

TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
Eesti ja üldkeeleteaduse instituut
Arvutilingvistika eriala

Raul Sirel

Poolautomaatne teadmusbasiside konstrueerimine dialoogsüsteemidele

Magistritöö

Juhendaja Margus Treumuth

Tartu 2011

Sisukord

Sissejuhatus	4
1. Tehisintellekti olemus	6
1.1. Intellekt ja intellektitehnika	6
1.2. Kas arvuti võib olla intelligentne	7
1.3. Küberneetika ja intellekt	8
1.4. Dialoogsüsteem kui rakenduslik tehisintellekt.....	10
2. Teadmus	11
2.1. Teadmuse mõiste.....	11
2.2. Teadmuse esituse meetodid.....	12
2.2.1. Produktsiooniline meetod.....	12
2.2.2. Loogikal põhinevad meetodid.....	14
2.2.3. Semantilised võrgud ja ontoloogiad.....	15
2.2.4. Freimid	18
3. Informatsiooni kogumine vabatekstist	21
4. Eksperiment.....	22
4.1. Ülevaade.....	22
4.2. Kasutatud tarkvara ja keeleressursid.....	23
4.2.1. Morfoloogiline ühestaja	23
4.2.2. Wordnet.....	24
4.3. Algoritmide kirjeldused	26
4.3.1. Esimese katse algoritm.....	26
4.3.2. Teise katse algoritm	27
4.3.3. Kolmanda katse algoritm	28
4.3.4. Võtmesõnade hulga piiramine	30
5. Tarkvara kirjeldus	31
5.1. Tehniline kirjeldus	31
5.2. Veebiliides.....	32
5.3. Nõuded sisendfailile.....	32
5.4. Väljundi kirjeldus.....	33
6. Eksperimendi tulemuste analüüs.....	34
6.1. Katsetes kasutatud konfiguratsioon.....	34
6.2. Esimese katse tulemuste analüüs	34
6.3. Teise katse tulemuste analüüs	35
6.4. Kolmanda katse tulemuste analüüs	36
6.5. Katsetulemuste võrdlus	36
7. Võimalusi tulemuste parandamiseks.....	39
Kokkuvõte	40
Kasutatud kirjandus.....	41
Abstract	44
Lisa 1. Väljavõte eksperimendis kasutatud küsimuste-vastuste komplektidest.....	45
Lisa 2. Väljavõte esimese katse käigus saadud võtmesõnade komplektidest	47
Lisa 3. Väljavõte teise katse käigus saadud võtmesõnade komplektidest	49
Lisa 4. Väljavõte kolmanda katse käigus saadud võtmesõnade komplektidest.....	51
Lisa 5. Laserplaat eksperimendi sisend- ja väljundtekstidega	

Joonised

Joonis 1. <i>Turingi masina graafiline esitus</i>	13
Joonis 2. <i>Semantilise võrgu näide roomajatest</i>	15
Joonis 3. <i>Näide taksonoomilisest puust (Collins ja Gillian 1969)</i>	16
Joonis 4. <i>Näide HPSG-st</i>	19
Joonis 5. <i>Tüüpiline freim</i>	20
Joonis 6. <i>Esimese katse algoritmi graafiline esitus</i>	26
Joonis 7. <i>Teise katse algoritmi graafiline esitus</i>	27
Joonis 8. <i>Kolmanda katse algoritmi graafiline esitus</i>	29
Joonis 10. <i>Veebiliidese vaade</i>	32
Joonis 11. <i>Näide programmi väljundist</i>	33

Sissejuhatus

Käesolev magistritöö käsitleb teadmuse esitamist ja kogumist kui interdistsiplinaarset probleemi. Tegemist on rahvusvahelisel tasandil kestvalt aktuaalse teemaga, kuna teadmus ning selle esitamise tehnikad kuuluvad mitmesuguste keeletehnoloogiliste aplikatsioonide tööks vajalike eelduste hulka. Sellisteks rakendusteks võivad olla näiteks dialoogi- või muud tehisintellektisüsteemid, mis kasutavad tööks teadmusbaase.

Dialoogsüsteemi (inglise keeles *dialogue agent*) all mõistame arvutiprogrammi, mis suudab inimesega teatud teema(de)l suhelda, kusjuures suhtlus toimub loomulikus keeles ning kõne või teksti vahendusel.¹

Teadmus (eesti keeles ka *teadmised*, inglise keeles *knowledge*) on süstemaatiliseks kasutamiseks organiseeritud faktide, sündmuste ja tõdemuste kogum.² Teadmusbaasiks (inglise keeles *knowledge base*) nimetatakse aga reeglite ja faktide kogumit ainevalla kirjeldamiseks, mis koos tuletusmehhanismiga võimaldab vastata ka sellistele küsimustele, mille vastus ei sisaldu teadmusbaasis ilmutatud kujul.³

Maailmas ja ka Eestis on loodud arvukalt dialoogsüsteeme ning dialoogsüsteemide raamistikke⁴, mis kasutavad enda töös teadmusbaase. Teadmusbaaside konstrueerimine on osutunud aga küllaltki aega- ja ressursinõudvaks ülesandeks: nende loomiseks on tarvilik leida konkreetsele süsteemile sobiv formalism teadmuse esitamiseks ning seejärel esitada vajalikud teadmised valitud formalismi abil. Teadmiste formaalseks kirjapanemiseks on tarvis läbi töötada suures mahus informatsiooni ning transformeerida see valitud formalismile sobivale kujule. See kõik võib osutada küllaltki aeganõudvaks ülesandeks, mistõttu on mõistlik üritada vähemalt osaliselt seda protsessi automatiseerida.

Kirjeldatud probleemi lahendamiseks pakutakse käesolevas magistritöös välja kolm poolautomaatset meetodit teadmusbaaside loomise lihtsustamiseks, kasutades selleks vabatekstianalüüsi. Meetodite demonstreerimiseks ning testimiseks luuakse

¹ Koit 2003. Lk 119

² IT terministandardi projekti sõnastik

³ Liikane, Kesa 2006

⁴ Treumuth 2010

eksperimentaalne tarkvara, mis võimaldab küsimuste ja vastuste komplektidest kolmel erineval meetodil automaatselt võtmesõnu genereerida, et neid saaks kasutada dialoogsüsteemide teadmusbases. Töös kasutatud meetodeid nimetatakse poolautomaatseteks, kuna nende abil loodud teadmusbasis vajavad enne kasutuselevõttu inimesepoolset järelkontrolli ja kohaldamist.

Eestikeelsetest tekstidest teadmuse ekstraheerimist sellel eesmärgil seni teadaolevalt katsetatud ega realiseeritud ei ole.

Magistritöö käsitlus on interdistsiplinaarne: kokkupuutepunkte on lisaks lingvistikale ka arvutiteaduse, psühholoogia, neurobioloogia ja teadusfilosoofiaga. Käsitletakse tehisintellekti olemust, selle relatsiooni inimintellektiga ja peamisi intellektitehnika probleeme.

Magistritöö on jaotatud seitsmeks peatükiks, millest esimeses tutvustatakse tehisintellekti olemust, selle relatsioone inimintellektiga ja peamisi intellektitehnika probleeme. Teises peatükis antakse ülevaade tuntumatest teadmuse esituse meetoditest ning kolmandas käsitletakse lühidalt vabateksti analüüsiga seotud probleemistikku. Neljas peatükk hõlmab magistritöö raames läbi viidud eksperimendi ning selleks kasutatud meetoodika ja ressursside kirjeldamist. Viiendas jaos antakse lühike tehniline ülevaade loodud tarkvarast, kuuendas analüüsitakse eksperimendi tulemusi ning seitsmendas peatükis arutletakse võimaluste üle katsetulemuste parandamiseks.

Magistritöö praktilisse poolde kuuluv tarkvara on loodud kasutades programmeerimiskeeli Python ja PHP. Viimase abil on programmile loodud ka veebiliides, mille abil on lõppkasutajal võimalik tarkvara kasutada.

1. Tehisintellekti olemus

1.1. Intellekt ja intellektitehnika

Intellekt on üks neist mõistetest, mida inimene igapäevaselt kasutab ning mille olemusest arusaamist iseenesest mõistetavaks peab, ent mille defineerimisel või selgitamisel alatihti raskustesse satub. Seda ka täiesti põhjendatult. Kuigi Õigekeelsussõnaraamat selgitab intellekti lühidalt ja kuivalt kui *mõtlemisvõimet* või *aru*⁵, ei ole maailmatuntud teadlased intellekti defineerimise osas sugugi ühel meelel.

1994. aastal tegid 52 intellekti uurivat teadlast eesotsas Linda S. Gottfredsoniga pöördumise, milles defineerisid intellekti kui väga üldist mentaalset võimet, mis hõlmab võimeid arutleda, planeerida, lahendada probleeme, abstraktselt mõelda, hoomata keerukaid ideid, õppida kiiresti ning teha seda kogemustest. Intellekt ei pidavat olema pelgalt õppimisvõime või kitsas akadeemiline oskus, vaid pigem võime hoomata ümbrust, seda mõista ning sellele reageerida⁶. Samas on välja pakutud ka oluliselt lihtsamaid ning üldisemaid definitsioone. Nii näiteks pakkus inglise psühholoog Sir Cyril Lodowic Burt välja, et intellekt on lihtsalt sünnipärane kognitiivne võime⁷, hiljem on Sternberg ja Salter sõnastanud intellekti kui eesmärgile orienteeritud adaptiivse käitumise⁸ ja Gottfredson kirjeldas intellekti ka kui võimet tulla toime kognitiivse keerukusega⁹.

Tuleb täheldada, et enamik definitsioonidest kasutab sõna *võime*, kuid mitte ükski neist ei omanda seda *võimet* pelgalt inimesele. Vaid Reuven Feuerstein on väitnud, et intellekt on inimesele omane unikaalne võime välismaailmaga adapteerumiseks oma kognitiivse talitluse modifitseerimise läbi¹⁰. Seetõttu kerkivad ka õigustatud küsimused: kas intellekt on vaid inimõistusele omane või on teisedki liigid intellektivõimelised, ning kas intellekt on võimena ka tehnikult realiseeritav?

Viimasele küsimusele üritab vastata filosoofiaharu, mis kannab tehisintellekti filosoofia (inglise keeles *philosophy of artificial intelligence*) nime. Selle alaga töötavad teadlased

⁵ Õigekeelsussõnaraamat 2006

⁶ Gottfredson 1997. Lk 1

⁷ Burt 1931

⁸ Sternberg, Salter 1982. Lk 24

⁹ Gottfredson 1998. Internetaallikas

¹⁰ Haywood, Tzuriel 1992. Lk 9

uritavad vastata veel teistelegi üsna eksistentsiaalsetele küsimustele – kas arvuti on võimeline arukalt (intelligentselt) käituma ning lahendama samu probleeme, mida inimene lahendab mõtlemisega? Kas arvutil võib olla mõistus ja teadlikkus nagu inimesel ning kas see on võimeline tundeid omama? Ning mis vahest kõige olulisem: kas inimese ja arvuti intellekt oleksid üldse omavahel võrreldavad?¹¹

Kuigi inglisekeelne termin *artificial intelligence* ei erista uurimisvaldkonda ning uurimisobjekti, on eesti keeles soovitatav neil siiski vahet teha. Intellektitehnikaks nimetatakse nimelt interdistsiplinaarset uurimisvaldkonda, mis uurib mudeleid ja süsteeme funktsioonide täitmiseks, mida üldjuhul peetakse inimhõimusele omasteks; tehisintellektiks peetakse aga eelkirjeldatud valdkonna uurimisobjekti.¹²

1.2. Kas arvuti võib olla intelligentne

Olgugi, et tehisintellektide jagamine üld- ja rakenduslikeks tehisintellektideks ei ole kuigi levinud praktika, tundub see üldistava faktorina olevat siiski mõnevõrra õigustatud. Nii nimetatakse üldtehisintellektideks (inglise keeles *general artificial intelligence*) neid tehisintellekte, mille vaimsed võimed ei jää inimese omadele alla või koguni ületavad neid, ning rakenduslikeks tehisintellektideks (inglise keeles *applied artificial intelligence*) selliseid süsteeme, mis on loodud kitsasse valdkonda kuuluvate probleemide lahendamiseks.¹³

Kuna eelmainitud jagunemise aluseks on tegelikult intelligentsuse määr, siis võib täiesti põhjendatult esitada küsimuse – mille abil on võimalik välja selgitada, kas ja kui intelligentne arvuti on? Palju on räägitud Turingi testist kui intelligentsuse määrajast – kui arvuti on loomulikus keeles suheldes inimesest eristamatu, ongi tegu intelligentse masinaga¹⁴. Taoline test osutub aga ebaefektiivseks, kui arvutisüsteem ongi programmeeritud inimsuhtlust imiteerima; või teise äärmusena võib eksisteerida nutikaid süsteeme, mis saavad äärmiselt keeruliste ülesannete lahendamisega hakkama, kuid ei mõista inimkeelt.

¹¹ Nath 2009. Lk 11

¹² IT terministandardi projekti sõnastik

¹³ Copeland 2000. Internetiallikas

¹⁴ Turing 1950. Lk 384

Arvutite intelligentsuse osas on oluliselt skeptilisem aga tuntud keeleteoloog John Searle, keda teatakse lingvistide hulgas eelkõige kõneaktiteooria peamise edasiarendajana. Searle nimelt pakkus välja hiina ruumi (inglise keeles *chinese room*) nimelise mõttelise eksperimendi, mis tema meelest tõestab, et arvutid ei saagi olla võimelised intellektiks¹⁵. Eksperiment on lihtsustatult ümber seletatuna järgmine: Searle kujutab end arvutina suletud ruumis, kuhu libistatakse ukse alt hiinakeelseid tekste. Ruumi saabuvad tekstid ei oma tema jaoks mingit tähendust, kuna ta lihtsalt ei mõista hiina keelt, kuid sellegipoolest suudab ta mingit fiktiivset programmeeringut kasutades anda veenvaid hiinakeelseid vastuseid. See näitab, et välismaailmale on võimalik pelgalt programmeeringu järgi reageerida ning usutava reaktsiooni saavutamiseks ei pea eksisteerima mingeid kognitiivseid võimeid, vaid piisab kõigest heast programmist.

