUMINE	0,00%	0,00%	0,00%
VR: INFO OSUTAMINE UUEKS	0,00%	0,00%	0,00%
VTE: MUU	0,00%	0,00%	0,00%
VTJ: MUU	0,00%	0,00%	0,00%

Lisa 3. DARECi kasutajaliidese hindamise küsimustik

https://docs.google.com/spreadsheet/viewform?formkey=dDY0WENvOEt3d2pqbktQYlNpdktrT1E6MQ

DAREC kasutajaliidese analüüs

Käesolevaga palun Teil leida veidi aega ja hinnata dialoogiaktide poolautomaatse märgendamisvahendi DAREC kasutajaliidest (http://ats.cs.ut.ee/darec/www1/). Küsimustiku eeskujuks on valitud G. Perlmani ankeet (vt. ingliskeelset algvarianti http://hcibib.org/perlman/question.cgi?form=PHUE). Küsimuste korral saatke palun need e-kirjaga aadressile sven.aller@ut.ee.
Lugupidamisega
Sven Aller

Õpitavus
1 - väga halb, 5 - väga hea

	1	2	3	4	5
Abiinfo ja dokumentatsioon (süsteemi intuitiivsus, kasutavus ilma dokumentatsioonita, ülesandele orienteeritud abiinfo)	○	○	○	○	○
Kasutajapärasus (selge keelekasutus, žargooni vältimine)	○	○	○	○	○
Lihtne ja loomulik suhtlus kasutajaga (puuduvad ülearused tegevused, järjekord on loogiline)	○	○	○	○	○
Ekspertkasutaja võimalused (spetsiaalsed võtted, kiirkäsud jms)	○	○	○	○	○

Kommentaare õpitavuse kohta

Kasutajasõbralikkus
1 - väga halb, 5 - väga hea

	1	2	3	4	5
Ülevaade hetkeolukorrast (selgus, kuhu saab liikuda, mida saab teha, kuidas tagasi pöörduda eelmise juurde)	○	○	○	○	○
Selgus, millised operatsioonid on võimalikud ja millised mitte	○	○	○	○	○
Tegevuste ja juhtelementide (ikoonid, nupud) sobivus	○	○	○	○	○
Meelespidamise vajadus (operatsioonid tehtavad ilma varasemaid valikuid mäletamata)	○	○	○	○	○
Süsteemi ühtsus, standardite järgimine (terminite ühtne kasutamine, harjumuspärased kasutamisvõtted)	○	○	○	○	○
Disaini meeldivus (värvilahendus, ülearuse info puudumine jne)	○	○	○	○	○

Kommentaare kasutajasõbralikkuse kohta

Tagasiside ja veakäsitlus

1 - väga halb, 5 - väga hea

	1	2	3	4	5
Tagasiside (hetkeseis selgelt arusaadav)	○	○	○	○	○
Vigade ennetamine (vähe võimalusi juhuslikult eksida, kinnituste küsimine jne)	○	○	○	○	○
Veateated (selge sõnastus, soovitused edasiseks tegevuseks)	○	○	○	○	○
Selged väljumisvõimalused programmist, vigade parandatavus	○	○	○	○	○

Kommentaare tagasiside ja veakäsitluse kohta

Kasutajaliidest puudutavad negatiivsed märkused

Kasutajaliidest puudutavad positiivsed märkused

Submit

Lisa 4. DARECi uuendatud kasutajaliidese abiinfo

DAREC: abiinfo

Sisukord

- Üldinfo
- Kasutajate haldamine
- Failide haldamine
- Dialoogiaktide märgendamine
- Kasutatud ikoonide tähendused

Üldinfo

Veebilehe eesmärgiks on võimaldada dialooge mugavalt märgendada. Dialoogifaile saab üles laadida, neid muuta või kustutada või valida märgendamiseks. Kõigepealt märgendab iga dialoogi automaatne märgendaja, seejärel on kasutajal võimalus kontrollida ja muuta saadud tulemust ning salvestada väljundfail.

Tekkinud küsimustest ja probleemidest, mida käesolev abiinfo ei puuduta, andke palun teada aadressil sven.aller@ut.ee.

Kasutajate haldamine

Kasutajate haldusliidese abil on võimalik pääseda juurde kasutajate saadetud sisend- ja väljundfailidele, kasutajaid lisada ja kustutada. Kasutanimed võivad sisaldada vaid nö. inglise tähestiku tähti, numbreid, side- ja alakriipsu ning punkti, muud sümbolid kas teisendatakse sarnaseks sümboliks (näit. sümbol "õ" sümboliks "o") või asendatakse alakriipsuga. Kui samanimeline kasutaja on juba olemas, katkestatakse operatsioon ja antakse sellest teada.

