TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika eriala

Sven Aller

Loomulikus eesti keeles suhtlus avalike veebiteenustega

Bakalaureusetöö (6 EAP)

Juhendaja: Margus Treumuth

Autor: ..... "....." juuni 2010
Juhendaja: ..... "....." juuni 2010
Lubada kaitsmisele
Professor Mare Koit ..... "....." juuni 2010

TARTU 2010

Sisukord

Sissejuhatus

Viimaste aastate jooksul on arvuti ja võrguühendus saavutanud koha enamikus Eestimaa kodudest: kui 2005. aastal oli veel veidi üle 200 000 leibkonna koduse võrguühendusega, siis 2009. aastaks oli nende arv kerkinud juba üle 350 000 (*Statistikaamet*, 2010). Seoses sellega on tavaliseks saanud, et pöördume info saamiseks just veebiressursside poole. Reeglina tähendab see mitmesuguste otsimismootorite kasutamist, mis ühest küljest nõuavad arvutipärast lähenemist (sisestame otsimismootori vastavasse lahtrisse sobivaid märksõnu, mida arvame otsitava infoga veebilehel olevat), teisest küljest saame ka tulemuse mitte arusaadava kokkuvõttena, vaid potentsiaalselt sobivate veebilehtede loeteluna, millest enamik tõenäoliselt ei anna meile vajalikku vastust ning eesmärgini jõudmine võtab liialt palju jõudu ja aega. Lisaks sellele: kuigi mingit osa vajalikust teabest leiame oma emakeeles, on hulgaliselt rohkem meile vajalikku hoopis võõr-, enamasti inglise keeles. Oma senistele kogemustele tuginedes võtame paratamatuna, et saame Internetist teavet võõrkeeles, kuid oluliselt lihtsam oleks, kui meieni jõuaks sisestatud konkreetsele küsimusele korralikult sõnastatud emakeelne vastus.

Süsteeme, mis võimaldavad kasutajaga loomulikus keeles suhelda, on loodud juba alates 1960-ndatest aastatest. Enamasti on sellised süsteemid autonoomsed (vastused koostatakse süsteemi enda andmete põhjal) ning piiritletud ühe konkreetse teema ja andmebaasiga (näiteks rongipiletite müük vms). Kindlasti on autonoomse süsteemi stabiilsus paremini tagatud ning tema hooldus lihtsam: omanik kontrollib ise ka andmebaasi ning selle struktuuri muutuste korral on võimalik samaaegselt muuta ka dialoogsüsteemi ülesehitust. Samas tähendab see ka andmehaldust (kontroll ja täiendamine) ning vastava valdkonna kompetentsi, mis teeb kogu süsteemi ülalhoidmise ebaefektiivseks ja sageli dubleerivaks juhul, kui vastav andmebaas on kusagil juba loodud. Teisest küljest piirab ühe teema keskne dialoogsüsteem kasutajat: saadav teave on ühekülgne ja kasutaja fraaside lahtimõtestamisel ei suuda süsteem olla väga paindlik. Samal ajal leidub hulgaliselt andmebaase avalike veebiteenuste (*web service*) näol meid huvitava ning sageli tasuta jagatava infoga.

Käesoleva bakalaureusetöö eesmärgiks on luua dialoogsüsteem, mis lubaks kasutajal suhelda erinevate veebiteenustega loomulikus eesti keeles. Süsteem kasutab info otsimisel kuut veebiteenust, kuid teiste analoogsete teenuste lisamine on tänu programmi universaalsele ülesehitusele lihtne. Lisaks neile on kasutusel ka mitmesugused abivahendid:

firma Filosoft eesti keele lemmatiseerija, tõlketeenus *Google Translate,* kaarditeenus *Google Maps* ja *IPInfoDB* kasutaja asukoha tuvastamiseks IP-numbri järgi.

Töö on jagatud kolmeks peatükiks. Esimene neist kirjeldab loomuliku keele ja arvuti seoseid, sh. kirjeldab ka dialoogsüsteeme. Teine osa võtab vaatluse alla veebiteenuste leviku ja kasutamise. Töö kolmas peatükk keskendub käesoleva bakalaureusetöö raames valminud dialoogsüsteemi ülesehituse ja funktsioonide kirjeldamisele, testimistulemustele ja edasiarendusvõimalustele. Lisadena on toodud süsteemiga peetud dialoogide näidised, süsteemisisene sõnastik, suhtlusfraaside nimekiri ja programmi kood CD-plaadil.

Dialoogsüsteem on kirjutatud programmeerimiskeeles PHP, kasutajalt nõutakse vaid graafilist veebibrauserit. Programmi testversioon asub aadressil http://dab.hostei.com/.

1. Loomulik keel ja arvuti

Loomuliku keele töötlus arvuti abil on vaatluse all olnud pikka aega. Kui elektronarvutite ajalugu ulatub 1940-ndate aastate keskpaigani, siis esimesed katsetused keele automaattöötluse osas toimusid küllaltki varsti pärast seda, kui 1954. aastal viidi New Yorgis läbi esimene masintõlkega seotud eksperiment. Algselt suhteliselt lihtsaks peetud keelega seotud ülesanded (masintõlge, dialoogsüsteemid jms) osutusid arvatust aga oluliselt raskemaks ning viimase rohkem kui poole sajandi jooksul pole jõutud soovitud eesmärkideni, näiteks kvaliteetne täisautomaatne tõlge, suulise kõne sünkroontõlge, loomulikus keeles suhtlus andmebaasidega.

Samas on keele automaattöötluses saavutatud olulisi tulemusi. Valminud on hulk masintõlkesüsteeme, mis suudavad anda sageli küllalt häid tulemusi, seda küll reeglina tehniliste tekstide puhul ja piiratud valdkonnas (Hutchins, 2005). Loodud on dialoogsüsteeme, mis oma piiratud valdkonnas saavad asendada inimest.

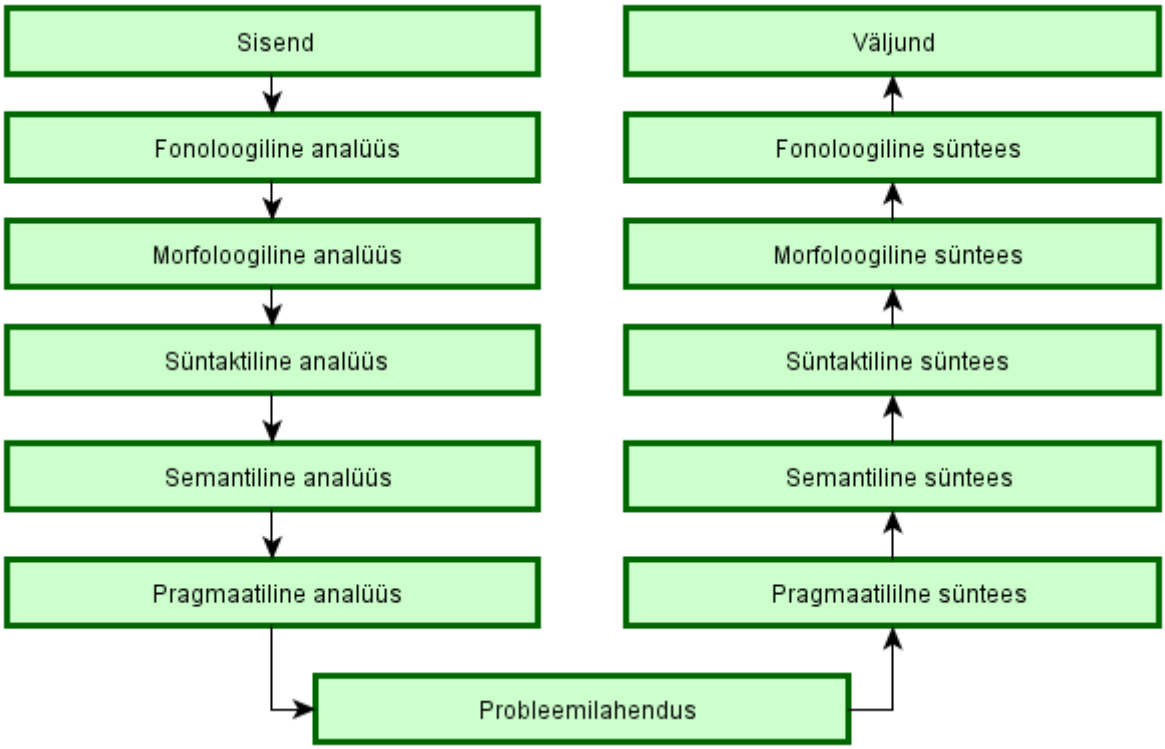
Lisaks täissüsteemidele on valminud hulk keeletöötluseks vajalikke abivahendeid, mida saavad kasutada tulevikus loodavad süsteemid: elektroonilised sõnastikud, tesaurused, paralleelkorpused, aga ka kõnetuvastus ja –süntees.