1.3. Küberneetika ja intellekt

Küberneetika seisukohalt on tehisintellekti saavutamiseks kaks erinevat suunda: musta kasti küberneetika ja neuroküberneetika. Esimese puhul arendatakse sellist süsteemi, mille sisemine struktuur ning seal toimuvad funktsioonid ei ole teada ning selle vastu ei tunta huvi. Põhimõtteliselt peetakse oluliseks probleemide edukat lahendamist suvalist meetodikat kasutades. Teine lähenemine, neuroküberneetika, keskendub pigem sellele, et probleeme saaks lahendada sarnaselt inimesele – intellektuaalse tegevusega.¹⁶ Kuna antud töös käsitletakse tehis- ja inimintellekti käsikäes, siis keskendub ka käesolev peatükk pigem neuroküberneetilisele lähenemisele.

Kuna neuroküberneetika kaugel eesmärk on võime emuleerida inimese intellektuaalset tegevust, võib täiesti põhjendatult selle ala pühaks graaliks pidada tehislikku inimaju. Selle saavutamiseks tuleb ületada kaks peamist takistust: õppida inimaju piisavalt tundma ja seal toimuv piisaval määral kaardistada ning jõuda tehnoloogiliselt tasemeni, kus arvutite jõudlus oleks selle ülesande täitmiseks piisav.

Vastupidiselt üldlevinud arvamusele on aju-uuringutes tegelikult saavutatud täiesti arvestatavaid tulemusi. Nii näiteks simuleeriti 2006. aasta lõpuks projekti Blue Brain raames roti neokortikaalset sammast. Selline samm on oluline osa imetaja ajukoore ehituses ning inimese ja roti omad on tegelikult üsnagi sarnased, suurim erinevus on

¹⁵ Preston 2002. Lk 17

¹⁶ Wiener 1964. Lk 1

vaid seal leiduvate neuronite arv: rotil 10 tuhat ning inimesel 60 tuhat. Miljonid sellised sambad moodustavad hallaine mis omakorda moodustab umbes 80% inimese ajust ning seda peetakse vastutavaks inimese võime eest mõtelda, mäletada, suhelda ning planeerida.¹⁷

Olgugi, et juhtivad bioloogid suhtusid ettevõtmisesse äärmise ebausuga, osutus eksperiment oodatust oluliselt edukamaks ning autorite sõnul takistab hetkel miljardi inimaju sünapsi simuleerimist antud mudelil kõigest sobivate arvutite puudumine – tarvis oleks läbi töötada umbes 500 petabaiti (1 petabait on 10^{15} baiti) informatsiooni. Praeguse tehnoloogiataseme juures ei ole nii suure andmemahu läbitöötamine aga võimalik, kuna sellise jõudlusega arvuti peaks olema paari jalgpalliväljaku suurune.

Millal aga sellise ülesande lahendamiseks tarvilik arvutusvõimsus saavutatakse, on üsna keeruline ennustada ning teema kuulub pigem futuristide-visionääride pärusmaale. Kuigi Moore'i seaduse järgi kahekordistub mikroprotsessorites kasutatavate transistorite arv iga 24 kuu jooksul, ei saa selline areng kesta igavesti, kuna taoline pöörane kasv ei lähtu mitte mikroprotsessorite suurenemisest, vaid hoopis transistorite mõõtude kahanemisest: selleks, et protsessorile mahuks rohkem transistoreid, tehakse neid järjest väiksemana. Tuntud futuristi Ray Kurzweili hinnangul saab selline kasv otsa 2020. aastaks, mil transistoreid ei saa enam pisemana konstrueerida, kuna need on selleks ajaks vaid mõne aatomi paksused. Samas ei tähenda see suure tõenäosusega mitte arengu peatumist, vaid hoopis paradigmapuhetust, mil areenile asuvad näiteks kvantarvutid.¹⁸

Suure tõenäosusega aga inimintellekti ja –teadvuse täielikuks emuleerimiseks pelgalt tehisajust siiski ei piisa, kuna inimene kogub teadmisi nii ratsionalistlikul (faktidest järeldades) kui ka empiirilisel (kogedes) meetodil. Viimase toimimiseks on vajalikud aga meeled, et välismaailma kogeda ning sealt teadmisi hankida. Sir Edmund Leach leiab lausa, et meelte vahel peab eksisteerima veel mingisugune eriline loogiline mehhanism, mis võimaldab transformeerida ühe meelega poolt omandatud informatsiooni teise meelega tajutavaks (näiteks visuaalseid sõnumeid helilisteks või kombitavaid

¹⁷ Blue Brain Project. <http://bluebrain.epfl.ch/>

¹⁸ Kurzweil 1999. Lk 27–28

sõnumeid lõhnalisteks), kuna inimesel on olemas võime visualiseerida kuuldot või muuta kirjutatud teksti kõneks¹⁹.

1.4. Dialoogsüsteem kui rakenduslik tehisintellekt

Kuna maailm on suur ning vahest inimesegi jaoks raskesti hoomatav, on tavaks konstrueerida selliseid tehisintellektisüsteeme, mis tegelevad vaid konkreetsete probleemide lahendamisega ning orienteeruvad vaid oma pisikeses maailmas. Selliste suletud ja komplikatsioonivabade mikromaailmade heaks näiteks on SHRDLU, robotjäset juhtiv arvutiprogramm, mis ei oma mingeid teadmisi välisilmast, kuid opereerib adekvaatselt ja iseseisvalt oma suletud mikromaailmas värviliste klotsidega.²⁰

Nagu paar peatükki eespool juttu oli, nimetatakse selliseid konkreetsete probleemidega tegelevaid süsteeme rakenduslikeks tehisintellektisüsteemideks. Kuigi dialoogsüsteemid ei pruugi oma töös alati rakendada tuntumaid intellektitehnikaid, võib neid autori hinnangul sellegipoolest liigitada rakenduslikeks tehisintellektideks, kuna neid konstrueeritakse enamasti siiski lahendamaks inimestega suheldes konkreetseid probleeme (näiteks ekspertsüsteemid).

¹⁹ Leach 2010. Lk 30

²⁰ Copeland 2000. Internetiallikas

2. Teadmus

2.1. Teadmuse mõiste

Kui inglise keeles kasutatava sõna *knowledge* puhul on üsna keeruline (kui mitte üldse võimatu) eristada teadmist ja teadmust, siis eesti keele puhul on sellise eristuse tegemine täiesti võimalik. Nii näiteks kirjeldab Õigekeelsussõnaraamat teadmust kui teadmiste kogumit²¹. Ka käesolevas töös kasutatakse neid termineid diakriitiliselt ning teadmistest räägitakse kui teadmuse koostisosadest.

Teadmus on mõiste, mida käsitlevad maailmas väga mitmed teadusvaldkonnad, nii näiteks uurivad psühholoogid, kuidas teadmus inimeses eksisteerida võiks, samuti üritavad psühholingvistid välja mõtelda, kuidas toimub inimese keeleline mõtlemine ning milline on selle relatsioon teadmusega. Lisaks on paari tuhande aasta vältel filosoofid-epistemoloogid uurinud ja arutlenud inimteadmuse loomuse ja päritolu üle ning informaatikud-arvutiteadlased mõnikümme aastat kaalunud võimalusi, kuidas võiks inimese intellektuaalset tegevust ja teadmust arvutitel emuleerida.

Ka ajalooliselt on teadmised olnud alati kõrgelt hinnatud, sõnas juba 16. sajandil Sir Francis Bacon kuulsad sõnad *scientia potentia est* (teadmised on jõud). Niisamuti on nad olnud olulisel kohal mitmetes religioonides, näiteks on Vana Testamendi koosseisu kuuluvas 1. Moosese raamatus juttu hea ja kurja tundmise puust, milles sisalduvad teadmised eristavat inimest ja Jumalat²². Samuti on mõnedes kristlikes konfessioonides teadmised üheks seitsmest Püha Vaimu kingitusest inimestele ning muslimite pühakiri Koraan väidab, et teadmised tulenevat Jumalast enesest²³.

Filosoofias jagatakse teadmised klassikaliselt kaheks: apriorseteks (kogemustest sõltumatud, mõistusliku tuletuse abil saadud teadmised) ning aposterioorseteks (empiirilised ehk kogemuste läbi saadud teadmised)²⁴. Platonit on tõlgendatud dialoog Theaetetus seadvat teadmisele kui uskumusele kaks tingimust: see peab olema tõene

²¹ Õigekeelsussõnaraamat 2006

²² Piibel. 1Mo 3:22

²³ Koraan 2:239

²⁴ Internet Encyclopedia of Philosophy. Internetiallikas

ning põhjendatud²⁵. Teadmust võib seega kokkuvõtlikult ja ülevaatlikult kirjeldada kui põhjendatud uskumuste kogumit maailma (või ainevalla) kohta.

Arvutiteadustes kasutatakse teadmuse mõistet enamasti intellektitehnika või keeletehnoloogia valdkonnast rääkides, kuna just nendes teadusharudes tegutsevad teadmusinsenerid, kelle ülesanneteks on teadmuse kogumine ja teadmusbaaside projekteerimine. Kuigi teadmusbaasid jagunevad tegelikult oma sihtgrupi alusel inimestele kasutamiseks mõeldud teadmusbaasideks (infotelefonid, *help desk*'id) ja arvutisüsteemidele kasutamiseks mõeldud teadmusbaasideks, käsitleb käesolev töö ainult viimast ning teadmusbaasidest rääkides mõeldakse just intellektisüsteemide tarvis loodavaid teadmusbaase.

Teadmusbaaside loomise juures räägitakse ka teadmuse esitusest ja selle meetoditest. Mõned meetodid baseeruvad pigem loogikale ja matemaatikale, kuid osa teadlasi leiab, et õigem oleks siiski teadmusbaase struktureerida sarnaselt inimajule ning talletada teadmust inimkeeles. Tegemist on olulise vaidlusalaga, kuna kerkib põhimõtteline küsimus: kas olulisem on arvutuslik efektiivsus või võimalikult tõelähedane inimteadmuse matkimine?²⁶

2.2. Teadmuse esituse meetodid

2.2.1. Produktsiooniline meetod

Kuigi tehisintellektisüsteemides hakati teadmuse esituseks produktsioonilist meetodit kasutama 1970ndatel aastatel, defineeris produktsiooni esmakordselt juudi päritolu matemaatik ja loogik Emil Leon Post juba 1943. aastal. Posti järgi on $A \rightarrow B$ operaator, mis asendab sisendsõnes osasõne A osasõnega B.²⁷

Produktsioone võib lihtsamalt kujutada „kui A, siis B“ stiilis olevate reeglitena, kus A on tingimus ning B on mingi tegevus, mis sooritatakse, kui tingimus A on täidetud. Levinuim programmeerimiskeel, mis põhineb produktsioonidel, on intellektitehnikas ning arvutilingvistikas kasutatav Prolog²⁸.

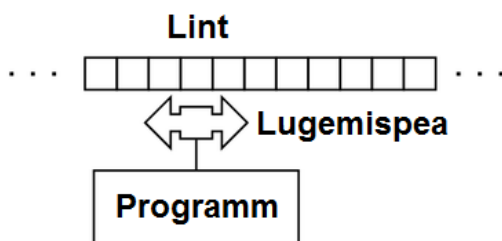
²⁵ Internet Encyclopedia of Philosophy. Internetiallikas

²⁶ Davis, Shrobe, Szolovits 1993. Lk 23

²⁷ Post 1943. Lk 197–215

²⁸ Prolog. <http://www.swi-prolog.org/>

Produksioonidel põhinevad näiteks ka Turingi masinate ümberkirjutusreeglid. Turingi masinateks nimetatakse Alan Turing'i poolt 1937. aastal kirjeldatud abstraktseid arvutusmasinaid, mille eesmärgiks oli uurida võimalike arvutuste ulatust ning piiranguid neile²⁹. Turingi masinat kirjeldatakse üldjuhul kujul $M = \langle K, \Sigma, s, \delta \rangle$, kus K on lõplik seisundite hulk, Σ on lõplik alfabeet, s on lähteseisund ($s \in K$) ning δ on funktsioon hulga $K \times \Sigma$ hulka $K \times (\Sigma \cup \{L, R\})$. Sellises süsteemis esitatakse Turingi masinate ümberkirjutusreeglid kujul $(q_i, a_j) \rightarrow (q_k, X)$, kus q_i ja q_k on seisundid, a_j alfabeedisümbol, mida parasjagu loetakse, ning X kas alfabeedisümbol, mis kirjutatakse a_j asemele või mõni erisümboleist, mis juhib lugemispea liikumist vasakule või paremale (L või R). Graafiliselt kujutatakse Turingi masinaid enamasti sisendlindi, lugemispea ja programmi (vt. joonis 1), kus programmi näol ongi tegemist ümberkirjutusreeglitega, mis omakorda on sisuliselt produktsioonid.



Joonis 1. *Turingi masina graafiline esitus*

Produksioonilisel meetodil põhinevaid teadmusbasse kasutavad mitmed ekspertsüsteemid. Tuntuim neist on ilmselt DENDRAL, mille eesmärgiks oli tundmatute orgaaniliste molekulide identifitseerimine massispektrit analüüsides ning muid keemiaalaseid teadmisi kasutades³⁰. Projekt sai alguse Stanfordini Ülikoolis ning selle eestvedajateks olid Edward Feigenbaum, Bruce Buchanan, Joshua Lederberg ja Carl Djerassi. DENDRAL'ist endast pärinevad veel mitmed teised ekspertsüsteemid, näiteks MYCIN, MOLGEN, MACSYMA, PROSPECTOR, XCON ja STEAMER.

Produksioonilise meetodi suurimaks eeliseks on lihtsus, kuna produktsioonide formuleerimine ei ole kuigi keeruline tegevus. Samuti on formuleeritud reeglid kergesti ning üheselt mõistetavad.

²⁹ Stanford Encyclopedia of Philosophy. Internetiallikas

³⁰ Stefik 1995. Lk 388

Produktsoonilise meetodi puudusteks on juhtimise keerukus ning teadmusbaasi vastuolulisuse välistamine. Esimese probleemi näol on tegemist olukorraga, kus teadmusbaas sisaldab samade tingimuste, kuid erinevate tegevustega produktsioone (ühel situatsioonil võib olla mitu lahendust). Sellistel juhtudel tuleb teha võimalike produktsioonide hulgast valikuid, mis aga langetavad probleemilahenduse efektiivsust. Teine raskus seisneb garanteerimises, et suures teadmusbaasis sisalduv informatsioon ei oleks omavahel vastuolus. Kuna suurte ainevaldade kirjeldamiseks kasutatavad teadmusbaasid võivad olla väga mastaapsed, on produktsioonide mittevastuolulisuse kontrollimine üpris keeruline ülesanne.