Kasutaja kustutamisel eemaldatakse automaatselt ka kõik tema kataloogis olnud failid. Enne kustutamist küsitakse kasutajalt kinnitust.

Failide haldamine

Kasutajate failid on jagatud sisend- ja väljundfailideks. Faile saab avada nii vaatamiseks (hiireklõps faili nimel või ikoonil selle järel) kui toimetamiseks (ikoon). Failide nimed on varustatud infoga suuruse ja muutmise kuupäeva kohta, et orienteerumine oleks lihtsam. Ühe või mitme faili märgendamiseks tuleb ära märkida vastavad ruudud nime ees ja vajutada nuppu "Märgenda".

Vastava liidese abil saab üles laadida ka uusi dialoogifaile (tekstifailid, txt-laiendiga, vt.dialooginaide.txt), iga faili tohib olla kuni 500 kB. Failinimed võivad sisaldada vaid nö. inglise tähestiku tähti, numbreid, side- ja alakriipsu ning punkti, muud sümbolid kas teisendatakse sarnaseks sümboliks (näit. sümbol "õ" sümboliks "o") või asendatakse alakriipsuga. Kui samanimeline fail on juba olemas, katkestatakse operatsioon ja antakse sellest teada.

Failide kustutamiseks tuleb vastavas alajaotuses märkida ära vastavate failide ees olevad ruudud ja vajutada nupule "Kustuta". Enne kustutamist küsitakse kasutajalt kinnitust. Väljundfaile (laiendiga ***.darec***) saab nii vaadata kui kustutada sarnaselt sisendfailidele.

Dialoogiaktide märgendamine

Dialoogifaili märgendamisele saatmisel tehakse kõigepealt automaatne dialoogiaktideks jagamine ja märgendamine. Kasutajal on soovitav jagamine üle kontrollida, teha vajalikud parandused ja korrata automaatset märgendamist.

Märgendamiseks valitud dialoogid asuvad eraldi hallides raamides, mida saab sulgeda ja avada hiireklõpsuga selle päises oleval failinimel. Algselt on kõik raamid avatud. Raami sees on iga dialoog esitatud ridadena lausungite kaupa, igas reas on näha lausungi tekst koos võimalike operatsioonidega. Kontekstitundliku abiinfo saab sisse lülitada märkeruudu "Näita nõuandeid" abil, sel juhul tekib objektide juurde sildike objekti kohta käiva infoga.

Hiireklõpsuga lausugi tekstil saab seda jagada kaheks, lausungi liitmiseks eelmisega tuleb kasutada ikooni .

Tulemuse eelvaate tulbas on näha lausungile omistatud märgendid e. märgendid, mis on valitud tõenäosuslike ja kõikide märgendite tulpades. Nende muutmiseks tuleb leida uus märgend tõenäosuslike või kõigi märgendite tulbast, kustutamiseks aga valida ikoon . Kõigi märgendite hulgast õige kiiremaks leidmiseks võib kasutada klaviatuuri, trükkides soovitud märgendi esitähte seni, kuni soovitud märgend on valitud.

Ikooni abil saab automaatset märgendamist korrata.

Ikoon näitab, et antud rida on taasmärgendamiseks avatud, ikoon aga seda, et nende lausungite puhul valitud märgendeid taasmärgendamisel ei muudeta. Lukustust saab muuta hiireklõpsuga vastaval ikoonil.

Rea lukustamisel ja automaatse märgendamise kordamisel lisatakse ka kasutaja poolt valitud märgendid tõenäosuslike märgendite hulka. Neist mitme märgenduse lisamiseks võib valida tõenäosuslike märgendite hulgast mitu varianti, kasutades valiku tegemiseks või tühistamiseks Ctrl-klahvi. Valitud märgendit saab valikus ülespoole tõsta topeltklõpsuga.

Pärast jagamist, ühendamist ja kordusmärgendamist uuendatakse veebilehte, viimati vaadatud reale saab tagasi liikuda ikooni abil.

Nupu "Salvesta tulemus tööpinki" abil salvestatakse kõik hetkel avatud dialoogid raamistikku, salvestamisel lisatakse failinime lõppu laiend ***.darec***. Salvestatud faile saab näha kasutaja väljundfailide hulgas.

Kasutatud ikoonide tähendused

- kasutajate haldamine

- failide haldamine

- abiinfo

- liikumine viimati vaadatud reale

- faili vaatamine

- faili redigeerimine

 - kontekstitundlik abiinfo

 - lausungi liitmine eelmise lausungiga

 - lausung on automaatseks märgendamiseks avatud

 - lausung on automaatseks märgendamiseks suletud, analüüs on fikseeritud

 - automaatse märgendamise kordamine

 - valitud märgendite kustutamine