Eestiski on loodud palju erinevaid keeletöötlusvahendeid. Riikliku programmi Eesti keele keeletehnoloogiline tugi 2006-2010 raames on arendatud erinevaid projekte ning selle tulemusel on valminud hulk keeletöötlusvahendeid (*Eesti keele...*, 2006). Teemad on olnud seotud masintõlke, keelekorpuste, dialoogikorpuste, dialoogsüsteemide, foneetilise, morfoloogilise, süntaktilise ja semantilise analüüsiga, tesauruse, kõnetuvastuse ja –sünteesiga jne. Valminud on muuhulgas eesti keele õigekirjakontroll, poolitaja, tesaurus, lemmatiseerija, morfoloogiline analüsaator, süntesaator, keelekorpused, sõnastikud, dialoogsüsteeme jne. Ka Eesti Haridus- ja Teadusministeeriumi avaldatud dokumendis Eesti keele arendamise strateegia 2004-2010 (*Eesti keele...*, 2004) on muuhulgas ära toodud eesti keele tehnoloogiliste vahendite arendamise tähtsus ja ülesanded.

1.1. Dialoogsüsteemid ja küsimus-vastussüsteemid

Dialoogsüsteemide (*dialog system*) all mõistetakse süsteeme, mis suudavad inimesega suhelda, kasutades näiteks teksti, kõnet, graafikat, kompimist, žeste, miimikat jms. nii

sisendi kui väljundina. Ideaaljuhul võib dialoogsüsteem koosneda sisendi analüsaatorist (morfoloogilisel, süntaktilisel, semantilisel tasemel, vajadusel ka kõnetuvastus), sellele järgneb vajaliku info otsimine ja vastuse genereerimine (vt. Joonis 1), kuid sageli jäetakse vajaduse puudumisel ja süsteemi lihtsustamise eesmärgil mõned etapid ära.



Joonis 1. *Dialoogsüsteemi skeem*

Esimene tuntuim dialoogsüsteem nimega Eliza valmis juba 1966. aastal ning põhines võtmesõnade äratundmisel. Edasised arengud viisid selliste süsteemide täiustumiseni, 1991. aastast korraldatavad võistlused Loebneri auhinnale kasutavad nn. Turingi testi ja püüavad motiveerida autoreid looma võimalikult inimkaaslasele sarnast dialoogsüsteemi (Jurafsky & Martin, 2009).

Ka eesti keeles on valminud dialoogsüsteeme. Eliza analoogist Liisbet on jõutud infot pakkuvate programmideni nagu Reisiagent (Treumuth, 2002), Alfred (Alfred, 2010) ja hambaravi kohta nõu andev Zelda (Zelda, 2010).

Küsimus-vastussüsteemid (*question answering*) on üheks dialoogsüsteemide alaliigiks. Need lubavad kasutajal sisestada loomulikus keeles küsimusi ja püüavad neile samal viisil ka vastata. Andmetena kasutatakse kas eelnevalt vastavalt vajadusele struktureeritud andmebaasi, loomulikus keeles veebis olevaid või lokaalseid tekstikorpusi. Teemade valik

võib olla piiratud, kuid ei pruugi. Sageli on selliseid süsteeme nimetatud ka otsimismootorite järgmiseks põlvkonnaks.

2. Veebiteenused

Enamik veebis olevast infost on kättesaadav ainult läbi infot omavate organisatsioonide veebilehtede. Samas on sarnast teavet vaja ka teistel organisatsioonidel, kasutades seda siis eraldi või kombineerides teiste andmetega. Nii inim- kui tehnilise ressursi kokkuhoiu mõttes on mõistlik kasutada ressursse ühiselt. Vähenevad kulud vajalikele andmebaasidele (näiteks riikide infot hoitakse ja uuendatakse vaid ühes kohas) ning seadmetele (näiteks ilmaennustusega seotud andmeid kogutakse ainult ühes kohas). Sel põhimõttel pakuvad infot veebiteenused (*web service*).

Veebiteenused on serverite poolt pakutavad ressursid, mis on kasutatavad kaugarvutite poolt HTTP (*Hypertext Transfer Protocol*) abil. Sageli on veebiteenused tasuta, väikese tasu eest võidakse pakkuda suuremat hulka päringuid, paremat töökindlust või struktuurimuutustest teavitamist. Kasutaja küsimus antakse serverile edasi HTTP-päringuna näiteks aadressile atribuutide lisamise ja väärtustamise abil. Server annab vajaliku päringu edasi andmebaasile ja saadud vastus edastatakse kasutajale, seda sageli XML (*Extensible Markup Language*) kujul. Selline lahendus pole serverile väga koormav (infohulk, mida vahendatakse, on väike), samuti säilitatakse andmebaasi turvalisus: päringuvõimalused on serveri poolt rangelt piiratud. Ka väljund on standardiseeritud, tagastades iga teenuse puhul sarnase struktuuriga faili.

XML on hulk reegleid semantiliste märgendite defineerimiseks, mille abil jagatakse dokument osadeks ja märgendatakse need (Harold, 2004). Märgendid pole eelnevalt defineeritud, kasutaja võib need ise vastavalt vajadusele määrata. Tänu oma lihtsusele, paindlikkusele ja vähenõudlikkusele on formaadi XML populaarsuse pidevalt kasvanud. Seepärast võib arvata, et oluliselt kasvab ka XML-formaadis veebiteenuste hulk, mida tulevikus saab kasutada.

Veebiteenuste tarvitamisel ei puudu ka negatiivsed küljed. Erinevalt terviklikust lokaalsest süsteemist, kuhu on peale vahendusmooduli integreeritud ka kogu kasutatav andmebaas, ei saa sel puhul garanteerida info kättesaadavust: infot vahendav süsteem võib töötada, kuid infot sisaldav süsteem mitte, samuti võib olukord olla vastupidine. Samuti ei oma näiteks dialoogsüsteemi arendaja kontrolli pakutava teenuse struktuuri üle: teenusepakkuja võib

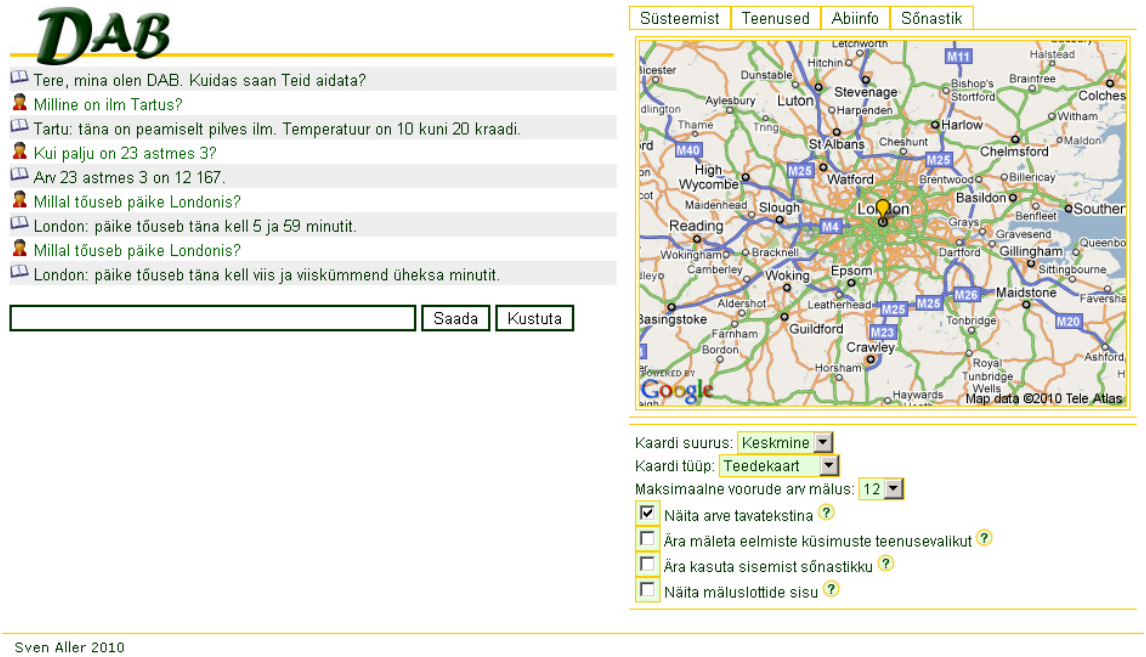
muuta suvalisel ajal enda poolt pakutavate andmete struktuuri, see omakorda võib viia rivist välja teenust kasutanud dialoogsüsteemi.

Sarnased riskid on ka alternatiivsetel võimalustel: ka muude veebis veebilehe või dokumendina jagatavate materjalide puhul tekivad analoogsed probleemid. Lisaks sellele on veebilehtede puhul vajaliku info kättesaamine oluliselt töömahukam (dialoogsüsteem peab muu teksti seest leidma üles õige fraasi), muud tüüpi dokumentide puhul lisanduvad failiformaati puudutavad probleemid. Mõlemal juhul ei saaks küllalt tõenäoliselt ehitada ka universaalset süsteemi e. süsteemi, millele uute teenuste lisamine oleks lihtne ja mugav: lehekülje struktuur ja info paiknemine sellel võib olla äärmiselt erinev.

3. Programmi kirjeldus

Programmi eesmärgiks on pakkuda võimalust suhelda mitme avaliku veebiteenusega loomulikus eesti keeles. Kõik süsteemi poolt kasutatavad veebiteenused on tasuta ja annavad väljundina XML-formaadis faili, tänu standardsusele need on hästi integreeritavad, samuti on uute veebiteenuste lisamine lihtsam.