2.2.2. Loogikal põhinevad meetodid

Loogikaks nimetatakse teadusharu, mis uurib mõtlemise kõige fundamentaalsemaid aspekte. Kuigi tihtipeale eristatakse filosoofilist ja matemaatilist ehk formaalset loogikat, on nad siiski üksteist täiendavad distsipliinid ning üksteisest sõltumatult neid ilmselt käsitleda ei saakski. Loogika üritab vastata küsimustele, kuidas inimene üldse mõtleb ning kuidas see tegevus struktureeritud võiks olla.³¹ Olgugi, et tegemist on äärmiselt mahuka ja keerulise distsipliiniga, on selles töös antav ülevaade väga kompaktne ja lakooniline, kuna loogikal põhinevad teadmuse esituse meetodid töö praktilises osas kasutust ei leia.

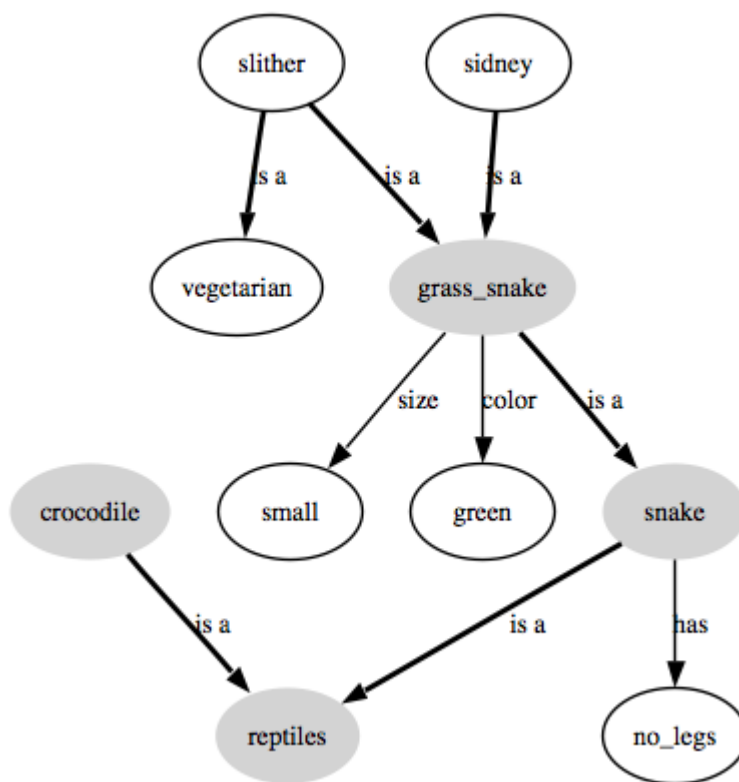
Üldiseim ning vahest ka lihtsaim loogikavorm, mida loogikakursustes enamasti käsitletakse, on lausearvutus. Selline meetod jagab arutluskäigu lauseteks ehk propositsioonideks ning neid on omakorda võimalik jagada osalauseteks. Kui lausearvutus ehk lauseloogika opereerib lausetega, siis selle edasiarendus, predikaatloogika, analüüsib lauseid juba oluliselt detailsemalt. Selleks jagatakse laused objektideks ja predikaatideks, nii näiteks on lauses *Mati istub matil* objektideks *Mati* ja *matil* ning predikaadiks ehk seoseks objektide vahel *istub*. Sellise määratlusega on võimalik teha juba oluliselt üldisemaid järeldusi kui lauseloogikas. Lisaks lause- ja predikaatarvutusele leidub veel mitmeid loogikavorme, näiteks modaal- või hāgusloogika.

³¹ Tamme, Tammet, Prank 1997. Lk 2

Seni on teadmuse esitamiseks enim kasutatud esimest järku predikaatrvutisi, mille valemid võimaldavad väga täpselt defineerida arutluses kasutatavaid objekte ja nendevahelisi seoseid, kuid intellektitehnikas kasutatakse loogikaid lisaks teadmuse esitamisele veel näiteks ka teoreemide automaatseks tõestamiseks.

2.2.3. Semantilised võrgud ja ontoloogiad

Semantilise võrgu (inglise keeles *semantic network*) näol on tegemist orienteeritud graafiga, mille otsad ja kaared on märgendatud, kusjuures otstes on vastavasse ainevaldkonda kuuluvad objektid või mõisted ning kaarteks on nendevahelised semantilised seosed (vt. joonis 2).³²



Joonis 2. Semantilise võrgu näide roomajatest

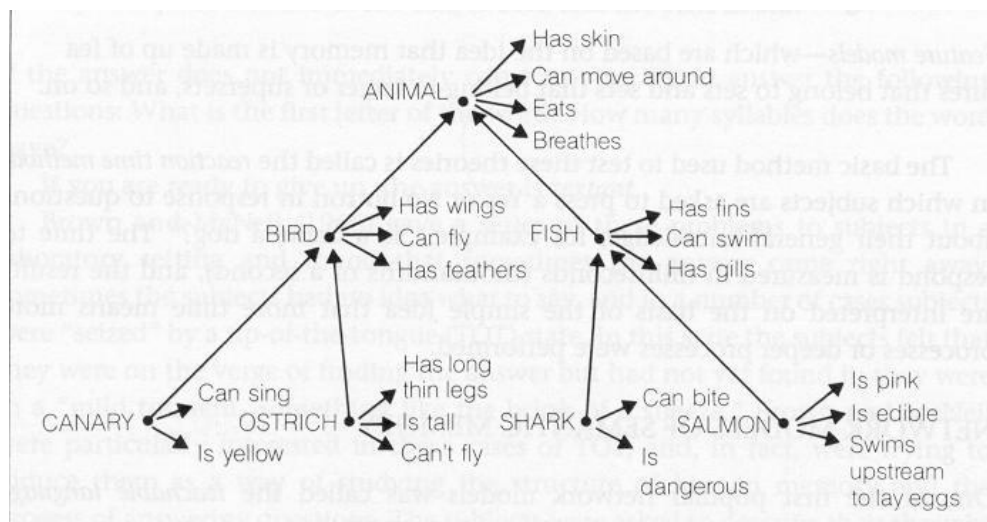
Kuigi arvutiteaduses on semantilistest võrkudest esmakordselt juttu 1950ndatel aastatel, kui neid üritati R. Richensi ideel masintõlke alal omamoodi vahekeelena kasutada³³, on semantilisi võrke tegelikult kasutatud filosoofias, psühholoogias ning keeleteaduses

³² Cercone, McCalla 1983. Lk 8

³³ Mel'čuk 1963. Lk 52.

juba oluliselt varem, nimelt leidsid John R. Anderson ja Gordon Bower endi sõnul viiteid semantilistest võrkudest juba Aristotelese ajast³⁴.

Terminina on semantilist võrku esmakordselt kasutanud Ross Quillian 1968. aastal oma doktoridissertatsioonis, kus ta üritas esitada sõnatähendusi võimalikult sarnaselt sellele, kuidas need võiksid esineda inimese mälus. Võrke esitatakse graafiliselt enamasti IS-A hierarhiate (vt. joonis 2) või taksonoomiliste puudena (vt. joonis 3).



Joonis 3. Näide taksonoomilisest puust (Collins ja Gillian 1969)

Semantiliste võrkude kui kontseptsiooni tuntumateks realiseerimiseks on erinevad leksikaal-semantilised andmebaasid ehk *wordnet*'id. *Wordnet* ehk teaurus on liik mõistelist sõnaraamatut, kus mõisted ei ole organiseeritud mitte alfabeetiliselt, vaid nendevaheliste semantiliste suhete alusel³⁵. Teauruse põhiliseks üksuseks on sünohulk (inglise keeles *synonym set*, *synset*), mille moodustavad ühte mõistet tähistavad sünonüümsed sõnad ja sõnaühendid. Oluline on sünohulkade puhul eristada termineid *mõiste* ning *sõna*, kuna ühel mõistel võib keeles leiduda mitu tähistajat (sünonüümia).

Kuigi *wordnet*'e leidub mitmes keeles, on kõigi nende aluseks Princetoni Ülikoolis loodud teaurus, mis sisaldab üle 117 tuhande sünohulga ja 155 tuhande sõna. Princetoni *wordnet*'i projekt algas 1985. aastal ning on töös veel tänapäevalgi, projekti on aastate jooksul finantseerinud mitmed riigiasutused, kes on huvitatud olnud masintõlke arendamisest.³⁶

³⁴ Anderson, Bower 1980. Lk 9

³⁵ G. A. Miller, Beckwith, Fellbaum, Gross, K. J. Miller. 2008. Lk 327

³⁶ Princeton WordNet. <http://wordnet.princeton.edu/>

1998. aastast alates on Tartu Ülikooli arvutilingvistika uurimisrühmas koostatud eesti üldkeele teaurus ehk TEKsaurust, mis sisaldab praegusel hetkel veidi üle 29 tuhande sünohulga ning 44 tuhande leksikaalse üksuse. Projekti autorite sõnul peaks tesauruses sisalduvad kirjed hõlmama enam-vähem tervet eesti keele põhisõnavara tähenduste hulka³⁷. Eesti *wordnet* kuulub ka semantiliste võrgustike süsteemi EuroWordNet, mis ühendab mitme keele tesauruseid läbi keeltevahelise indeksi (inglise keeles *interlingual index*).

Semantiliste võrkude kasutamist teadmuse esitamisel saab põhjendada kognitiivse teooriaga, mis kujutab inimese pikaajalist mälu raamistikuna, kuhu saab vajadusel teadmisi lisada. Samuti väärivad märkimist psühholoogilised eksperimendid, mille tulemustest järeldati, et mõiste omaduse tuletamise kiirus sõltub omaduse ja mõiste omavahelisest kaugusest hierarhias. Nii näiteks kulub inimesel üsna vähe aega järeldamiseks, et hai on ohtlik, kuid järeldusele, et hail on uimed, jõudmine võtab aega juba oluliselt rohkem. Väidetavalt on põhjuseks see, et hierarhiliselt asub omadus *has fins* mõistest *shark* märksa kaugemal kui omadus *is dangerous* (vt. joonis 3). Seega võiks arvata, et inimese pikaajaline mälu on struktureeritud *wordnet*'ile üsna sarnaselt.³⁸

Semantiliste võrkude problemaatika teadmuse esituse meetodina seisneb peamiselt liiga meelevaldsetes struktuurides. Nimelt on võimalik samade tippude ning kaartega luua mitu erinevat graafi. Samuti raskendab semantiliste võrkude rakendamist asjaolu, et nende genereerimine on üsna aeganõudev ja keeruline ning nõuab üsna põhjalikku analüüsi. Sellele vaatamata on *wordnet*'id väga väärtuslikud leksikaalsed ressursid, mis aitavad kaasa keelelise mõtlemise mõistmisele.

Kuigi semantilisi võrke kasutatakse mitmel pool ka tänapäeval, on arvutiteaduses ja informaatikas viimasel ajal räägitud rohkem hoopis ontoloogiatest. Ontoloogia termin on võetud filosoofiast (olemisõpetus) ning seda defineeritakse kui kontseptualisatsiooni detailset spetsifikatsiooni³⁹. Kui aga abstraktne definitsioon kõrvale jätta, võib ontoloogiat kirjeldada kui semantiliste võrkude edasiarendust, millega üritatakse üle saada semantiliste võrkude peamisest raskusest – meelevaldsusest. Iga ontoloogia opereerib eelnevalt defineeritud piiratud arvu objektidega ning nendevaheliste seostega,

³⁷ TEKsaurus <http://www.cl.ut.ee/ressursid/teksaurus/>

³⁸ Collins, Quillian 1972. Lk 117

³⁹ Gruber 1993. Lk 199

samuti on erinevalt semantilistest võrkudest objekte võimalik argumentidega kirjeldada. Ontoloogiad on semantilistest võrkudest oluliselt formaalsemad ka selle poolest, et nende koostamisel ja kodeerimisel kasutatakse konkreetseid formaalseid keeli nagu OWL (Web Ontology Language), OKBC (Open Knowledge Base Connectivity) või KM (Knowledge Machine).

Kui semantilisi võrke kasutatakse tänapäeval pigem psühholoogias ja psühholingvistikas, siis ontoloogiaid pigem informaatikas ja arvutiteaduses ainevaldade ehk domeenide kirjeldamiseks.

2.2.4. Freimid

Freimi mõiste on kasutusel mitmetes teadusdistsipliinides, mistõttu on veidi komplitseeritud ka selle ühene defineerimine. Tuleneb ta inglisekeelsest sõnast *frame*, mis tähendab raami, struktuuri või kaadrit.

Keeleteaduses käsitletakse freime peamiselt semantiliste struktuuriskeemidena, mis kirjeldavad situatsioonide struktuure, tuues välja komponendid ja nendevahelised (rolli)seosed⁴⁰. Freimid on seega sisuliselt situatsioonide skemaatilised representatsioonid. Veelgi lihtsamalt võiks freimi kirjeldada kui struktuuri, mis seob puhtlingvistilist informatsiooni ning välismaailma kohta omatavaid teadmisi. Lisaks semantikale on freimid kasutuses ka mõnedes süntaksikäsitlustes, üheks neist on HPSG ehk Head-Driven Phrase Structure Grammar (vt. joonis 4).

Freimi kontseptsioon ei pärine algselt aga mitte üldsegi keeleteadusest, vaid hoopis psühholoogiast. Sellesse distsipliini ilmus freimi idee skeemi nime all juba 1932. aastal, initsiaatoriks oli Sir Frederic Bartlett⁴¹. Kognitivist George Mandleri sõnul aga kirjeldas skeemi juba Immanuel Kant 1781. aastal⁴².

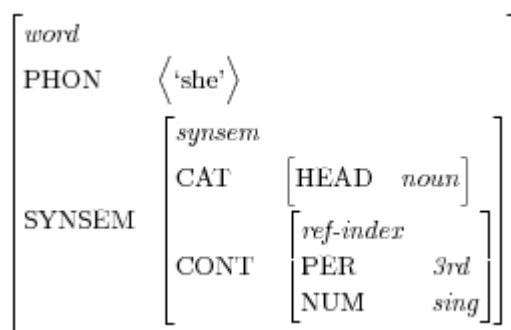
Tänapäevase freimiteooria nime all tuntav on aga sisuliselt skeemiteooria teine tulemine, nimelt võttis mõiste *frame* selles tähenduses esmakordselt kasutusele Marvin Minsky 1974. aastal. Tema teooria põhiolemus toetus hüpoteesile, et inimene talletab

⁴⁰ Orav 1998

⁴¹ Tracey, Morrow 2006. Lk 52

⁴² D'Andrade 1995. Lk 122

kõiki situatsioone oma mälus stereotüüpsetena, freiminimeliste struktuuridena, mida saab vajadusel aktiveerida või modifitseerida. Freimides sisalduv informatsioon on mitmekülgne: see võib ütelda, mis situatsioonides ja kuidas kõnealust freimi rakendada, mida mingist situatsioonist oodata ning kuidas neis käituda. Freim koosneb freimi nimest ning paljudest informatsiooni sisaldavatest terminalidest (vt. joonis 5), mille kirjeldamiseks on inglise keeles kasutusel termin *slot*. Kuigi päris täpset eestikeelset vastet antud mõistele ei olegi, võib rääkida neist kui pesadest. Iga pesa puhul võib olla eraldi defineeritud, millist informatsiooni see sisaldada võib ning mis olukordades seda kasutada saab. Kõik teadmisi sisaldavad freimid moodustavad kokku freimisüsteemi või –võrgustiku (inglise keeles *frame-system*).⁴³



Joonis 4. Näide HPSG-st

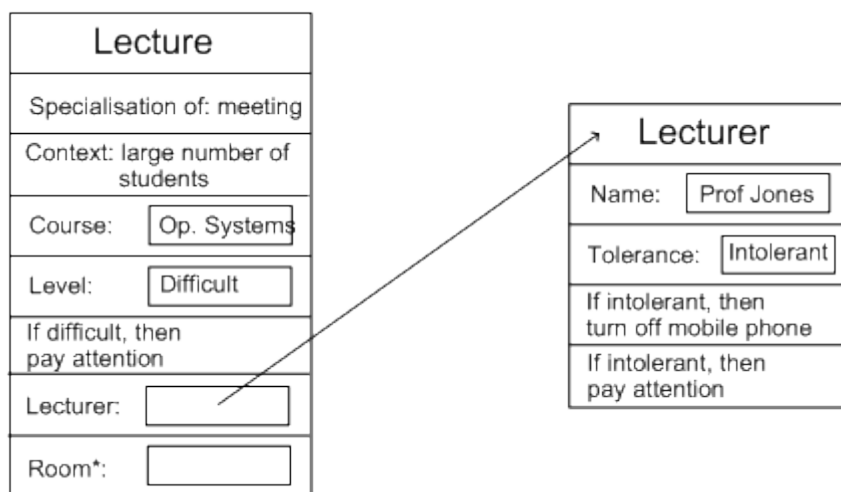
Omamoodi freimide süsteem on loodud Berkeley arvutiteaduste instituudis FrameNET-i projekti raames. Tegemist on veebipõhise leksikaalse andmebaasiga, mis sisaldab üle 11 tuhande leksikaalse üksuse, millest 6800 on täielikult anoteeritud rohkem kui 960 semantilises freimis ning tekstinäiteid on rohkem kui 150 tuhandes anoteeritud lauses. Anoteerimine tähendab teksti varustamist lisainformatsiooniga, mis aitab teksti paremini analüüsida ja realiseerida.⁴⁴

Kognitiivse lähenemise puhul freimiteooriale on oluline ka, kuidas toimub freimidesse informatsiooni lisamine ning freimide loomine. Selle paremaks kirjeldamiseks on 1975. aastal Charles Fillmore kasutanud freimi aktiveerimise mõistet. Näiteks teksti lugemist alustades aktiveeritakse (või kui sobivaid freime veel pole, luuakse) kõigepealt teemaga seotud freimid, kuid suurem osa nende freimide pesadest on tühjad. Teksti edasi lugedes

⁴³ Minsky 1974. Internetiallikas

⁴⁴ FrameNET. <http://framenet.icsi.berkeley.edu/>

või seda läbi töötades hakatakse aga järjest aktiveeritud freimidesse informatsiooni lisama.⁴⁵



Joonis 5. Tüüpiline freim

Freimide kasutamisel teadmuse esituse meetodina on mitmeid eeliseid: informatsioon on näiteks ülevaatlik ning väga mugavasti kättesaadav. Samuti on selle meetodi kasutamine veenvalt põhjendatav antud alal tehtud psühholoogialaste uuringutega. Sarnaselt aga semantiliste võrkudega on selle elegantse meetodi raskusteks meelevaldsus (freimidel ehk mõnevõrra vähem) ning freimisüsteemide koostamise aeganõudvus.

⁴⁵ D'Andrade 1995. Lk 123

3. Informatsiooni kogumine vabatekstist

Kui selge süvastruktuuriga andmebaasidest on informatsiooni ekstraheerimine suhteliselt komplikatsioonivaba ja lihtne tegevus, siis loomulikus keeles kirja pandud tekstidest millegi kasuliku kättesaamine on juba oluliselt keerulisem ülesanne. Viimasest lähtuvalt tegeletakse enamasti kitsendatud probleemidega, milleks on näiteks nimeüksuste, terminoloogia või semantiliste relatsioonide ekstraheerimine.

Üheks heaks näiteks semantiliste relatsioonide tuvastamisest on NELL ehk *Never-Ending Language Learning*, mis kujutab endast Carnegie-Melloni Ülikoolis välja töötatud arvutisüsteemi, mis tegeleb internetist kättesaadavatest tekstidest semantiliste seoste leidmisega. Nagu nimest järeldada võib, on tegemist katkematu protsessiga – Yahoo! superarvutiklastril jooksev NELL on 2010. aasta algusest saati pidevalt töötanud ning oktoobrikuuks 2010 oli arvutisüsteem ära õppinud 440 tuhat fakti. Sarnaselt inimese õppimisprotsessile on ka NELL võimeline infohulga suurenedes varem leitud fakte ümber hindama.⁴⁶

Käesoleva magistritöö praktiline osa on tihedalt seotud terminoloogia ekstraheerimisega (inglise keeles *terminology extraction*) vabatekstidest. Tegemist on andmekaeve alamülesandega, mistõttu võib seda nimetada ka terminoloogia kaevamiseks (inglise keeles *terminology mining*). Ülesande lahendamiseks kasutatakse peamiselt mitmeid keele automaattöötlusvahendeid nagu morfoloogilised ja süntaktilised analüsaatorid (nii statistilistel kui reeglipõhistel formalismidel põhinevad), fraasipiiride määravad jne. Neid vahendeid kasutatakse peamiselt selleks, et identifitseerida vabatekstist osi (sõnu või fraase), mis võiksid olla terminoloogilised üksused.

⁴⁶ NELL. <http://rtw.ml.cmu.edu/rtw/>

4. Eksperiment

4.1. Ülevaade

Käesoleva töö raames läbiviidava eksperimendi eesmärk on katseliselt välja selgitada, kas ja kui kvaliteetselt on võimalik mõne lihtsa algoritmiga küsimuste-vastuste komplektidest ekstraheerida võtmesõnu, mida oleks võimalik kasutada mõne dialoogsüsteemi töös selle teadmusbaasis.

Eksperimendi olulisus seisneb peamiselt selles, et kuigi eksisteerib väga mitmeid erinevatel formalismidel põhinevaid masinõppe meetodeid, mille abil on võimalik vabatekstidest informatsiooni ekstraheerida ning faktidevahelisi semantilisi relatsioone tuvastada, on need enamasti väga keerulised ning kognitiivsest aspektist ka pahatihti täiesti põhjendamatud. Samuti on eksperimendi väljund küllaltki valdkonnaspetsiifiline ning üldisest masinõppe metoodikast pisut kaugemale kalduv: dialoogsüsteemide teadmusbaaside poolautomaatse konstrueerimise katsetamine.

Eksperiment koosneb kolmest eraldiseisvast katsest, millest igaühes on kasutatud võtmesõnade ekstraheerimiseks erinevat algoritmi. Kuigi algoritmide tööpõhimõtte on üldjoontes üsna sarnane, leidub neis siiski põhimõttelisi erinevusi, mis võiksid tulemuste kvaliteeti oluliselt mõjutada. Kõigi kolme algoritmi põhjalikum kirjeldus asub peatükis 4.3.

Katsete edukaks läbiviimiseks on loodud programmeerimiskeeli Python ja PHP ning UNIXi tööriistu kasutades ka vajalik tarkvara, mis võimaldab veebiliidese vahendusel katseid kõigi kolme algoritmiga läbi viia või vajadusel korrata. Tarkvara täpsem kirjeldus on leitav peatükist 5.

Eksperimendis on keelematerjalina kasutatud 100 küsimuste-vastuste komplekti, mis on kogutud interneti vahendusel erinevatest korduma kippuvate küsimuste rubriikidest. Tegemist on rubriigiga, mis sisaldab vastuseid sagedasematele küsimustele, mida võib kohati ka triviaalseteks pidada. Materjal ei ole temaatiliselt piiratud, see tähendab, et kasutatud on erinevate valdkondade informatsiooni. Peamiselt on see tingitud sellest, et tulemusi ei saaks interpreteerida kui kontekstist või ainevallast sõltuvaid. Materjali kogudes on silmas ka peetud, et võimalikult paljudel juhtudel oleks küsimustele

vastatud täislauseliselt. Taoliste küsimus-vastus komplektide kasutamise poolt antud katsetes räägib peamiselt informatsiooni küllus. See tähendab, et internetis leidub väga suurel hulgal veebilehti, mis sisaldavad just selliseid triviaalsusi selgitavaid korduma kippuvate küsimuste lehekülgi.

4.2. Kasutatud tarkvara ja keeleressursid

4.2.1. Morfoloogiline ühestaja

Morfoloogilise analüsaatori näol on tegemist arvutiprogrammiga, mis on võimeline sõna vormist lähtudes määrama selle struktuuri, sõnaliigi ja käände või pöörde.⁴⁷ See tähendab, et näiteks sõna *lauale* puhul võimaldab analüsaator automaatselt eraldada selle tüve *laua* ning käändelõpu *-le* ning samuti on programm võimeline määrama, et tegemist on allatiivis oleva substantiiviga.

Näide morfoloogilisest analüüsist:

Mees *mees+0 // _S_ sg n, //*
mesi+s // _S_ sg in, //
peeti *peet+0 // _S_ adt, sg p, //*
pida+ti // _V_ ti, //
kinni *kinni+0 // _D_ //*

Morfoloogiline analüsaator aga ei lahenda mitmesuse probleemi, see tähendab, et näiteks sõna *lood* puhul ei ole analüsaator enam võimeline eristama, kas tegemist on tööriistaga või mitmuse nominatiiviga sõnast *lugu*. Selle probleemi lahendamiseks on tarvis tekst morfoloogiliselt ühestada. Morfoloogiline ühestamine on protsess, mille käigus leitakse kõigi võimalike vormide hulgast üks, mis on konkreetses kontekstis korrektne.⁴⁸

Näide morfoloogilisest ühestamisest:

Mees *mees+0 // _S_ sg n, //*
peeti *pida+ti // _V_ ti, //*
kinni *kinni+0 // _D_ //*

⁴⁷ Kaalep 1997. Lk 6

⁴⁸ Kaalep, Vaino 1998. Lk 1

Käesolevas töös kasutatakse morfoloogilist ühestajat sõnavormide lemmade ehk algvormide leidmiseks ning sõnaliikide määramiseks. Lemmade kasutamine on vajalik selleks, et vähendada eesti keele suhteliselt mahukatest käände- ja pöördeparadigmadest tulenevat leksikoni suurust. See tähendab, et näiteks verbi *laulma* vormid *laulab* ja *laulavad* on küll erinevad, kuid antud ülesande seisukohalt ei oma käände- ja pöörde kategooriad nii suurt tähtsust, et neid analüüsis arvestada. Seetõttu analüüsitakse mõlemaid sõnavorme lemmana *laulma*.

Kui üldjuhul kasutatakse lausete lemmatiseerimiseks tavalist lemmatiseerijat, siis antud töös on kasutatud ülesande lahendamiseks morfoloogilist ühestajat, mille väljund konverteeritakse vajalikule kujule (eemaldatakse näiteks lõpumorfeemid jms). Morfoloogilise ühestaja kasutamine lemmade leidmiseks on vajalik selleks, et vähendada väljastatavate lemmade arvu, kuna kõiki neid arvestatakse potentsiaalsete võtmesõnadena. See tähendab, et ühestaja kasutamine annab võimaluse elimineerida lemmatiseerimise väljundist lemmad, mis ei ole vastavas kontekstis sobivad.

Sõnaliikide (inglise keeles *part of speech*) määramine on vajalik etapp kolmanda katse algoritmis. Lähem kirjeldus asub peatükis 4.3.3.

4.2.2. Wordnet

Käesolevas peatükis käsitletakse leksikaal-semantilise andmebaasi *wordnet* realiseerimist antud magistrیتöös. Selle veidi detailsem kirjeldus ja informatsioon sealt leitava kohta on loetavad peatükist 2.2.3.

Kuna *wordnet* sisaldab endas mõisteid ja nende vahelisi semantilisi relatsioone, on selle kasutamine antud magistrیتöö praktilises osas igati õigustatud. Teatavasti võib ühel maailma objektile eksisteerida mitu keelelist tähistajat. See tähendab, et ühe mõistega võib vastavuses olla mitu sõna; kuivalt nimetatakse sellist situatsiooni sünonüümiaks. Nii näiteks võib tekstis esineda sõna *laev*, kuid samahästi võib selle asemel mingis kontekstis olla ka *alus*. *Wordnet*'i kasutades on võimalik eelnevaid teadmisi laevandusterminoloogiast omamata välja selgitada, et need sõnad võivad osutada samale objektile.

Sageli jääb muidugi võimalus, et sõnad ei ole täissünonüümid ning objekt, millele sõna viitab, selgub alles konteksti lähemalt analüüsid. Protsessi, mille käigus määratakse konteksti arvestades sellised viitesuhted, nimetatakse semantiliseks ühestamiseks (inglise keeles *semantic disambiguation*). Eesti keele jaoks paraku selliseid arvutiprogramme veel saadaval ei ole.