Suhtlus toimub dialoogina: pärast tervitust antakse kasutajale võimalus esitada küsimus, programm püüab sellele vastata, ootab kasutajalt uut küsimust jne. Dialoogsüsteem hoiab meeles vastavalt kasutaja valikule 6, 12 või 18 küsimust või vastust, et lubada paremat ekraanikasutust ning vajadusel vähendada ülearuse info säilitamist ja transporti üle arvutivõrgu. Kuna enamik käsitletavaid teemasid on seotud geograafilise asukohaga, siis on lisatud ka geograafiline kaart. Arvulisi andmeid on võimalik lasta arvutil näidata ka sõnalisel kujul, samuti saab keelata või lubada eelmistest voorudest tuvastatud teenuste kasutamist uuele küsimusele vastamisel, süsteemisisese sõnastiku tarvitamist ning mäluslottide sisu esitamist koos kasutatava teenuse aadressiga. Joonisel 2 on esitatud programmi ekraanipilt.



Joonis 2. *Programmi DAB ekraanipilt*

Süsteem pakub kasutajale infot kuuelt teenuselt:

- *GeoNames Country Info* – riikide kohta käiv info;

- *GeoNames Wikipedia Fulltext Search* – Wikipedia artikli algus linna kohta;
- *GeoNames Timezone* – kohalik aeg.
- *Google Maps Geocoding* – linnade koordinaadid ja asukohariik;
- *Weather Underground* – ilmateade;
- *Calculator* – kalkulaator.

Lisaks ülaltoodud kuuele teemale kasutab programm veel teisi teenuseid kasutaja tuvastamisel, sisenditöötlusel ja kaardi kuvamisel:

- *Eesti keele lemmatiseerija* – sisendsõnade algvormide leidmiseks;
- *Google Translate* – inglise ja eesti keele vaheliseks tõlkimiseks;
- *Google Maps* – kaardi kuvamiseks;
- *IPInfoDB* – IP-numbri põhjal kasutaja asukoha määramiseks.

3.1. Infoteenused

Kõik valitud teenused on oma kasutusviisilt sarnased: nad saavad päringuks vajalikud argumendid URLi osana ja annavad vastuse XML kujul. Argumentideks on enamasti sisendfraasist leitud või selle alusel tuletatud linnanimed, kahetähelised riigitähised või arvud.

3.1.1. GeoNames Country Info

Annab etteantud riigitähise alusel infot küsitava riigi kohta.

- aadress: http://ws.geonames.org/countryInfo
- argumendid: riigitähis (näit. *country=EE*)
- päringu vastusest kasutab süsteem riigi nime, pealinna, pindala, rahvaarvu ja rahaühikut
- märksõnad: pealinn, pindala, suurus, rahvaarv, rahvastik, elanik, inimene, valuuta, rahaühik

3.1.2. GeoNames Wikipedia Fulltext Search

Annab etteantud linna kohta käiva Wikipedia artikli alguse, asukohariigi tähise, elanike arvu, koordinaadid jne.

- aadress: http://ws.geonames.org/wikipediaSearch

- argumendid: linn (näit. *q=tartu*)
- päringu vastusest kasutab süsteem Wikipedia artikli algust linna kohta ja vastava lehekülje aadressi
- märksõnad: wikipedia, artikkel, lühiinfo, info, teave, teadma

3.1.3. GeoNames Timezone

Annab etteantud linnas momendil kehtiva kuupäeva ja kellaaja, ajanihke ja ajatsooni identifikaatori.

- aadress: http://ws.geonames.org/timezone
- argumendid: linna koordinaadid (näit. *lat=58&lng=27*)
- päringu vastusest kasutab süsteem kuupäeva ja kellaaega ühendavat sõnet
- märksõnad: kuupäev, kell, kellaaeg

3.1.4. Google Maps Geocoding

Annab etteantud linna koordinaadid ning asukohariigi nime ja tähise.

- aadress: http://ws.geonames.org/timezone
- argumendid: linna nimi (näit. *address=Tartu*)
- päringu vastusest kasutab süsteem linna koordinaate ja asukohariigi nime
- märksõnad: kus, asuma, nimetus, laiuskraad, pikkuskraad, koordinaat

3.1.5. Weather Underground

Annab etteantud linna ilmateate, ajatsooni, maksimum- ja miinimumtemperatuuri, päikese-tõusu ja –loojangu kellaaja.

- aadress: http://api.wunderground.com/auto/wui/geo/ForecastXML/
- argumendid: linna nimi (näit. *query=tartu*)
- päringu vastusest kasutab süsteem ilmateadet, maksimum- ja miinimum-temperatuuri, päikesetõusu ja –loojangu kellaaega.
- märksõnad: ilmateade, ilm, õhutemperatuur, temperatuur, soe, külm, päike, pilv, vihm, lumi, lörts, äike, loojuma, päikeseloojang, loojang, tõusma, päikesetõus, tõus

3.1.6. Calculator

Annab etteantud arvude põhjal arvutustehete tulemused: liitmine, lahutamine, korrutamine, jagamine, astendamine, juurimine ja absoluutväärtuse leidmine.

- aadress: http://wswebservice.awardspace.info/calculator.php
- argumendid: kaks arvu (näit. *value1=2&value2=2*)
- päringu vastusest kasutab süsteem tehete vastuseid
- märksõnad: arvutamine, liitma, liitmine, pluss, summa, lahutama, lahutamine, miinus, vahe, korrutama, korrutamine, kordama, korrutis, jagama, jagamine, jagatis, aste, astendama, juur, juurima, ruutjuur, kuupjuur, absoluutväärtus

3.2. Abiteenused

Lisaks nendele teenustele, mis annavad otseselt kasutajat huvitavat infot, kasutab süsteem ka nelja nö. abiteenust: eesti keele lemmatiseerija, *Google Translate*, *Google Maps* ja *IPInfoDB*.

3.2.1. Eesti keele lemmatiseerija (http://www.filosoft.ee/lemma_et/)

Firma Filosoft poolt loodud eesti keele lemmatiseerija abil leitakse sisendist saadud sõnade algvormid. Ilma selleta oleks teenuseid ja fraase esilekutsuvate märksõnade kirjeldamine keerulisem, samuti oleks võimatu tavalause kujul olevast sisendist info (linnade, riikide ja arvude) eristamine. Lemmatiseerija tagastab päringule tavalisel kujul veebilehe, seepärast peab süsteem töötlema tulemust nii, et vajalik info eristataks muust tekstist.

3.2.2. Google Translate (http://translate.google.com/)

Enamik veebiteenustest tagastab ingliskeelset infot, seega on vaja seda enne eestikeelse lause moodustamist tõlkida. Antud dialoogsüsteem kasutab selleks teenust *Google Translate*, mis on üks suuremaid tasuta tõlketeenusepakkujaid. Sarnaselt teistele tõlke-süsteemidele peab arvestama siingi, et tulemus on sageli ebakorrektne. Eriti on seda näha pikemate lõikude puhul ning see on paratamatu, näiteks ühes teenuses pakutavate Wikipedia lühiartiklite tõlkimine on ebakvaliteetne ja seepärast kasutatav ainult tekstist ülevaate saamiseks. Lühemate puhul on süsteemis kasutusele võetud väike sisemine sõnastik, kuhu on lisatud probleemsed kohanimed ja ilmateates kasutatavad väljendid. Viimaste puhul aitas nimekirja koostamisel teenusepakkuja poolt pakutud nimekiri võimalikest fraasidest (vt. Lisa 2).

3.2.3. Google Maps (http://maps.google.com/)

Aitamaks kasutajal mugavamalt dialoogsüsteemist aru saada, on programmile lisatud ka geograafiline kaart teenuse *Google Maps* abiga. Kaardil on markeriga tähistatud kõnealune linn, kaardi esitamise päringus antakse edasi linna koordinaadid. Kasutatud on staatilist kaarti, nii koormab teenuse kasutamine vähem serverit: kasutaja ei saa kaarti dünaamiliselt skaleerida või tüüpi valida. Samas on dialoogsüsteemi rippmenüüde abil võimalik valitud pildi suurust ja tüüpi muuta, valida on viie erineva suuruse ja sarnaselt dünaamilise kaardiga nelja erineva tüübi vahel.

3.2.4. IPInfoDB (http://www.ipinfodb.com/)

Antud teenust vajatakse kasutaja asukoha kohta käiva info saamiseks: IP-numbri alusel tuvastatakse asukoha riigitähis, riigi nimi, linn ja selle koordinaadid. Juhul, kui IP-aadressi põhjal saadud info pole korrektne (süsteem asub samas lokaalses võrgus), kasutatakse serveri IP-aadressi.

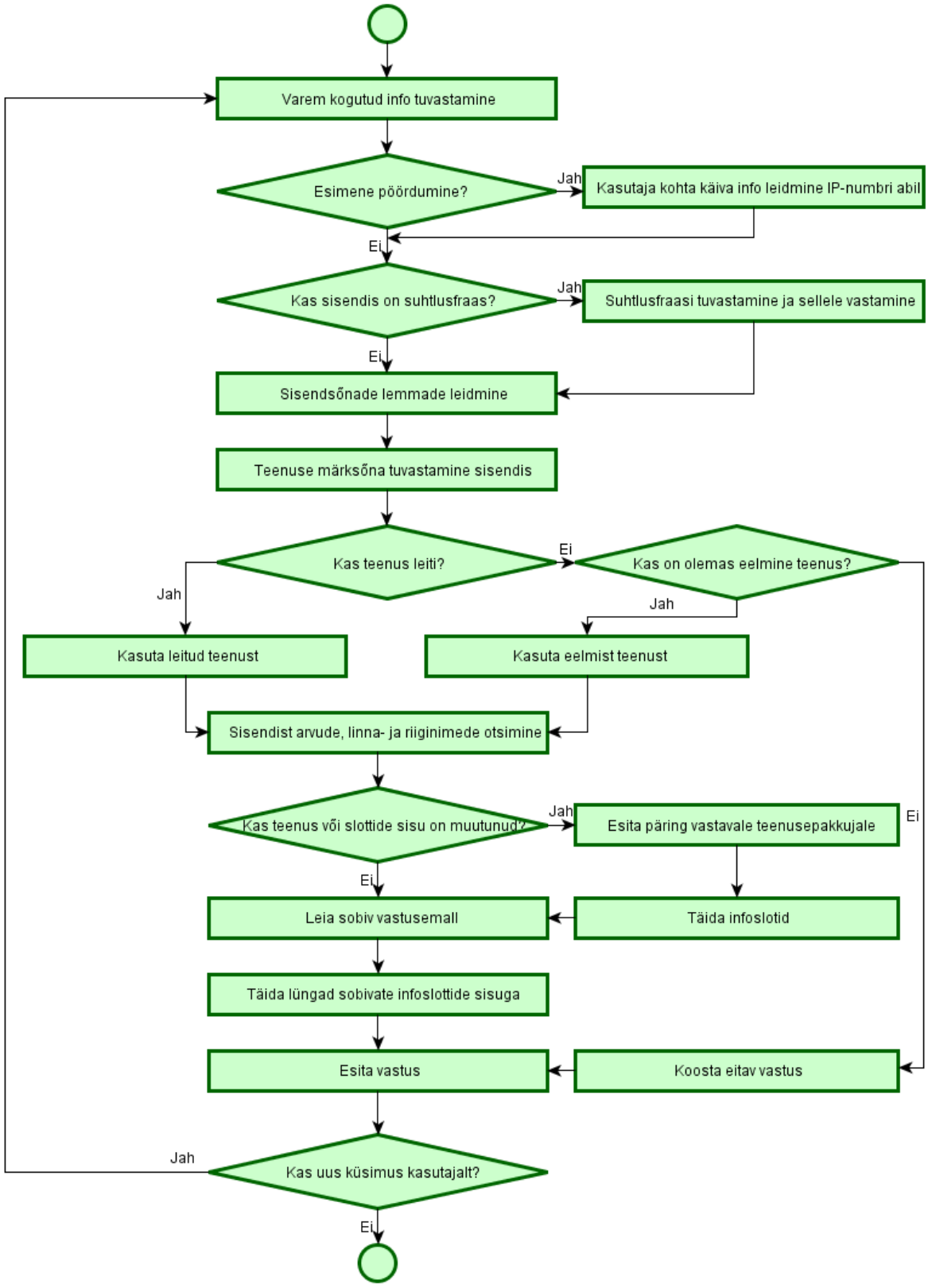
3.3. Ülesehitus

Programmi sisenemisel tervitatakse kasutajat, kirjeldatakse süsteemi ja pakutakse võimalust tutvuda teenuste loetelu ja abiinfoga. Pärast esimest kasutaja sisendit otsitakse teenuse *IPInfoDB* abil infot kasutaja kohta arvuti IP-numbri alusel, eesmärgiks ennetada olukorda, kus kasutaja võib esitada küsimusi oma linna või riigi kohta, jättes koha aga täpsustamata. Saadud infot kasutatakse hilisemate päringute puhul, kui kasutaja sisendist ei leita kohta täpsustavat infot. Vaikeväärtustega täidetakse ka ülejäänud mäluslotid, mida kasutatakse päringute tegemisel argumentide väärtustamiseks.

Süsteemi reaktsioon vastavalt kasutaja sisendile koosneb järgmistest etappidest:

- sisendi töötlemine;
- päringu esitamine;
- päringuvastuse töötlemine ja info salvestamine slottidesse;
- väljundfraasi valik ja lünkade täitmine;
- vastuse koostamine ja esitamine.

Mälus hoitakse eelnenud vestlust (vastavalt kasutaja valikule 6, 12 või 18 vooru), viimase teenuse nimetust ja kasutajainfot. Samuti hoitakse alles kord juba täidetud infoslotte koos viimase infoga. Joonis 3 kirjeldab loodud programmi algoritmi.



Joonis 3. *Programmi algoritm*

3.3.1. Sisendi töötlemine

Sisenditöötlus koosneb viiest etapist. Esmalt otsitakse sisendist tervitus-, tänamis- ja hüvastijätufraase, nende leidmisel lisatakse tulevase vastuse algusesse või (hüvastijätu-fraasi puhul) lõppu sobivate hulgast juhuslikult valitud vastusefraas (vt. Lisa 3). Seejärel leitakse lemmatiseerija abil iga sisendsõna sõna algvorm või –vormid. Selline lähenemine vähendab ühest küljest teenuste või fraaside valiku jaoks vajalike märksõnade nimekirja, teisest küljest aga tagab, et päringu argumentidena esitatud sõnad oleksid kindlasti algvormis.

Kolmanda sammuna otsitakse teenuste hulgast esimene, mille märksõnadest mõni kattub mõne sisendsõna algvormiga. Kui sellist ei leita, kasutatakse eelmises voorus tuvastatud teenust või antakse teada, et küsimusest ei saadud aru. Eelmise teenuse tarvitamise saab kasutaja ka ära keelata: ühelt poolt on mugav, kui ei pea trükkima täislauseid (võimalikud on küsimused stiilis "Milline on ilm Tartus?", "Aga Tallinnas?"), nii on olemas sidusus eelmiste voorudega. Teiselt poolt aga võib teenuse kopeerimine eelmisest küsimusest anda ebakvaliteetsemaid vastuseid: süsteem ei tunnista, et ei saanud küsimusest aru, vaid reageerib eelmiste andmete põhjal. Kuna süsteem kasutab erinevaid teenuseid, peab vastavate märksõnade valik olema läbimõeldud: need ei tohi ühest küljest kattuda, teisest küljest aga peavad nad piisavalt hästi katma kõikvõimalikke teenusega seotud küsimuste esitamise variante. Kui märksõnad on valitud õnnestunult, suudab süsteem üheselt otsustada valitava teenuse üle, saades samas aru võimalikult erineval viisil esitatud küsimustest.

Neljandaks püütakse sisendlemmadest otsida teenuste kasutamiseks vajalikke sõnu: arvud (ainult numbrilisel kujul, nii positiivsed kui negatiivsed, täis- ja murdarvud), riigid ja linnad. Kahe viimase puhul eeldatakse liigse koormuse vältimiseks, et kohanimi algab suure tähega (sel juhul testitakse vaid lemmatiseerijast tulnud suure algustähega sõnu), kontrollimiseks kasutatakse teenust *GeoNames Country Info*. Juhul, kui leiti riik, kuid mitte linna, salvestatakse linnana selle riigi pealinna nimi. Leitud tekstiline info tõlgitakse sisemise sõnastiku või selles vastava fraasi puudumisel teenuse *Google Translate* abil inglise keelde.

Viienda sammuna salvestatakse sisendist leitud info vastavatesse mäluslottidesse.

3.3.2. Päringu esitamine

Pärast sisendi analüüsimist koostatakse päring, mille aluseks võetakse algselt leitud teenuse aadress ning millele lisatakse vajalikud argumendid koos sobivatest mäluslottidest leitud väärtustega. Juhul, kui pärast sisendi töötlemist leiti, et teenus pole eelmise päringuga võrreldes muutunud ning ka mäluslotid on säilitanud esialgse sisu, päringut enam ei teostata, vaid kasutatakse eelmisi andmeid.

3.3.3. Päringuvastuse töötlemine ja info salvestamine slottidesse

Pärast vastuse õnnestunud saamist leitakse saadud XML-failist väärtused vastava teenuse slottide täiteks. Kasutatud on vahendeid *SimpleXML* ning *xpath*, slottide massiivi elementide võtmed ühtivad vastavate *xpath*-väärtustega. Juhul, kui mõni väärtus puudub, jääb ka vastav slott tühjaks.

3.3.4. Väljundfraasi valik ja lünkade täitmine

Väljundfraaside hulgast valitakse vastavalt märksõnade ja sisendlemmade kattumisele esimene sobiv fraas. Kui vastavat fraasi ei leita, antakse vaikimisi fraas, mis üldjuhul hõlmab enamikku leitud infot (näiteks riigiinfo puhul nii riigi pealinna, pindala, rahvaarvu kui rahaühikut). Vajalikud lüngad täidetakse slottidest saadud teabega, tõlkides need vajadusel enne sisemise sõnastiku või teenuse *Google Translate* abil eesti keelde. Väljundfraasid on konstrueeritud nii, et lünkades olevad sõnad on alati nimetavas käändes või (erijuhtudel, näiteks ilmateate puhul) on need juba sisemises sõnastikus õiges vormis. Arvulist infot on vastavalt kasutaja valikule võimalik esitada ka tavatekstina, sel juhul teisendatakse kõik arvud sõnalisele kujule. Kui kõik vajalikud slotid on täidetud, lisatakse tulevasse vastusesse täidetud lünkadega fraas, vastasel juhul teatatakse info mitteleidmisest.