Käesolevas töös ei ole realiseeritud eesti *wordnet*'i esialgset kuju, vaid selle ontoloogiaks transformeeritud versiooni. Ontoloogia eksisteerib XML/RDF⁴⁹ vormingus ning on loodud EKKTT⁵⁰ projekti Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp⁵¹ raames.

Magistritöös kasutatakse *wordnet*'i praktilise osa kõigis kolmes algoritmis, et suurendada võimalike võtmesõnade hulka. Selleks sooritatakse pärast esialgsete võtmesõnade leidmist päring andmebaasi, kust leitakse kõigile seni leitud märksõnadele ka sünonüümid ning lisatakse need väljastatavatele võtmesõnadele. Semantilise ühestamise probleem on lahendamata, mistõttu võib märksõnadele lisanduda ka ebatäpseid sünonüüme.

Kuna *wordnet* sisaldab oluliselt rohkem semantilisi relatsioone kui pelgalt sünonüümia, on võimalik sealt leida ka näiteks hüpero- või hüponüüme ehk ülem- või alammõisteid. See tähendab, et mõiste *laev* järgi on võimalik leida hüperonüüm *veesõiduk* ning hüponüümid *allveelaev*, *aurulaev*, *galeer*, *jaht* jne. Koostatud tarkvara võimaldab lisaks sünonüümidele leida ka hüperonüüme, mis on ka katsetes realiseeritud. Kuigi võtmesõnade hulka ülemmõistete määramine võib tulemuse täpsusele negatiivselt mõjuda (genereeritakse ebatäpseid ja ebaolulisi mõisteid), võimaldavad hüperonüümid siiski domeeni üldisemalt kirjeldada.

Kasutatud andmestikust on võimalik leida ka alammõisteid, kuid seda käesolevas töös realiseeritud ei ole.

⁴⁹ RDF/XML Syntax Specification. <http://www.w3.org/TR/REC-rdf-syntax/>

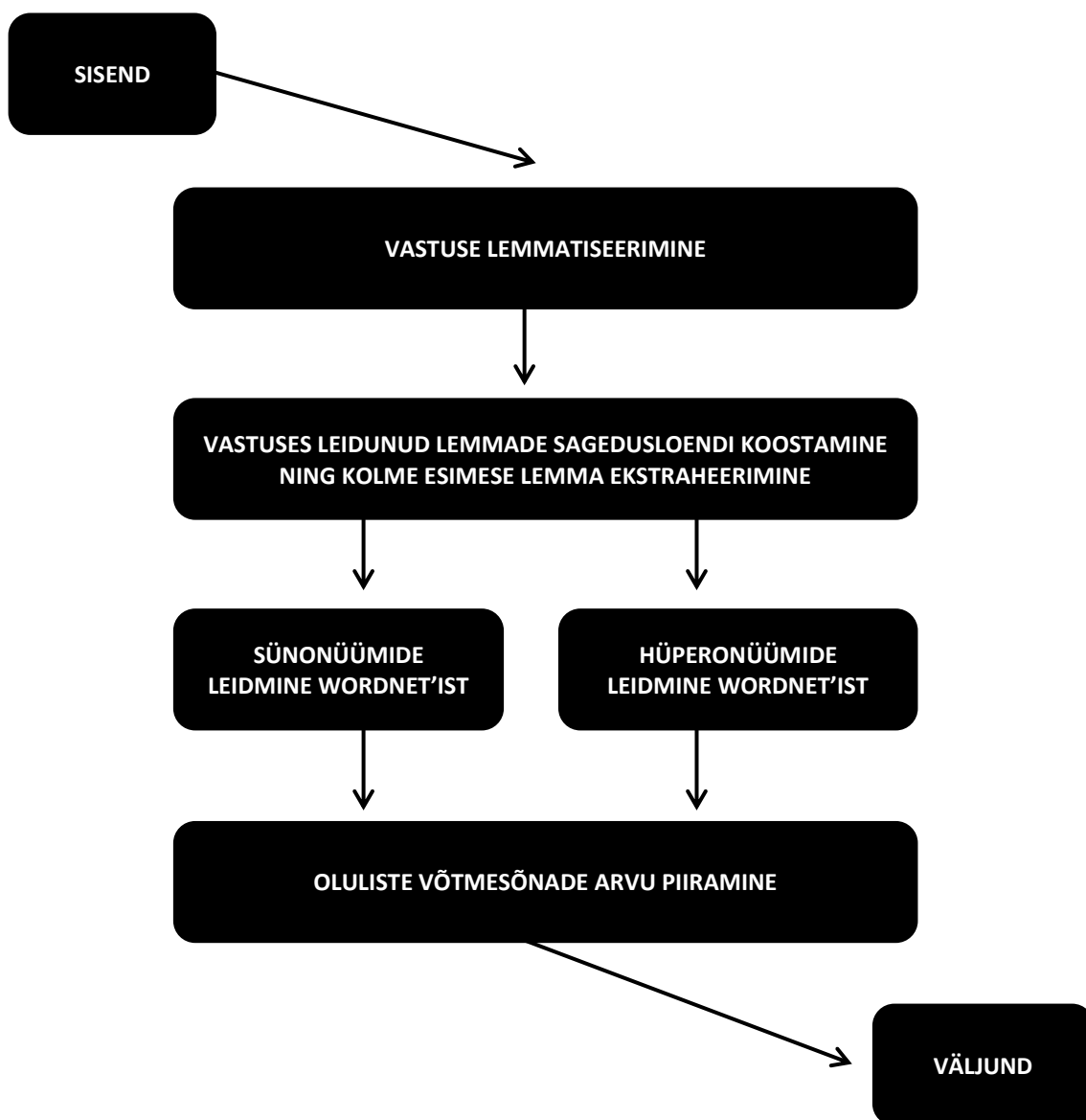
⁵⁰ EKKTT. <http://www.keeletehnoloogia.ee>

⁵¹ EKKTT09-66: Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp <http://ats.cs.ut.ee/semantika/wiki/index.php/Projektist>

4.3. Algoritmide kirjeldused

4.3.1. Esimese katse algoritm

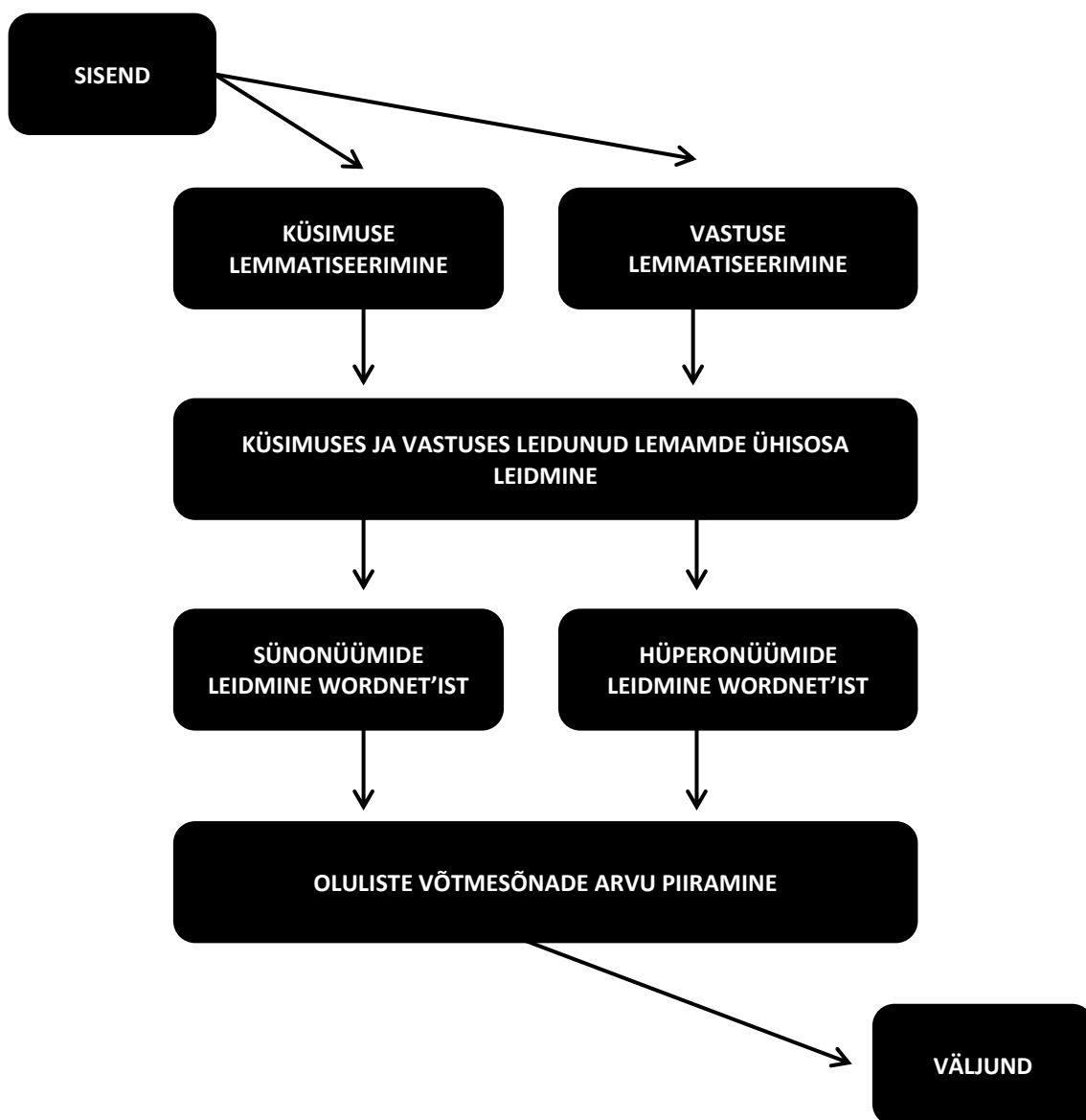
Ekspriimendi esimeses katses kasutatud algoritm on kolmest kasutatud algoritmist kõige lihtsam, see baseerub naiivsele hüpoteesile, et võimalikud kandidaadid võtmesõnadele esinevad vastuses sagedamini kui muud sõnad. See tähendab, et olulisemad sõnad võiksid esineda tihemini kui vähemolulised ning võtmesõnade leidmiseks on tarvis vaadelda vastust kui sõnesid sisaldavat hulka ning leida selles kõige sagedasemad elemendid. Võtmesõnade genereerimisel on sünonüümide ja hüperonüümide leidmiseks kasutatud ka *wordnet*'i. Võtmesõnade hulga piiramise meetodikat on kirjeldatud peatükis 4.3.4. Algoritm on esitatud graafilisena joonisel 6.



Joonis 6. Esimese katse algoritmi graafiline esitus

4.3.2. Teise katse algoritm

Teises katses kasutatud algoritm võtmesõnade leidmiseks on võrreldes esimesega mõnevõrra keerulisem. Selles kujutatakse komplektis sisalduvat küsimust ja vastust kui kahte sõnade hulka, millest tuleb leida ühisosa, kusjuures leitav ühisosa sisaldabki otsitavaid võtmesõnu. Ühisosa kasutamine märksõnade leidmiseks on põhjendatav järgmiselt: kui küsimustele vastatakse korrektselt ja täislauseliselt, mida võiks ka kvaliteetsetest küsimus-vastus komplektidest oodata, peaks oluline terminoloogia sisalduma nii küsimuses kui vastuses. Algoritm sisaldab ka võtmesõnadele *wordnet*'ist sünonüümide leidmist. Võtmesõnade hulga piiramise meetodikat on kirjeldatud peatükis 4.3.4. Algoritmi täpsem kirjeldus on nähtav graafilisena joonisel 7.



Joonis 7. Teise katse algoritmi graafilise esitus

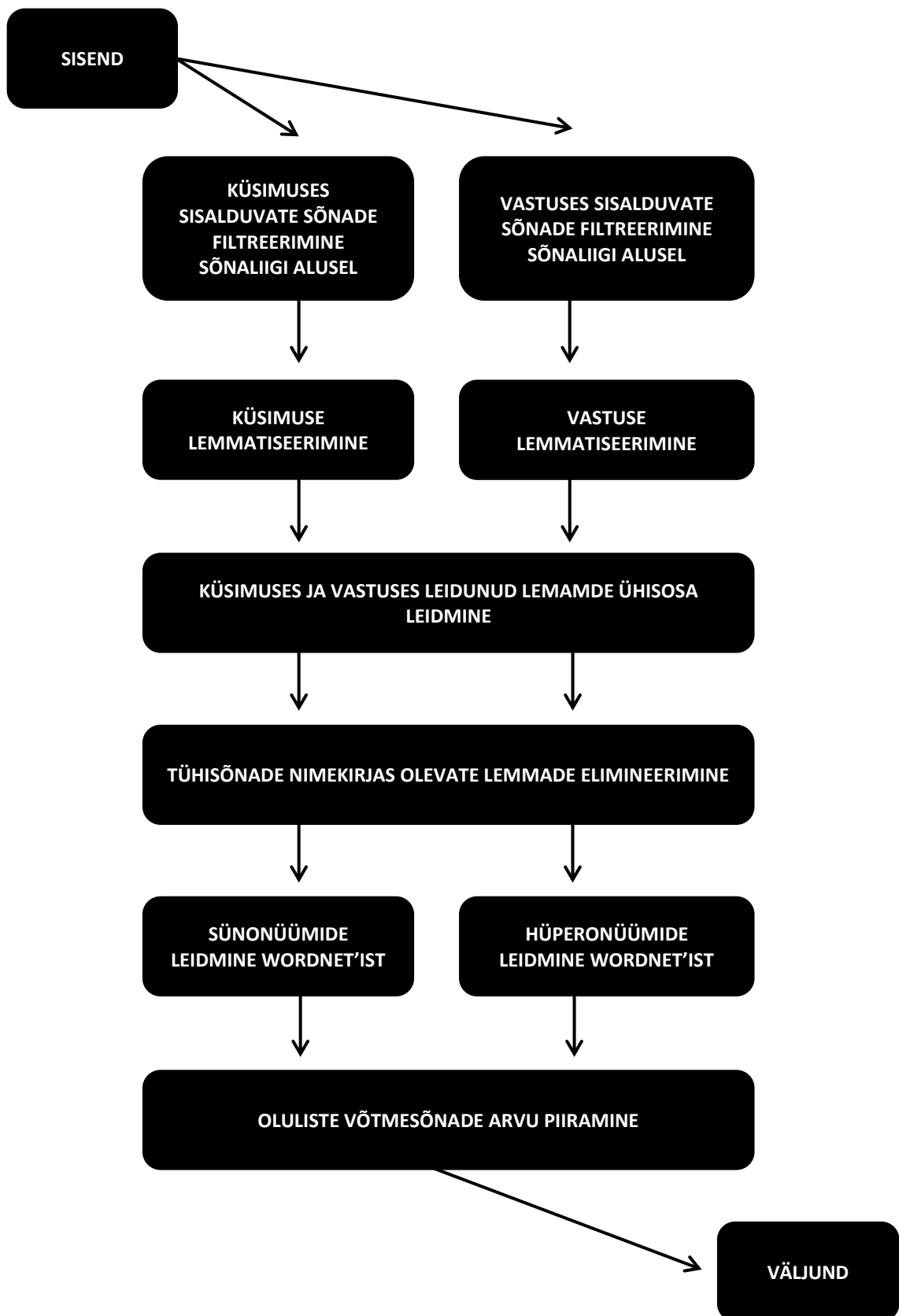
4.3.3. Kolmanda katse algoritm

Kolmandas katses kasutatud algoritm on väga sarnane algoritmile, mida kasutati teises katses, ehk leitakse küsimuses ja vastuses sisalduvate sõnede ühisosa. Põhimõttelise erinevusena on üritatud elimineerida võtmesõnade hulgast selliseid eksemplare, mis ei kannu olulist informatsiooni. Seda on üritatud saavutada kahel meetodil: esiteks sõnaliikide piiramise, teiseks nõ. tühisõnade (inglise keeles *stop words*) elimineerimise läbi.