3.3.5. Vastuse koostamine ja esitamine

Eelnenud protsessi käigus tekkis mingi hulk vastusefraase: suhtlusfraasid nagu tervitus, tänamine ja hüvastijätt, samuti sobiva fraasimalli lünkade täitmise tulemus või eitav vastus. Vastavalt nendele koostatakse vastus, arvestades, et tervitusele ja tänamisele vastamine oleks vastuse alguses ning hüvastijätt lõpus, päringutulemusi sisaldav osa aga nende vahel. Näiteks fraasile "Tere! Tartu ilm täna. Nägemist." vastab süsteem: "Tervitused ka minu poolt! Tartu: täna on vihmavõimalusega ilm. Temperatuur on 8 kuni 21 kraadi. Nägemiseni, oli tore suhelda!".

3.3.6. Lokaalsed ressursid

Valminud programm püüab infoallikatena kasutada võimalusel ainult väliseid teenuseid. Samas on kiiruse ja kvaliteedi tõstmiseks lisatud süsteemile riikide loetelu (riigi ingliskeelne nimetus ja lühend), samuti väike sõnastik (vt. Lisa 2). Vajaduse viimase järele põhjustas teenuse *Google Translate* ebapiisava kvaliteediga tõlge, mis tekitas probleeme programmi töös. Sõnastik koosneb teenuse *Google Translate* jaoks probleemsetest riigi-nimedest (*Google Translate*: Norra – *Norway*, kuid Taani – *Danish*) ja ilmateatega seotud fraasidest (*Google Translate*: *Chance of Snow* – Lumesaju). Sõnastiku tarvitamise saab kasutaja vastavalt soovile ka välja lülitada.

3.3.7. Kaardi esitamine

Vastavalt mäluslottidele koostatakse päring ka kaarditeenusele *Google Maps*, mille tulemusel näidatakse kõnealuse linna kaarti. Kaardi detailsuse ja tüübi saab valida kasutaja.

3.4. Tähtsamate funktsioonide kirjeldused

3.4.1. Funktsioon annaKasutajainfo

Antud funktsiooni kasutatakse pärast esimest kasutaja poolt sisestatud lausungit, tuvastamaks kasutaja asukohta tema arvuti IP-numbri järgi. Nii täidetakse paljud asukoha kohta käivad argumendislotid ja sel viisil kohaldab süsteem end kasutaja järgi. Tänu sellele suudab süsteem vastata kasutaja küsimustele ka konkreetset kohta mainimata: kasutaja ei pea oma esimeses küsimuses teada andma, et kellaaega küsides mõtleb ta just oma ajavööndit või ilmateate vastu huvi tundes täpsustama, et soovib seda saada just näiteks Tartu linna kohta.

Funktsioonis kasutatakse veebiteenust *IPInfoDB*, mis kasutab töökindluse tõstmiseks ka tagavaraserverit (*backup*). Saadud XML-kujul vastusest leitakse momendil oluline info: riigitähis, riigi nimi, regioon, linn ja asukoha koordinaadid. Kui sisendist teistsugust infot ei leita, arvestatakse asukohaga seotud päringute puhul nende andmetega, samuti kasu-tatakse laius- ja pikkuskraadi geograafilise kaardi näitamisel.

3.4.2. Funktsioon annaLemma

Funktsioon saab sisendina sõna ja (kasutades firma Filosoft lemmatiseerijat) tagastab sõna kõikvõimalike algvormide massiivi. Sõltuvalt sõnast on lemmade arv erinev: sõna "asub"

annab algvormiks sõna "asuma", samas nime "Berliini" lemmadeks on nii "Berliin" ja "berliini". Tänu suure algustähe säilitamisele nimede puhul saab hiljem sisendinfost kohanimede tuvastamisel päringute arvu madalal hoida, kuna kontrollitakse ainult suure algustähega sõnu. Teisest küljest kohustab see kasutajat olema korrektne: sõna "berliini" puhul kohanimede otsingut läbi ei viida.

3.4.3. Funktsioon annaLemmadeMassiiv

Funktsiooni eesmärgiks on eristada kasutaja lausungist sõnad, kustutada kirjavahemärgid (v. a. komad arvude puhul) ja suunata need funktsiooni *annaLemma*. Väljundiks on lause kõigi sõnade kõik lemmad, hilisemas töötluses enam ei arvestata, milline oli sõna vorm sisendis.

3.4.4. Funktsioon sisendinfo

Funktsioon tuvastab sisendlemmade hulgast päringute tegemiseks olulist infot. Praeguses variandis otsitakse sisendist kolme sorti sõnesid:

- arvud (ka kümnendkoha eraldajaks olev asendatakse koma punktiga);
- riigid;
- linnad.

Riikide ja linnade puhul valitakse välja vaid suure tähega algavad sõnad (suur täht säilib nimede puhul ka pärast lemmatiseerijat) ja kontrollitakse neid teenuse *GeoNames Country Info* abil. Kui leitakse riik, aga ühtegi linnanimetust sisendis ei tuvastata, võetakse linnaks riigi pealinn ning koha koordinaatideks pealinna koordinaadid. Kui leitakse linn ja riiki pole seni leitud, salvestatakse mällu linn ja selle koordinaadid ning riik. Mällu jäetakse kogu sisendist üks linn, üks riik ja kaks arvu.

3.4.5. Funktsioon taidaSlotid

Antud funktsioon koostab vastavalt valitud teenusele päringu, kasutades argumentide väärtustamisel eelmistest voorudest jäänud või viimasest sisendist saadud andmeid. Vastuseks saadud XML-faili alusel täidetakse kasutatava teenusega seotud infoslotid. Kui vastav väärtus vastuses puudub, kustutatakse vana väärtus ikkagi, et säilitada andmete terviklikkust.

3.4.6. Funktsioon annaVastus

Funktsioon *annaVastus* koostab etteantud malli ja infoslottide sisu põhjal ning kasutaja valikutele tuginedes infot sisaldava vastusfraasi. Arvulised andmed vormindatakse vastavalt Eestis kasutatavale mallile (koma kui kümnendkohtade eraldaja, tühik eraldab kolmenumbrilisi gruppe) ja ujukomaarvud teisendatakse normaalkujule. Juhul, kui kasutaja on valinud arvude teisendamise tavatekstiks, rakendatakse siin ka funktsiooni *arvTekstiks*. Mittearvulised andmed tõlgitakse funktsiooni *tolkimine* abil eesti keelde. Seejärel täidetakse mallis olevad lüngad. Kui vastuse koostamisel leitakse vajalike väärtuste hulgast tühi slott, katkestatakse mallilünkade täitmine ja teatatakse info mitteleidmisest.

3.4.7. Funktsioon annaSuhtlusfraas

Kasutaja sisendist loetakse peale muu info välja ka nö. suhtlusfraase: tervitusi, hüvastijättu, tänamist. Vastavalt sellele suudab antud funktsioon lisada kas muu info ette või järele sobiva vastuse (vt. Lisa 3), sellistel puhkudel on vastus mitmelauseline.

3.4.8. Funktsioon tolkimine

Selle funktsiooni abil teostatakse tõlkimist, ette antakse tõlgitav tekst ning vajadusel siht- ja algkeel (vaikimisi inglise keelest eesti keelde). Kui kasutaja pole häälestuses seda ära keelanud, kasutatakse kõigepealt süsteemisisest sõnastikku (vt. Lisa 2), vastava fraasi puudumisel pöördutakse funktsiooni *googleTranslate* poole.

3.4.9. Funktsioon googleTranslate

Funktsioon kasutab teenust *Google Translate*. Etteantud teksti ning vajadusel siht- ja algkeele (vaikimisi eesti ja inglise keel) põhjal tagastatakse tõlgitud tekst. Sõltuvalt sisendist on kvaliteet äärmiselt kõikuv, eesmärgiks on parema võimaluse puudumisel teksti mõttest aru saada.

3.4.10. Funktsioon arvTekstiks

Antud funktsioon teisendab etteantud numbrilisel kujul oleva arvu tekstilisele kujule nimetavas käändes. Funktsioon suudab arvestada ka kümnendmurdude ja negatiivsete arvudega. Arv jagatakse täis- ja murdosaks ning neid töödeldakse eraldi. Kumbki osa jagatakse kolmenumbrilisteks komponentideks ja iga komponenti töödeldakse sarnaselt.

3.5. Tehnilised nõuded

Programm on kirjutatud programmeerimiskeeles PHP (*Hypertext Preprocessor*), kasutatud on kujunduse eesmärgil ka stiililehti (*CSS, Cascading Style Sheets*). Veebiserverisse peab olema installeeritud vähemalt PHP 5. Veebiteenuste kasutamine nõuab kaugserveris olevate failide avamiskäskude *file_get_contents* ja *fopen* lubamist. Lisaks sellele kasutab salvestatud voorude vaatamine vahendit GD (*Image Processing...*, 2010), mille abil näidatakse diagrammidena küsimuste statistikat.

Süsteemi majutuse ja installeerimise lihtsuse huvides on loobutud muudest nõuetest serverile (sh. näiteks andmebaas). Programmi installeerimiseks serverisse piisab failide kopeerimisest avalikult ligipääsetavasse kausta. Juhul, kui soovitakse kasutada ka dialoogiaktide salvestamist, tuleb failis *index.php* omistada muutujale *$logimine* väärtus *true* ja anda kataloogile *log* kõigile kasutajatele lisaks muule ka kirjutusõigus, selles kataloogis hakatakse hoidma vastavaid logifaile.