Esimese meetodi abil lubatakse lemmasid leida vaid sõnadest, mis on saanud morfoloogiliselt ühestamise käigus adjektiivi, adverbil, substantiivi või verbi staatuse. Sõnaliikide piiramine on vajalik selleks, et väljundist oleks võimalik elimineerida sõnu, mis endas olulist informatsiooni ei sisalda. Nii näiteks ei ole ratsionaalne lubada võtmesõnade hulka konjunktsioone või pronomeneid. Lubatavate sõnaliikide hulk on kindlaks määratud tarkvara testimise käigus saadud tulemuste subjektiivse hindamise alusel.

Teise meetodi tarvis on konstrueeritud lühike tühisõnade nimekiri, milles sisalduvaid sõnu ei analüüsita. See on vajalik, et elimineerida väljundist sõnu, mis küll sõnaliigi poolest väljundisse sobiksid, kuid sellegipoolest relevantset informatsiooni ei kannu. Nimekiri on koostatud tarkvara katsetamise ja tulemuste observatsiooni tagajärjel.

Algoritm sisaldab ka sünonüümide leidmist *wordnet*'i andmebaasist. Võtmesõnade hulga piiramise meetodikat on kirjeldatud peatükis 4.3.4. Algoritmi graafiline kirjeldus on nähtav joonisel 8.



Joonis 8. Kolmanda katse algoritmi graafiline esitus

4.3.4. Võtmesõnade hulga piiramine

Dialoogsüsteemide teadmusbases, mis kasutavad teadmuse esituseks siinkasutatavat formalismi, on tavaks võtmesõnade hulga ülempiir fikseerida. Käesolevas töös tehtava eksperimendi raames tuleb seda arvestada juba seetõttu, et ühe küsimus-vastus komplekti kohta ei loodaks liiga palju võtmesõnu. Märksõnade hulga ülempiir on kasutaja poolt fikseeritav veebiliidese kaudu.

Võtmesõnade hulga piiramine toimub kõigi kolme algoritmi puhul vahetult enne võtmesõnade massiivi väljastamist kasutajale. Kuna aga ühtegi heuristikut, mille abil saaks üht võtmesõna teisele selles kontekstis eelistada, autorile teadaolevalt ei eksisteeri, siis toimub piiramine märksõnade päritolu järgi. See tähendab, et eelistatakse neid märksõnasid, mis pärinevad vastusest (esimeses katses) või küsimusest ja vastusest (katsed 2 ja 3) endast ning sünonüüme ja hüperonüüme soositakse vähem, kuna need on enamasti ebatäpsemad.

Tarkvaraliselt on probleem lahendatud nõnda, et eksisteerib tühi massiiv, kuhu esmajärjekorras lisatakse tekstist leitud võtmesõnad, seejärel sünonüümid ning viimasena hüperonüümid. Kui võtmesõnu leitakse kokku rohkem kui kasutaja on veebiliideses fikseerinud, antakse massiivi algusest kasutaja poolt fikseeritud arv märksõnasid. Kui aga märksõnasid on kasutaja poolt valitud piirarvust vähem, väljastatakse kõik leitud võtmesõnad.

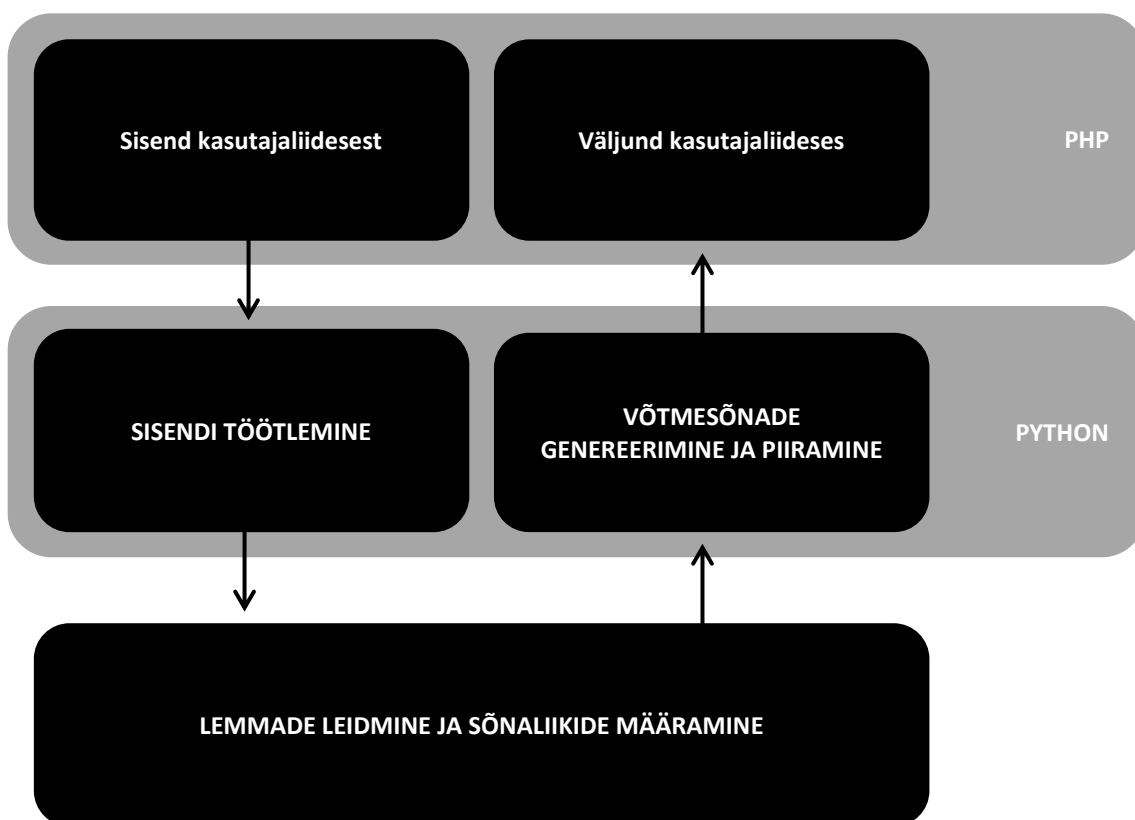
5. Tarkvara kirjeldus

5.1. Tehniline kirjeldus

Eksperimendis kasutatava tarkvara loomisel on kasutatud programmeerimiskeeli Python ja PHP. Esimese abil on loodud programmi põhiosa, mis tegeleb sisendi töötlemise, morfoloogilise ühestaja käivitamise, selle tulemuste töötlemise ja võtmesõnade leidmise ning nende väljastamisega kasutajaliidesele. PHP'd on kasutatud kasutajaliidese programmeerimisel. Tarkvara üldine tööskeem on nähtav joonisel 9.

Tarkvara töötamiseks on vajalik mõne Linuxi operatsioonisüsteemiga server, mis on varustatud PHP (vähemalt PHP4) ja Pythoniga (vähemalt Python 2.4). Kuigi tarkvara võib töötada ka varasema Pythoni versiooniga, ei ole seda testitud.

Kuna loodud programm kasutab enda töös Filosoft OÜ⁵² poolt loodud tarkvaralahendusi, mis ei ole vabavaralised, ei ole ka selle lähtekood vaikimisi saadaval ning tarkvara kasutamine on võimalik ainult veebiliidese vahendusel.



Joonis 9. Tarkvara tööskeem

⁵² OÜ Filosoft. <http://www.filosoft.ee/>

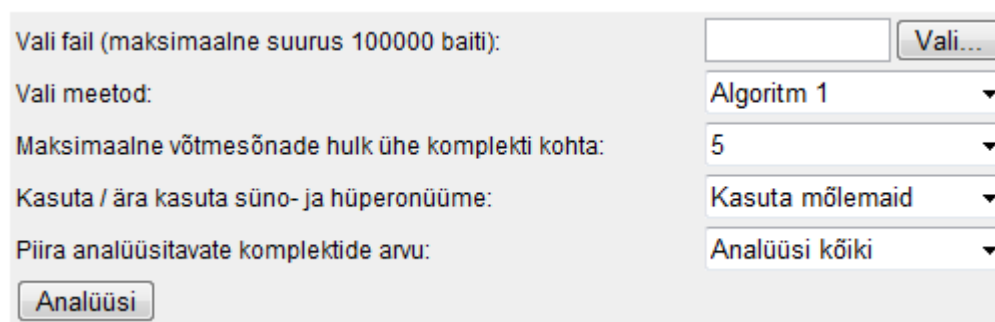
5.2. Veebiliides

Tarkvarale loodud veebiliides võimaldab kasutaja arvutist laadida sisendfaili ning selles sisalduvaid küsimuste-vastuste komplekte peatükis 4.3. kirjeldatud meetoditel analüüsida. Nõutav sisendfaili formaat on kirjeldatud peatükis 5.3.

Liideses on võimalik seadistada järgmised parameetrid (vt ka joonis 10):

- meetod sisendfaili analüüsimiseks (peatükis 4.3. kirjeldatud 3 algoritmi)
- ühe komplekti kohta väljastatavate võtmesõnade piirarv (3, 5 või 10; vaikeväärtus 5)
- sünonüümide ja hüponüümide kasutamine / mittekasutamine (võimalik kasutada ainult sünonüüme või süno- ja hüponüüme koos, samuti võimalik kumbagi kasutamata analüüsida; vaikeväärtusena mõlema kasutamine)
- komplektide arv faili algusest, mida soovitakse analüüsida (5, 10, 25 või kõik komplektid; vaikeväärtusena viimane ehk terve faili analüüsimine)

Viimast opsiooni on soovitatav kasutada mahukama sisendfaili puhul, kuna võtmesõnade leidmine võib olla üsna aeganõudev protsess.



Vali fail (maksimaalne suurus 100000 baiti):	<input type="text"/>	Vali...
Vali meetod:	Algoritm 1 ▾	
Maksimaalne võtmesõnade hulk ühe komplekti kohta:	5 ▾	
Kasuta / ära kasuta süno- ja hüperonüüme:	Kasuta mõlemaid ▾	
Piira analüüsitava komplektide arvu:	Analüüsi kõiki ▾	
<input type="button" value="Analüüsi"/>		

Joonis 10. Veebiliidese vaade

5.3. Nõuded sisendfailile

Sisendfaili nõutav kodeering on UTF-8, see peab olema .txt laiendiga ning turvakaalutlustel ei tohi faili suurus ületada 100000 baiti.

Fail peab sisaldama küsimuste-vastuste komplekte, mis peavad olema eraldatud tühja reaga. Reavahetusena on aktsepteeritavad nii $\backslash n$ kui ka $\backslash r \backslash n$. Küsimuste-vastuste komplektid peavad olema struktureeritud nõnda, et küsimus ja vastus on eraldatud tühja reaga. Lisamärgendus ei ole vajalik.

Näide sisendfaili struktuurist:

Kus saab parkida Kaitsepolitseiametisse tulles?

Kaitsepolitseiameti hoovis pargivad vaid ametiautod. Külalistele parkimine toimub lähedalasuvates parklastes või linna tasulise parkimise alal.

Kui suur on kaitsepolitsei eelarve?

Kaitsepolitseiameti 2010. a eelarve suurus on toodud 2010. aasta riigieelarve seaduses. Kaitsepolitseiameti eelarve liigendus on riigisaladus.

Kui palju on kaitsepolitseis töötajaid?

Kaitsepolitseiameti koosseis ja selle suurus on riigisaladus.

5.4. Väljundi kirjeldus

Programmi väljundiks on analüüsitud sisendfailis olnud küsimus-vastus komplektidest leitud võtmesõnad koos nendele sobivate vastustega, kusjuures väljundist on automaatselt eemaldatud komplektid, millele sobivaid võtmesõnu valitud meetodil ei leitud. Joonisel 11 on näha näide programmi väljundist, mille puhul on sisendina kasutatud peatüki 5.3. lõpus leiduvat näidet. Lisaks on võimalik jooniselt näha, et kolmekomplektilisele sisendile vastab ainult kahekomplektiline väljund. See tähendab, et sisendi viimasele komplektile võtmesõnu valitud meetodil (3. algoritm) ei leitud, mis tuleneb sellest, et küsimuses ja vastuses puudus lemmade ühisosa.

Võtmesõnad:	kaitsepolitseiamet ; parkima
Vastus:	Kaitsepolitseiameti hoovis pargivad vaid ametiautod. Külalistele parkimine toimub lähedalasuvates parklastes või linna tasulise parkimise alal.
Võtmesõnad:	eelarve ; paber ; dokument
Vastus:	Kaitsepolitseiameti 2010. a eelarve suurus on toodud 2010. aasta riigieelarve seaduses. Kaitsepolitseiameti eelarve liigendus on riigisaladus.