Kasutajalt nõutakse vaid graafilist veebibrauserit. Süsteemi arendamisel on hoidutud muudest kliendipoolsetest vahenditest (näiteks *JavaScript*, küpsised (*cookie*) vms), mis võiksid nõuda programmi kasutajalt oma veebibrauseri ümberseadistamist, kõik tööks vajalik on lahendatud serveripoolsete vahenditega.

3.6. Uute veebiteenuste lisamine süsteemile

Nagu eespool kirjeldatud, on käesolev süsteem suhteliselt universaalne. Tänu sellele on uue sarnase teenuse lisamine lihtne, seda eeldusel, et ka uus teenus on samuti XML-väljundiga ja vajalikke argumente antakse päringus edasi URLi abil. Kui sisendinfos piirdutakse kahe arvu, linna ja riigi tuvastamisega, siis piisab järgmistest operatsioonidest:

- Lisada massiivi *teenused* uus element koos nime, kirjelduse, veebiaadressi, argumentide loetelu, märksõnade ja testaadressiga. Tähele tuleb panna, et ükski märksõna ei kattuks varasemate teenuste märksõnadega, vastasel korral kasutatakse nimekirjas eespool asetsevat sama märksõna sisaldavat teenust.
- Lisada massiivi *slotid* uus element, mis koosneb massiivist, mille elementide võtmeteks on väärtuseni viiv tee (*xpath*). Kui soovida, et konkreetse infosloti väärtuse muutumisel muutuks automaatselt ka mõni argumendina kasutatava mälusloti sisu, tuleb vastava argumendisloti võti sisestada kui infosloti väärtus.

- Lisada massiivi *fraasid* uus element, millel on olemas vähemalt üks järglane võtmega *default*, kuid võib olla ka teisi ilma võtmeta järglasi. Iga järglase e. fraasi puhul tuleb määrata märksõnad (mis ühe teenuse piires ei tohi korduda, küll aga võib kasutada samu märksõnu teiste teenuste fraasidega), infoslottide massiiv, mida see fraas kasutab (tee XML-failis õige väärtuseni e. *xpath*), ning mall, milles lüngad on tähistatud [*n*], kus *n* on vahemikus 0 kuni kasutatavate infoslottide arv miinus 1.
- Vajadusel ja võimalusel tuleb modifitseerida ka süsteemisisest sõnastikku, et väljundfraasid oleksid võimalikult kvaliteetsed.

Kui uus teenus nõuab sisendinfost ka muude andmete peale ühe riigi, ühe linna ja kahe arvu tuvastamist, tuleb lisada vastav kood ka funktsiooni *sisendinfo*. Sarnaselt praegustele lahendustele tuleb leida viis, kuidas vajalik info üles leida. Praegusel kujul on arve tuvastatud eeltöötluse järel PHP funktsiooni *is_numeric* abil, linnade ja riikide puhul on kasutatud teenust *Google Maps Geocoding*.

3.7. Testimistulemused ja probleemid

Programmi testimiseks on süsteemile lisatud küsimuste ja vastuste salvestussüsteem: alles hoitakse päringu aeg, arvuti IP-number, teenuse nimetus, küsimus ja vastus sellele. Esimese kahe eesmärgiks oli pakkuda võimalust grupeerida ja seostada dialoogiakte. Katsetamisel osales kümme inimest (nii nais- kui meessoost vanuses 22 kuni 62 aastat, nii informaatikuid, muusikuid, filolooge kui teisi), kelle suhtlust pärast süsteemi kasutamist analüüsiti ning kellel paluti ka vastata küsimustikule. Arvestades testijate väikest hulka ei saa teha kaugeleulatuvaid järeldusi, vaid võtta tulemusi kui juhuslikke arvamusi.

Küsimustiku vastustest selgus, et kasutajate arvamused erinevad suuresti. Muuhulgas paluti hinnata küsimuste mõistmist oma valdkonna piires, vastuste sisulist ja keelelist korrektsust, kasutusmugavust ja disaini. Kasutusel oli viiepalline hindamisskaala, kus arv 1 tähendas hinnangut "nõrk", 5 hinnangut "väga hea". Kõige äärmuslikumalt hinnati küsimuste mõistmist, hinnangud ulatusid kõige madalamast hindest kõige kõrgemani, keskmiselt 3,1. Kui vaadata esitatud küsimusi, tundub üheks peamiseks probleemiks olevat puudulik sissejuhatava teksti ja abiinfoga tutvumine: üritati küsida asjasse puutumatuid küsimusi ("Mis on kolmnurga pindala?", "Beethoveni sünniaasta", "Mis värvi on armastus?"), kohanimede puhul ei kasutatud suurt algustähte, arvutusülesannetes anti arvud tekstilisel kujul ("5 korda 5", mitte "viis korda viis"), samas tehtemärgid märkide, mitte sõnadena ("2 + 2", mitte "2 pluss 2"). Ilmselt arvestati, et süsteemiga võib vabalt ükskõik

millisel teemal ja vormis suhelda, mõne vooru järel reeglina tutvuti ilmselt abiinfoga ja kvaliteet enamasti paranes oluliselt. Vastuste sisulist ja keelelist korrektsust hinnati kõrgemalt (vastavalt 3,6 ja 4,1), viimase kohta märgiti küll ära konarlikkust, ühest küljest on see seotud teenuse *Google Translate* puudustega, seda eriti Wikipedia artiklite tõlkimisel, aga ka linnanimede puhul, mida tõlgiti sageli käändesse (nime "Berlin" tõlge oli "Berliin" asemel "Berliinis"). Teisest küljest aga ka vastuste konstruktsiooniga: kõiki andmeid sai kasutada vaid nimetavas käändes ning seetõttu polnud lauseehitus sageli parim ("Riigi Taiwan pealinn on Taipei", "Riigi Holland rahvaarv on 16 645 000"). Parima hinnangu (vastavalt 4,5 ja 4,3) said kasutusmugavus ja disain.

Positiivsetest külgedest toodi välja võimalus esitada küsimusi lakooniliselt (ilma täislauseta), geograafiline kaart, kiiresti haaratav kasutusjuhend. Ära märgiti ka mugav kasutajaliides ja lihtne disain. Täienduste poolelt pakuti välja loomulikus keeles arvude sisestamist ja kõnesünteesi lisamist, reeglina olid mõtted siiski sellised, mis praeguse ülesehituse puhul pole rakendatavad: rahaühikute tõlkimine loomulikku keelde, mitmeosaliste nimede tuvastamine, tõstutundlikkuse kaotamine nimede puhul.

Kokkuvõtvalt võib öelda, et testimine andis väga olulist tagasisidet, osa saadud märkustest ja ideedest sai arvestada koheselt (märksõnade lisamine, abiinfo täiendamine jne). Samuti lisati süsteemile ka võimalus eelmistest voorudest tuvastatud teenuste arvestamine uue vastuse puhul välja lülitada: nii tuleb küsimused esitada küll täislausega, kuid arusaamatu küsimuse korral ei püüa arvuti igal juhul mingit positiivset vastust välja pakkuda. Selle valiku puhul muutub dialoog küll ebamugavamaks, kuid korrektsemaks.

Üheks oluliseks probleemiks on sisenditöötluse piiratus. Võimatu on küsida infot mitmesõnaliste objektide kohta, näiteks Los Angelese või San Marino kohta, sest sisendit töödeldakse ühe sõna haaval. Kuna linnade ja riikide tuvastamine toimub pärast lemma leidmist *Google Maps* abil, oleks lisaks ühesõnalistele ka kahesõnaliste fraaside läbiproovimine teenusele oluliselt koormavam, nimelt tehtaks päringuid oluliselt rohkem ning samuti oleks tulemus ebamäärasem – leitaks ülearuseid geograafilisi asukohti, mida küsija silmas ei pidanudki.

Ka teenuste töökorras olek tekitas probleeme: süsteemi loomise käigus tekkis olukord, kus üks teenustest oli paar nädalat ebastabiilne. Kuna tegemist oli ühega abiteenustest, mille

korrasolek oli süsteemi jaoks hädavajalik, sai see asendatud teise teenusega, mis polnud küll nii sobiv, kuid eelmisest töökindlam.

3.8. Edasiarendusvõimalusi

Lisaks uutele teenuste lisamisele on süsteemil olulisi edasiarendusvõimalusi. Süsteem ei peaks piirduma ühe vooru puhul vaid ühe teenusega, vaid korraga võiks vastata ka mitme päringu abil leitud info abil. Samuti saaks lisada sisendituvastusele mitme linna, riigi vm. mäluslottidesse salvestamise, päringud tehakse sel juhul iga leitud väärtuse kohta eraldi ning esitatakse kas mitme lause või liitlausena.

Täiustamist vajaks kindlasti sisenditöötlus. Kui praegu analüüsib süsteem sisendit ühe sõna haaval, siis on teenuste väljakutsumiseks ja fraaside valikuks võimalik kasutada vaid ühesõnalisi tunnusfraase, samuti peavad ka linna- ja riiginimetused olema vaid ühesõnalised. Nii on võimatu küsida näiteks New Yorki ilma või Los Angelese kellaaega.