Joonis 11. Näide programmi väljundist

6. Eksperimendi tulemuste analüüs

6.1. Katsetes kasutatud konfiguratsioon

Kõigi kolme katse puhul on kasutatud erinevat algoritmi, kuid muud seadistatavad parameetrid on kõigi katsete puhul samad (vt. tabel 1).

Kasutajaliidesest määratav parameeter	Väärtus
Maksimaalne võtmesõnade hulk ühe komplekti kohta	5
Kasuta / ära kasuta süno- ja hüperonüüme	Kasuta mõlemaid
Piira analüüsitavate komplektide arvu	Analüüsi kõiki

Tabel 1. Katsetes kasutatud parameetrid

6.2. Esimese katse tulemuste analüüs

Esimese katse tulemusena saadi sisendina kasutatud 100 küsimus-vastus komplektile 100 komplekti võtmesõnu koos vastustega (vt. tabel 2), mis tähendab, et võtmesõnad leiti kõigile sisendkomplektidele. Tulemus on põhjendatav algoritmi põhimõttega: nimelt valiti vastusest 3 kõige sagedasemat lemmat ning leiti nende süno- ja hüperonüümid, mistõttu ei ole ka võimalik situatsioon, kus mõnele komplektile võtmesõnu ei leita. Ühe sisendis olnud küsimus-vastus komplekti kohta leiti keskmiselt 3,83 võtmesõna (vt. tabel 2).

	Tulemus
Sisendis kasutatud komplektide arv	100
Leitud võtmesõnade arv	383
Komplektide arv, millele võtmesõnu leiti	100
Keskmine võtmesõnade arv komplekti kohta	3,83

Tabel 2. Esimese katse tulemused numbriliselt

Kuigi võtmesõnad leiti kõigile saajale sisendkomplektile, võib esimese algoritmi peamiseks puuduseks pidada tähenduselt mitteoluliste lemmade rohkust võtmesõnade

seas. Selliste lemmade hulka kuuluvad näiteks lemmad nagu *olema, kui, ära, või* jne. Probleemiks võib pidada ka osasünonüümide suhteliselt suurt määra võtmesõnade hulgas. See tähendab, et *wordnet*'ist on sünonüümidenä pakutud tähenduselt sarnaseid, kuid mitte samaväärseid sõnu. Nii näiteks on sõnale *toetus* pakutud sünonüümiks sõna *tugi* või *kaasabi*.

Subjektiiivse hinnanguna tuleb tunnistada, et kuigi võtmesõnade hulgas on kohati üsnagi täpseid ja kvaliteetseid eksemplare, ei ole võtmesõnade üldine kvaliteet kuigi kõrge. See tähendab, et leidub hulgaliselt ebatäpseid ja ebaolulisi võtmesõnu, mis tuleb protsessi hilisemas etapis käsitsi välja filtreerida.

6.3. Teise katse tulemuste analüüs

Teise katse käigus saadi sisendiks olnud 100 küsimus-vastus komplektile väljundiks 90 komplekti võtmesõnade ja vastustega (vt. tabel 3), mis tähendab, et võtmesõnu leiti 90 protsendile sisendis olnud komplektidest. Tulemus on põhjendatav sellega, et algoritm põhineb eeldusel, et olulisemad võtmesõnad võiksid olla küsimuse ja vastuse ühisosas. See tähendab, et kümnel protsendil sisendis olnud komplektidel puudus küsimuse ja vastuse ühisosa. Ühe sisendis olnud komplekti kohta leiti keskmiselt 3,69 võtmesõna (vt. tabel 3), mis on 0,14 võrra vähem kui esimeses katses.

	Tulemus
Sisendis kasutatud komplektide arv	100
Leitud võtmesõnade arv	332
Komplektide arv, millele võtmesõnu leiti	90
Keskmine võtmesõnade arv komplekti kohta	3,69

Tabel 3. Teise katse tulemused numbriliselt

Kuigi väljundis pakutavad võtmesõnad võivad tunduda kohati üsnagi mõistlikud, siis sarnaselt esimese katsega ilmneb võtmesõnade hulgas üsna suurel määral müra, mistõttu on ka selle meetodi puhul tarvilik üpris arvestatav väljundi järeltöötlus. Sarnaselt esimesele katsele leidub ka selle algoritmi väljundis ebatäpseid sünonüüme, mille väljafiltreerimine ei ole kasutatud meetodi puhul võimalik.

6.4. Kolmanda katse tulemuste analüüs

Kolmanda katse käigus leiti võtmesõnad sisendiks olnud 100 küsimuse-vastuse komplektist 81 komplektile (vt. tabel 4), see tähendab, et võtmesõnu leiti 81 protsendile sisendkomplektidest. Võrreldes 2. katsega on leitud võtmesõnu 9 võrra vähemale arvule komplektile, põhjendatav on see algoritmis realiseeritud kitsendustega, mille eesmärk oli vähendada ebatäpsete ja mittevajalike võtmesõnade hulka.

Ühe küsimus-vastus komplekti kohta leiti keskmiselt 3,54 võtmesõna (vt. tabel 4), mis on 0,15 võrra vähem kui teises katses ning 0,29 võrra vähem kui esimeses katses.

	Tulemus
Sisendis kasutatud komplektide arv	100
Leitud võtmesõnade arv	287
Komplektide arv, millele võtmesõnu leiti	81
Keskmine võtmesõnade arv komplekti kohta	3,54

Tabel 4. Kolmanda katse tulemused numbriliselt

Sarnaselt esimese ja teise katse tulemustele on võtmesõnade hulgas endiselt suurel määral ebatäpseid sünonüüme, mis on tarvis järeltöötamise käigus elimineerida. Väljundi hulgas leidub endiselt müra, kuid selle määr on oluliselt väiksem kui esimeses kahes katses, nii näiteks on elimineeritud mittedobivad sõnaliigid ning sõnad, mis kuulusid nõ. tühisõnade nimekirja, nendeks olid verbid *olema*, *tahtma* ja *võima*.

6.5. Katsetulemuste võrdlus

Üldjoontes on näha, et esimese kahe katsega võrreldes on kolmandas leitud oluliselt vähem võtmesõnu, mis ei ole sobilikud. Filtreerimise tulemusel on ka teises ja kolmandas katses vähenenud komplektide arv, millele leiti vähemalt üks võtmesõna (vt. tabel 5). Tuleb aga märkida, et keskmine leitud võtmesõnade arv ühe komplekti kohta on kõigi kolme katse puhul enam-vähem võrdne.

Subjektiiivselt võib hinnata, et kuigi kolmandas katses leiti kõige vähemale arvule komplektidele vähemalt üks võtmesõna, on leitud võtmesõnade kvaliteet võrreldes kahe esimese katse tulemustega oluliselt kõrgem.

	Katse 1	Katse 2	Katse 3
Sisendis kasutatud komplektide arv	100	100	100
Leitud võtmesõnade arv	383	332	287
Komplektide arv, millele võtmesõnu leiti	100	90	81
Keskmine võtmesõnade arv komplekti kohta	3,83	3,69	3,54

Tabel 5. *Katsetulemuste võrdlus numbriliselt*

Kõigi kolme katse puhul on probleeme ebatäpsete sünonüümide leidmisega, seda eeskätt seetõttu, et üheski algoritmis ei ole realiseeritud kontekstianalüüsi. Eelmainitud ebatäpsustest saab aimu alltoodud näitest.

Näide sünonüümide genereerimise negatiivsest mõjust kolmandas katses:

Võtmesõnad: *olev ; hoiatustrahv ; liising ; auto ; essiiv*

Vastus: *Kui kiiruskaamera fikseerib liisingus oleva auto kiiruseületamise, saadetakse hoiatustrahv sõiduki vastutavale kasutajale. Rendiautode puhul saadetakse hoiatustrahvi teade rendifirmale käitub edasi juba rendilepingus kokkulepitud korras.*

Näitest on näha, et verbile *olev* on leitud antud kontekstis ebaõige sünonüüm, milleks on *essiiv*. Tegemist on keeleteaduses kasutatava terminiga, millega tähistatakse olevat käänat. Antud kontekstis ei ole selle kasutamine korrektne.

Kuigi sünonüümide leidmine võib väljundi kvaliteedile kohati negatiivset mõju avaldada, tuleb ka tunnistada, et tihtipeale on siiski leitud väga korrektseid sünonüüme ja hüperonüüme, mille koht võtmesõnade hulgas on igati õigustatud. Selliste olukordade kirjeldamiseks on lisatud allolev näide.

Näide sünonüümide genereerimise positiivsest mõjust kolmandas katses:

Võtmesõnad: nügema ; foto ; ülesvõte ; päevapilt ; pilt

Vastus: *Trahviteatele koopiat fotost ei lisata. Mootorsõiduki eest vastutava isiku taotlusel saadetakse talle koopia fotost, mille abil tegu tuvastati. Taotluse esitamine foto saatmiseks ei peata trahviteate tasumiseks antud 30-päevast tähtaega. Foto saadetakse vastutava kasutaja poolt näidatud elektron- või tavaposti aadressile. Juhul, kui kiiruskaamera poolt salvestatud fotol on sõidukisalongis näha peale sõidukijuhi ka reisijaid, siis hägustatakse isikuandmete kaitsmise eesmärgil kõigi isikute kujutised peale sõidukijuhi.*

Näitest võib näha, et sõnale *foto* on leitud sünonüümid *ülesvõte* ja *päevapilt*. Kuigi viimane ei ole tänapäeval väga laialdaselt kasutatav, on tegu siiski antud kontekstis korrektse võtmesõnaga. Samuti on leitud sobilik hüperonüüm *pilt*.

7. Võimalusi tulemuste parandamiseks

Kuigi tarkvara poolt väljastatavad märksõnade komplektid lihtsustavad ilmselt dialoogsüsteemi teadmusbaasi loomist üsna olulisel määral, tuleb siiski ka nentida, et programmi väljundis leidub mõningal määral ebatäpseid võtmesõnu, mis tuleb käsitsi välja filtreerida. Samuti võib kohati tunduda, et võtmesõnu pole leitud piisaval hulgal, seda eriti teise ja kolmanda algoritmi puhul.

Kuna kõigi kolme kasutatud algoritmi puhul ilmneb suurim osa ebatäpsustest süno- ja hüperonüümide lisamise käigus, on lihtsaim meetod väljundi täpsustamiseks ilmne: mitte kasutada süno- ja hüperonüüme. Teha saab seda kasutajaliideses metoodikat konfigureerides (kirjeldatud peatükis 5.2.).

Teine ja kolmas algoritm baseeruvad küsimuse ja vastuse lemmatiseerimisele ning nendes leidunud lemmade ühisosa leidmisele. Hüpoteesiliselt peaks olema võimalik suurendada leitavate võtmesõnade arvu üsna märkimisväärselt, kui olemasolevaid algoritme modifitseerida selliselt, et sünonüüme leitaks hoopis enne ühisosa leidmist, mitte aga pärast (nagu hetkel realiseeritud on). Taoline muudatus algoritmis võib aga esile kutsuda olukorra, kus oluliselt suureneb ebatäpsete võtmesõnade hulk. See tähendab aga, et sellise modifikatsiooni eelduseks peaks olema ebatäpsuste filtreerimise metoodika parandamine.

Kokkuvõte

Käesolevas magistritöös on käsitletud tehisintellektisüsteemide teadmuse esitamist ja teadmusbasiside konstrueerimist kui interdistsiplinaarset probleemi. On antud ülevaade üldlevinud teadmuse representeerimise meetodikast ning loodud eksperimentaalne tarkvara, mille eesmärgiks on lihtsustada dialoogsüsteemide teadmusbasiside konstrueerimist.

Loodud tarkvara võimaldab leida kasutaja poolt sisestatavatest küsimus-vastus komplektidest võtmesõnu kolmel erineval meetodil, mis on kirjeldatud peatükis 4.3. Tarkvara realiseerimisel on kasutatud kõrgtaseme programmeerimiskeeli Pyhton ja PHP, millest viimase abil on loodud programmile veebiliides.

Töö raames on loodud tarkvara kasutades läbi viidud kolm eraldiseisvat katset ning analüüsitud saadud tulemusi, et selgitada välja kasutatud algoritmide efektiivsus. Lisatud on ka ideid katsetulemuste parandamiseks.

Magistritöö sisaldab nelja lisa, milles leidub väljavõtteid eksperimendis kasutatud küsimus-vastus komplektidest ning katsetes saadud tulemustest. Tööle on lisatud ka laserplaat, millelt on võimalik leida katsetes kasutatud sisendi ning katsete käigus saadud väljundi terviktekstid.

Kasutatud kirjandus

Koit, M. 2003. Märgendatud dialoogikorpus kui keeletehnoloogiline ressurs. Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi Toimetised, 12. Lk 119-136.

IT terministandardi projekti sõnastik.

<http://www.keeleeveeb.ee/dict/speciality/itstandard/> (04.04.2011)

Liikane, L; Kesa, M. 2006. Arvutisõnastik.

http://lepo.it.da.ut.ee/~hkaalep/dict_arvutisonastik.html (04.04.2011)

Treumuth, M. 2010. A Framework for Asynchronous Dialogue Systems. *Frontiers in Artificial Intelligence and Applications: Human Language Technologies — The Baltic Perspective*. Lk 107-114.

Erelt, T; Leemets, T; Mäearu, S; Raadik, M. 2006. Eesti õigekeelsussõnaraamat ÕS 2006. Tallinn: Eesti Keele Sihtasutus.

Gottfredson, L.S. 1997. Foreword to „Intelligence and Social Policy“. *Intelligence* 24.

Burt, C. 1931. The Differentiation Of Intellectual Ability. – *The British Journal of Educational Psychology*.

Sternberg, R; Salter, W. 1982. *Handbook of Human Intelligence*.

Gottfredson, L.S. 1998. The General Intelligence Factor. *Scientific American Presents*.

<http://www.psych.utoronto.ca/users/reingold/courses/intelligence/cache/1198gottfred.html> (04.04.2011)

Haywood, C; Tzuriel, D. 1992. *Interactive Assessment*. New York: Springer-Verlag.

Nath, R. 2009. *Philosophy of Artificial Intelligence: A Critique of the Mechanistic Theory of Mind*. Boca-Raton: Universal Publishers.