Parandamist vajaks ka väljundfraaside kvaliteet. Praegusel juhul suudab süsteem kasutada kas leitud tekstilise info algvormis või sisemise sõnastiku abil määratud vormis vastust. Seega on vastavalt lahendusele võimalik koostada fraas nii, et info oleks nimetavas käändes ("Riigi Holland pindalaks on...", mitte "Hollandi pindalaks on...") või piirduksid vastusevariandid mingi piiratud nimekirjaga (näiteks ilmaga seotud fraasid, mis on antud sisemise sõnastiku kirjetena). Näiteks ingliskeelse süsteemi puhul sellist probleemi poleks, kuna tihti saaks kasutada eessõnu ja vastus ise jääks muutumatuks. Abiks oleks eesti keele morfoloogilise süntesaatori kasutuselevõtt.

Süsteemi töökindluse huvides oleks vajalik ühest küljest valida võimalikult stabiilseid teenuseid, teisest küljest aga püüda neid dubleerida. Dublant ei pruugi olla tingimata samasuguse struktuuriga, kuid peab katma vastusfraaside infovajaduse.

Süsteemi mugavamale kasutamisele aitaks kaasa ka administreerimisliides, mis lubaks administraatoril lisada teenuseid veebiliidese abil, arvatavasti tuleks siis andmete hoidmiseks kasutada andmebaasi või (süsteemi majutuse universaalsuse ja hoolduse lihtsuse huvides) XML-faile. Viimane on eriti soovitatav, kuna andmete lugemine on sagedane, kuid muutmine toimub harva.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli luua dialoogsüsteem, mis suudaks vastata kasutaja loomulikus eesti keeles esitatud küsimustele, võttes aluseks erinevat infot pakkuvad veebiteenused. Prototüüpsüsteemi jaoks valitud veebiteenuste hulka kuulusid riikide kohta infot andev *GeoNames Country Info*, Wikipedia artikleid vahendav *GeoNames Wikipedia Fulltext Search*, konkreetse linna ajatsoonile vastavat kuupäeva ja kellaaega pakkuv *GeoNames Timezone*, linnade koordinaate ja asukohariiki esitav *Google Maps Geocoding*, ilmateadet andev *Weather Underground* ja lihtsaid arvutustehteid tegev *Calculator*. Lisaks neile kasutab programm ka teisi abiteenuseid sõnade algvormide leidmisel, tõlkimisel, kaardi kuvamisel ja IP-numbri alusel kasutajainfo leidmisel.

Töös on lühidalt tutvustatud loomuliku keele töötluseks mõeldud vahendeid ja eesti keeleruumi jaoks loodud vastavaid ressursse, samuti dialoogsüsteemide ülesehitust ja põhimõtet. Samuti on antud ülevaade veebiteenustest ning nende positiivsetest ja negatiivsetest külgedest.

Töö põhiosa koosneb valminud süsteemi ülesehituse tutvustusest. Välja on toodud kasutatud teenuste kirjeldused, põhilised funktsioonid ja tehnilised nõuded. Kirjeldatud on programmi katsetamise tulemusi ja kasutajate arvamusi, nende põhjal on arutletud süsteemi probleemide ja edasiarendusvõimaluste üle.

Communication with Public Web Services
in Estonian Natural Language

Sven Aller

Summary

The aim of the thesis was to create a dialog system for communication with web services in natural Estonian language. Six different web services were selected for the prototype system: *GeoNames Country Info* (information about countries: capital, population, area, etc.), *GeoNames Wikipedia Fulltext Search* (articles about cities from Wikipedia), *GeoNames Timezone* (the date and time in a chosen city), *Google Maps Geocoding* (coordinates and location of a city), *Weather Underground* (weather reports by cities) and *Calculator* (simple calculations). In addition, the system also uses other web services for finding lemmas, translating, showing a map and getting information about a user by his IP number.

The thesis briefly describes the tools for processing the natural language, resources of automatic language processing in Estonian as well as basic principles and construction of dialog systems and web services. The main part of the thesis introduces the structure of the created dialog system. It points out the specifications of the web services used, basic functions and technical requirements. Also it outlines the results of testing the system and the opinions of the users, followed by a discussion about the problems and further development of the dialog system.

Kasutatud kirjandus

1. *Alfred* (2010). http://www.dialoogid.ee/alfred/. Kontrollitud 17.05.2010.
2. *Eesti keele arendamise strateegia 2004-2010* (2010). http://www.hm.ee/index.php?popup=download&id=4149. Kontrollitud 17.05.2010.
3. *Eesti keele keeletehnoloogiline tugi 2006-2010* (2010). http://www.keeletehnoloogia.ee/. Kontrollitud 17.05.2010.
4. *Eesti keele lemmatiseerija* (2010). http://www.filosoft.ee/lemma_et/. Kontrollitud 15.05.2010.
5. *Filosoft* (2010). http://www.filosoft.ee/. Kontrollitud 17.05.2010.
6. Harold, Elliotte Rusty (2004). *XML 1.1 Bible*. Wiley Publishing, Indianapolis.
7. Hennoste, Tiit; Gerassimenko, Olga, Kasterpalu, Riina; Koit, Mare; Rääbis, Andriela; Strandson, Krista; Valdisoo, Maret (2005). *Questions in Estonian Information Dialogues: Form and Functions*.
8. Hutchins, John (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, nr. 17.
9. *Image Processing and GD* (2010). http://php.net/manual/en/book.image.php. Kontrollitud 28.05.2010.
10. *IT terministandardi projekti (1998-2001) sõnastik* (2010). http://www.keeleveeb.ee/dict/speciality/itstandard/. Kontrollitud 17.05.2010.
11. Jackson, Peter; Moulinier, Isabelle (2002). *Natural Language Processing for Online Applications*. John Benjamins Publishing Co, Amsterdam.
12. Jurafsky, Daniel; Martin, James H. (2009). *Speech and Language Processing*. Pearson Education, New Jersey.
13. Koit, Mare (2002). *Märgendatud dialoogikorpus kui keeletehnoloogiline ressurss*.
14. Konchady, Manu (2006). *Text Mining Application Programming*. Charles River Media, Boston.
15. Sirel, Raul (2009). *Kohaldatava teadmusbaasiga eestikeelne dialoogsüsteem*. Bakalaureusetöö.
16. *Statistikaamet* (2010). http://www.stat.ee. Kontrollitud 15.05.2010.
17. *Zelda* (2010). http://www.igemeravi.ee/zelda/. Kontrollitud 17.05.2010.

18. Treumuth, Margus (2002). *Eestikeelne dialoogsüsteem lennuinfo andmebaasi liidesena*. Bakalaureusetöö.
19. Treumuth, Margus; Alumäe, Tanel; Meister, Einar (2006). A natural language interface to a theater information database. *Language Technologies, IS-LTC 2006: Proceedings of 5th Slovenian and 1st International Conference, 9th - 10 October, Ljubljana, Slovenia*. Ljubljana.

Lisad

Lisa 1. Näidisdialoogid

Kasutaja: Milline on Tallinna ilm?
Arvuti: Tallinn: täna on selge ilm. Temperatuur on 11 kuni 24 kraadi.

Kasutaja: Mis on Soome pealinn?
Arvuti: Riigi Soome pealinn on Helsinki.

Kasutaja: Aga Rootsi?
Arvuti: Riigi Rootsi pealinnaks on Stockholm, rahvaarv on umbes 9 045 000 ja pindala umbes 449 964 ruutkilomeetrit. Riigi valuutaks on SEK.

Kasutaja: Kui suur on Venemaa pindala?
Arvuti: Riigi Venemaa pindala on umbes 17 100 000 ruutkilomeetrit.

Kasutaja: Palju on 6 astmes 4?
Arvuti: Arv 6 astmes 4 on 1 296.

Kasutaja: Palju on kell Londonis?
Arvuti: Praegu on linnas London 2010/05/21 19:24.

Kasutaja: Kui palju sooja on praegu Moskvas?
Arvuti: Moskva: temperatuur on 13 kuni 24 kraadi.

Kasutaja: Kui suur on Prantsusmaa rahvaarv?
Arvuti: Riigi Prantsusmaa rahvaarv on 64 094 000.

Kasutaja: Milline on USA valuuta?
Arvuti: Riigi Ameerika Ühendriigid rahaühikuks on USD.

Kasutaja: Mis on Argentiina valuuta?
Arvuti: Riigi Argentiina rahaühikuks on ARS.

Kasutaja: Valga temperatuur?
Arvuti: Valga: temperatuur on 14 kuni 30 kraadi.

Kasutaja: Eesti Valga temperatuur?
Arvuti: Valga: temperatuur on 11 kuni 26 kraadi.

Kasutaja: Eesti Valga temperatuur?
Arvuti: Valga: temperatuur on 11 kuni 26 kraadi.

Kasutaja: Kui suur on arvude 345 ja 652 vahe?
Arvuti: Arvude kolmsada nelikümmend viis ja kuussada viiskümmend kaks vahe on miinus kolmsada seitse.

Kasutaja: Mis on Ungari pealinn?
Arvuti: Riigi Ungari pealinn on Budapest.

Kasutaja: Pindala?
Arvuti: Riigi Ungari pindala on umbes üheksakümmend kolm tuhat kolmkümmend koma ruutkilomeetrit.

Kasutaja: Mis on Albaania pealinn?
Arvuti: Riigi Albaania pealinn on Tirana.

Kasutaja: Mis on Hiina pealinn?
Arvuti: Riigi Hiina pealinn on Pekingis.

Kasutaja: Kui palju on kell Berliinis?
Arvuti: Praegu on linnas Berliinis 2010/05/22 22:33.

Kasutaja: Kus asub Belgrad?
Arvuti: Linn Belgrad asub riigis Serbia.

Kasutaja: Kus on Madriid?
Arvuti: Linn Madrid asub riigis Hispaania. Tema laiuskraadiks on umbes 40,4166909 ja pikkuskraadiks umbes -3,7003454.

Kasutaja: Mitu kraadi täna Tartus sooja on?
Arvuti: Tartu: temperatuur on 7 kuni 15 kraadi.

Kasutaja: Mitme tunni pärast saabub kesköö?
Arvuti: Vabandust, Teie lause oli kas teemaväline või minu jaoks liiga keeruline.

Kasutaja: Mis kell on?
Arvuti: Praegu on linnas Tartu 2010/05/27 15:06.

Kasutaja: Kui palju on ruutjuur kahekümneviiest?
Arvuti: 0. astme juur arvust 0 on 0.

Kasutaja: Milline on ilm Berliinis?
Arvuti: Berliinis: täna on äikesetormine ilm. Temperatuur on 11 kuni 24 kraadi.

Kasutaja: Kui palju on viis korda viis?
Arvuti: Arvude 0 ja 0 korrutis on 0.

Kasutaja: Kui palju on 5 korda 5?
Arvuti: Arvude 5 ja 5 korrutis on 25.

Kasutaja: Mis on ruutjuur 25-est?
Arvuti: 5. astme juur arvust 5 on 1,37972966146.

Kasutaja: Kus asub Seattle?
Arvuti: Linn Seattle asub riigis Ameerika Ühendriigid.

Kasutaja: Kus asub Mexico?
Arvuti: Linn Mexico City asub riigis Mehhiko.

Kasutaja: Kui suur on Tartu?
Arvuti: Vabandust, ma ei saanud Teist hästi aru.

Kasutaja: Kui suur on Tartu pindala?
Arvuti: Riigi Eesti pindala on umbes 45 226 ruutkilomeetrit.

Kasutaja: Mis rahaühik on Eestis kasutusel?
Arvuti: Riigi Eesti rahaühikuks on EEK.

Kasutaja: Mis koordinaatidel asub Tartu?
Arvuti: Linn Tartu asub riigis Eesti.

Kasutaja: Kui palju inimesi elab Eestis?
Arvuti: Riigi Eesti rahvaarv on 1 307 000.

Kasutaja: Kui palju inimesi elab USAs?
Arvuti: Riigi Ameerika Ühendriigid rahvaarv on 303 824 000.

Kasutaja: Kui palju on 5 pluss 7 miinus 2?
Arvuti: Arvude 5 ja 7 summa on 12.

Kasutaja: Mis on Valgevene pealinn?
Arvuti: Riigi Valgevene pealinn on Minsk.

Kasutaja: Mis kell tõuseb päike Tartus?
Arvuti: Praegu on linnas Tartu 2010/05/27 15:18.

Kasutaja: Kus asub Kärdla?
Arvuti: Linn Kärdla asub riigis Eesti.

Kasutaja: Mis värvi on armastus?
Arvuti: Vabandust, ma ei mõistnud Teid, püüdke oma küsimus ümber sõnastada.

Kasutaja: palun infot Madridi kohta
Arvuti: Madrid Wikipedia andmetel: Madridi (häältdatakse / mað id / hispaania, inglise keeles) on kapitali ja suurima linna Spain. Linn asub jõe Manzanares kesklinnas riigis. vahel autonoomsete Kastiilia ja Leóni ja Castilla-La Mancha (...) (vt. ka http://en.wikipedia.org/wiki/Madrid)

Kasutaja: Millised on La Paz'i koordinaadid
Arvuti: Vabandust, Teie lause oli kas teemaväline või minu jaoks liiga keeruline.

Kasutaja: Venemaa pealinn
Arvuti: Riigi Venemaa pealinn on Moskva.

Kasutaja: Ungari rahaühik
Arvuti: Riigi Ungari rahaühikuks on HUF.

Kasutaja: :), millised on La Pazi koordinaadid
Arvuti: Vabandust, ma ei saanud Teist hästi aru.

Kasutaja: kus asub La Paz
Arvuti: Vabandust, ma ei leinud sobivat infot!

Kasutaja: Aga mis on Boliivia pealinn?
Arvuti: Riigi Boliivia pealinn on La Paz.

Kasutaja: Tartu ilm?
Arvuti: Tartu: täna on vihmavõimalusega ilm. Temperatuur on 6 kuni 16 kraadi.

Kasutaja: Tartu info
Arvuti: Tartu Wikipedia andmetel: Tartu (ajaloolised nimed:. Saadud esialgne nimi Tarbatu.) On suuruselt teine linn Eestis. Erinevalt mitmed Eesti poliitilise ja rahalise kapitali Tallinn. Tartu peetakse tihti intellektuaalse ja kultuurilise jaoturiga. eriti kuna see on koduks mitmed Eesti vanima ja tuntuima ülikooli. Asub 186 km kagus Tallinn. Linnas on kesklinnas Lõuna-Eesti (...) (vt. ka http://en.wikipedia.org/wiki/Tartu)

Kasutaja: Võru info
Arvuti: Võru Wikipedia andmetel: Võru (ja valla Kagu-Eestis. Ta on pealinna Võru maakond ja Võru kesklinnast Parish. (...) (vt. ka http://en.wikipedia.org/wiki/V%C3%B5ru)

Kasutaja: Põltsamaa info
Arvuti: Vabandust, ma ei leinud sobivat infot!

Kasutaja: Eesti Põltsamaa info
Arvuti: Põltsamaa Wikipedia andmetel: Põltsamaa on linn Jõgeva maakonnas. Eestis. See asub umbes. (...) (vt. ka http://en.wikipedia.org/wiki/P%C3%B5ltsamaa)

Kasutaja: Mis on Hollandi rahvaarv?
Arvuti: Riigi Holland rahvaarv on 16 645 000.

Lisa 2. Süsteemisisene sõnastik

Inglise	Eesti
Afghanistan	Afganistan
Angola	Angoola
Argentina	Argentiina
Azerbaijan	Aserbaidžaan
Bahrain	Bahrein
Chance of Flurries	tuisuvõimalusega
Chance of Freezing Rain	jäite võimalusega
Chance of Rain	vihmavõimalusega
Chance of Sleet	lörtsivõimalusega
Chance of Snow	lumesajuvõimalusega
Chance of Thunderstorms	äikesetormi võimalusega
Chance of a Thunderstorm	äikesetormi võimalusega
Clear	selge
Cloudy	pilvine
Colombia	Kolumbia
Denmark	Taani
East Timor	Ida-Timor
Flurries	tuisune
Fog	udune
Freezing Rain	jäitevihma
Gibraltar	Gibraltar
Haze	uduvinene
Iraq	Iraak
Ivory Coast	Elevandiluurannik
Mostly Cloudy	peamiselt pilves
Mostly Sunny	peamiselt selge
Nepal	Nepaal
Oman	Omaan
Overcast	lauspilvine
Partly Cloudy	osaliselt pilves
Partly Sunny	osaliselt selge
Rain	vihmane
Scattered Clouds	hajuspilvine
Sleet	lörtsisajune
Snow	lumesaju
Sunny	selge
Thunderstorm	äikesetormine
Thunderstorms	äikesetormine

Lisa 3. Suhtlusfraaside nimekiri.

Dialoogi algus

- Tere, mina olen DAB. Kuidas saan Teid aidata?
- Tervist, mina olen DAB. Loodan, et saan Teile kuidagi kasulik olla.
- Tere, mina olen DAB. Tore oleks Teiega suhelda.

Reaktsioon tervitusele

- Tervitused ka minu poolt!
- Tervist! Mida Te soovite küsida?
- Tere!

Reaktsioon hüvastijätule

- Kohtumiseni, loodan, et suhtleme varsti jälle!
- Nägemiseni, oli tore suhelda!
- Näeme jälle, olge tubli!
- Head aega!

Reaktsioon tänamisele

- Võtke heaks!
- Olge lahke!
- Palun!

Reaktsioon mittearusaadavale fraasile

- Vabandust, ma ei saanud Teist hästi aru.
- Vabandust, ma ei mõistnud Teid, püüdke oma küsimus ümber sõnastada.
- Vabandust, Teie lause oli kas teemaväline või minu jaoks liiga keeruline.

Lisa 4. CD-plaat programmiga

Bakalaureusetööle lisatud CD sisaldab järgmisi faile:

- kataloog *dab* (sisaldab dialoogsüsteemi faile)
    - kataloog *graafika*
        - abi.gif
        - arial.ttf
        - ikoon_arvuti.gif
        - ikoon_kasutaja.gif
        - logo.jpg
        - pais.jpg
        - viide.gif
    - kataloog *log* (logifailide salvestamiseks)
    - fraasid.php
    - funktsioonid.php
    - gd_diagramm.php
    - index.php
    - kasutaja.php
    - log.php
    - metainfo.php
    - riigid.php
    - salvestalog.php
    - slotid.php
    - sonastik.php
    - stiil.css
    - suhtlusfraasid
    - teenused.php
- server.exe (sisaldab dialoogsüsteemi ja portatiivset veebiserverit operatsioonisüsteemile Windows: MicroApache 2.0.63, PHP 5.2.9, GD 2, vt. http://microapache.amadis.sytes.net/)
- kataloog *src*, Autorun.inf, Menu.exe (CD menüüga seotud failid)