Copeland, J. 2000. What is Artificial Intelligence? Turing Archive for the History of Computing.

http://www.alanturing.net/turing_archive/pages/Reference%20Articles/what_is_AI/What%20is%20AI02.html (04.04.2011)

Turing, A. 1950. Computing Machinery and Intelligence. – *Classical Selections on Great Issues Volume VIII: Science, Technology, and Society*. Lanham: University Press of America.

Preston, J. 2002. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*.

Wiener, N. 1964. *Progress in Brain Research Volume 2: Nerve, Brain and Memory Models*.

Blue Brain Project. <http://bluebrain.epfl.ch/> (04.04.2011)

- Kurzweil, R. 1999. *The Age of Spiritual Machines*. London: Orion Books Ltd.
- Leach, E. 2010. *Kultuur ja Kommunikatsioon*. Tallinn: Eesti Keele Sihtasutus.
- Piibel: Vana ja Uus Testament. Soome piibliselts, 1989.
- Koraan. Tallinn: Avita, 2007.
- Internet Encyclopedia of Philosophy. <http://www.iep.utm.edu/> (04.04.2011)
- Davis, R; Shrobe, H; Szolovits, P. 1993. What is a Knowledge Representation? *AI Magazine*.
- Post, E. 1943. Formal Reductions of the General Combinatorial Decision Problem. *American Journal of Mathematics* 65.
- Prolog. <http://www.swi-prolog.org/> (04.04.2011)
- Stanford Encyclopedia of Philosophy. <http://plato.stanford.edu/> (04.04.2011)
- Stefik, M. 1995. *Introduction to Knowledge Systems*. San Fransisco: Morgan Kaufman Publishers.
- Tamme, T; Tammet, T; Prank, R. 1997. *Loogika – mõtlemisest tõestamiseni*. Tartu: Tartu Ülikooli Kirjastus.
- Cercone, N; McCalla, G. 1983. *The Knowledge Frontier*. New York: Springer-Verlag.
- Mel'čuk, I. 1963. *Exact Methods in Linguistic Research*. Berkeley and Los Angeles: University of California Press.
- Anderson, J; Bower, G. 1980. *Human Associative Memory: A Brief Edition*. New Jersey: Lawrence Erlbaum Associates Inc.
- Miller, G. A; Beckwith, R; Fellbaum, C. D; Gross, D; Miller, K. J. 2008. *Introduction to WordNet: An On-line Lexical Database – Practical Lexicography*.
- Princeton WordNet. <http://wordnet.princeton.edu/> (04.04.2011)
- TEKsaurus. <http://www.cl.ut.ee/ressursid/teksaurus/> (04.04.2011)
- Collins, A. M; Quillian, M. R. 1972. Experiments on Semantic Memory and Language Comprehension. – *Cognition in Learning and Memory*.
- Gruber, T. R. 1993. Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* 5(2). Lk 199-220.
- Orav, H. 1998. Eesti keele direktiivverbide semantilise välja struktuur teaurusena. *Magistritöö*.

Tracey, D. H; Morrow, L. M. 2006. Lenses on Reading: An Introduction to Theories and Models. New York: The Guilford Press.

D'Andrade, R. 1995. The Development of Cognitive Anthropology. Cambridge: Cambridge University Press.

Minsky, M. 1974. A Framework for Representing Knowledge.
<http://web.media.mit.edu/~minsky/papers/Frames/frames.html> (04.04.2011)

FrameNET. <http://framenet.icsi.berkeley.edu/> (04.04.2011)

NELL. <http://rtw.ml.cmu.edu/rtw/> (04.04.2011)

Kaalep, H.-J. 1997. An Estonian morphological analyser and the impact of a corpus on its development. <http://www.cl.ut.ee/yllitised/chum1997.pdf> (04.04.2011)

Kaalep, H.-J; Vaino, T. 1998. Vale meetodiga õiged tulemused? Eesti keele morfoloogiline ühestamine statistika abil
http://www.cl.ut.ee/yllitised/kk_yhest_1998.pdf (04.04.2011)

RDF/XML Syntax Specification. <http://www.w3.org/TR/REC-rdf-syntax/> (04.04.2011)

Riiklik programm EKKTT. <http://www.keeletehnoloogia.ee> (04.04.2011)

EKKTT09-66: Nutika süvaveebi- ja veebiressurse kombineeriva infootsüsteemi prototüüp. <http://ats.cs.ut.ee/semantika/wiki/index.php/Projektist> (04.04.2011)

OÜ Filosoft. <http://www.filosoft.ee/> (04.04.2011)

Abstract

Semi-Automatic Knowledge Base Construction for Dialogue Agents

Raul Sirel

This Master's thesis disserts knowledge extraction and representation as an interdisciplinary subject. Among other applications, knowledge bases are used by dialogue agents to store information and protocols necessary for communication in verbal or written form.

Since knowledge base construction as a process is resource- and time-consuming, it is reasonable to try to automate at least some part of that task. This thesis proposes three semi-automatic methods to solve it by extracting information from Frequently Asked Questions' sets to produce keywords.

The main aim of this thesis was to develop software necessary to solve the problem of semi-automatic knowledge base construction, to test the developed software by applying it to numerous FAQ sets and analyzing the results. The software was developed using high-end programming languages such as Python and PHP.

Thesis also describes the underlying problems in the fields of artificial intelligence and knowledge base designing where the described task is situated in.

Lisa 1. Väljavõte eksperimendis kasutatud küsimuste-vastuste komplektidest

Mis kiirusest alates kiiruskaamera pildi teeb?

90 km/h alas on kaamerad seadistatud fikseerima kiiruspiirangu rikkumisi alates 97 km/h. Sellest kiirusest lahutatakse maha kaamerate mõõtemääramatus 4 km/h, mistõttu määratakse esimene hoiatustrahv sõidukiiruse ületamise eest 3 km/h. Iga lubatud sõidukiirust ületatud kilomeetri kohta tuleb tasuda 50 krooni, seega on minimaalseks hoiatustrahvi summaks 150 krooni. Hoiatustrahvi ülempiir on 3000 krooni – kui kiiruseületamisest tulenev trahv ületaks seda määra, suunatakse rikkumine üldmenetlusse. Üldmenetluse tulemusel tehtud trahviotsus on karistusõigusliku jõuga.

Kas iga kiiruskaamera tehtud pildi kohta saadetakse eraldi hoiatustrahv?

Kui sama sõiduki kiiruseületuse fikseerib järjestikku mitu kaamerat, tuleb sõiduki omanikul või vastutaval kasutajal tasuda iga fikseeritud kiiruseületuse eest. Sellisel juhul tuleb rikkujale postkasti ka iga fikseeritud kiiruseületuse kohta eraldi trahviteade.

Kas kiirabiautod, politsei ja tuletõrje hakkavad ka trahviteateid saama?

Kiiruskaamera fikseerib kiiruseületamise olenemata sellest, kas tegemist on tavasõiduki, politseisõiduki, kiirabiauto või tuletõrjega. Trahviteade saadetakse välja võrdsetel alustel kõigile, kes on kiirust ületanud. Kui tegemist oli alarmsõiduga ning sõidukil töötasid vilkurid, tuleb sõiduki vastutaval kasutajal täita hoiatustrahvi teatega kaasa tulnud vaidlustusvorm ning tõendada kiiruseületamise õiguspärane põhjus.

Mootorratastel on numbrimärk taga. Kuidas nende kiiruseületamisi fikseeritakse?

Praegu on kiiruskaamerad seadistatud mõõtma ainult lähenevate sõidukite kiirusi ning ei fikseeri endast kaugenevaid sõidukeid. Kuna mootorratastel on numbrimärk reeglina taga, siis mootorrataste kiiruseületamisi ei fikseerita. Vajadusel saab kaameraid seadistada fikseerima ka kaugenevaid sõidukeid, sh. mootorrattaid. Ära ei tasuks unustada ka seda, et lisaks kaameratele teevad tööd ka tavapatrullid, kes samuti mõõdavad liiklejate sõidukiirusi.

Kuidas hoiatustrahvi tasuda?

Hoiatustrahvi tasumiseks vajalikud andmed ja tingimused on kirjas trahviteatel. Hoiatustrahvi tasumiseks või selle vaidlustamiseks on aega 30 päeva. Hoiatustrahvi tasumata jätmisel suunatakse fikseeritud rikkumine täitemenetlusse.

Kas hoiatustrahvi saab vaidlustada?

Kui isik, kellele tuli trahviteade soovib trahvi vaidlustada, tuleb tal selleks täita teatega kaasas olev eeltäidetud vaidevorm. Hoiatustrahvi saab vaidlustada juhul, kui sõiduk või numbrimärk on varastatud (vajalik on pädeva asutuse kinnitus varguse kohta) või ilmneb mõni muu kiiruseületamise õigusvastasust välistav asjaolu.

Kaamera teeb pildi, kuid mulle tuleb vaid kirjalik teade. Kas ma saan väidetavalt enda rikkumisest tehtud fotot näha?

Trahviteatele koopiat fotost ei lisata. Mootorsõiduki eest vastutava isiku taotlusel saadetakse talle koopia fotost, mille abil tegu tuvastati. Taotluse esitamine foto saatmiseks ei peata trahviteate tasumiseks antud 30-päevast tähtaega. Foto saadetakse vastutava kasutaja poolt näidatud elektron- või tavaposti aadressile. Juhul, kui kiiruskaamera poolt salvestatud fotol on sõidukisalongis näha peale sõidukijuhi ka reisijaid, siis hägustatakse isikuandmete kaitsmise eesmärgil kõigi isikute kujutised peale sõidukijuhi.

Olen sõiduki omanik, kuid roolis oli keegi teine. Miks saadetakse trahv just mulle?

Vastavalt Liiklusseaduse § 11 peab sõidukiomanik, kui ta annab mootorsõiduki kasutada teisele isikule, säilitama mootorsõiduki kasutamise ajal ja teise isiku poolt mootorsõiduki kasutamise lõppemisest kuue kuu jooksul andmed sõiduki kasutaja kohta. Seega peab sõiduki omanik teadma, kes viimase kuue kuu jooksul tema omandis olevat sõidukit on kasutanud. Vastavalt Liiklusseadusele peab sõiduki omanik säilitama kasutaja ees- ja perekonnanime, aadressi, juhiloa numbri ja sünniaja või isikukoodi. Trahviteatega on kaasas eeltäidetud vaidlustusvorm, kuhu omanik saab kanda sõidukit tegelikult juhtinud isiku andmed ja vaidlustusvormi politseile tagasi saata.

Kas liisingus olevate autode juhid saavad ka hoiatustrahve?

Kui kiiruskaamera fikseerib liisingus oleva auto kiiruseületamise, saadetakse hoiatustrahv sõiduki vastutavale kasutajale. Rendiautode puhul saadetakse hoiatustrahvi teade rendifirmale käitub edasi juba rendilepingus kokkulepitud korras.

Kust saada kiiruskaamerateaga seonduvat lisainfot?

Politsei- ja Piirivalveameti kodulehel on lahti seletatud mitmed kiiruskaameraid puudutavad teemad ning püütud anda vastuseid korduma kippuvatele küsimustele. See info täieneb vastavalt vajadusele pidevalt. Samuti saab kiiruskaamerateaga seotud probleemide ja küsimustega pöörduda Politsei- ja Piirivalveameti infotelefoni poole numbril 612 3000.

Mis on koolitusosak?

Koolitusosak on sihtotstarbeline toetus töölase täiendkoolituse ostmiseks, mille eesmärgiks on tõsta mikro- ja väikeettevõtete konkurentsivõimet läbi parema ligipääsu koolitusteenusele.

Kellelt saab teenust osta?

Koolitusosaku abil saavad FIEd, mikro- ja väikeettevõtted hankida koolitusteenust riiklikelt koolitusasutustelt ja koolitusorganisatsioonidelt, kes on kantud teenusepakkujate nimekirja.

Kui suur on toetus?

Toetuse maksimaalne suurus on 15 000 krooni toetusesaaja kohta ning taotleda saab ainult üks kord 12 kuu jooksul.

Lisa 2. Väljavõtte esimese katse käigus saadud võtmesõnade komplektidest

Võtmesõnad: olema ; km/h ; kroon ; võra ; õiekroon

Vastus: 90 km/h alas on kaamerad seadistatud fikseerima kiiruspiirangu rikkumisi alates 97 km/h. Sellest kiirusest lahutatakse maha kaamerate mõõtemääramatus 4 km/h, mistõttu määratakse esimene hoiatustrahv sõidukiiruse ületamise eest 3 km/h. Iga lubatud sõidukiirust ületatud kilomeetri kohta tuleb tasuda 50 krooni, seega on minimaalseks hoiatustrahvi summaks 150 krooni. Hoiatustrahvi ülempiir on 3000 krooni – kui kiiruseületamisest tulenev trahv ületaks seda määra, suunatakse rikkumine üldmenetlusse. Üldmenetluse tulemusel tehtud trahviotsus on karistusõigusliku jõuga.

Võtmesõnad: kiiruseületus ; fikseerima ; tulema

Vastus: Kui sama sõiduki kiiruseületuse fikseerib järjestikku mitu kaamerat, tuleb sõiduki omanikul või vastutaval kasutajal tasuda iga fikseeritud kiiruseületuse eest. Sellisel juhul tuleb rikkujale postkasti ka iga fikseeritud kiiruseületuse kohta eraldi trahviteade.

Võtmesõnad: olema ; tulema ; tegemine ; sooritamine ; teostamine

Vastus: Kiiruskaamera fikseerib kiiruseületamise olenemata sellest, kas tegemist on tavasõiduki, politseisõiduki, kiirabiauto või tuletõrjega. Trahviteade saadetakse välja võrdsetel alustel kõigile, kes on kiirust ületanud. Kui tegemist oli alarmsõiduga ning sõidukil töötasid vilkurid, tuleb sõiduki vastutaval kasutajal täita hoiatustrahvi teatega kaasa tulnud vaidlustusvorm ning tõendada kiiruseületamise õiguspärane põhjus.

Võtmesõnad: sõiduk ; mootorratas ; ka ; veok ; sõiduvahend

Vastus: Praegu on kiiruskaamerad seadistatud mõõtma ainult lähenevate sõidukite kiirusi ning ei fikseeri endast kaugenevaid sõidukeid. Kuna mootorrattastel on numbrimärk reeglina taga, siis mootorrattaste kiiruseületamisi ei fikseerita. Vajadusel saab kaameraid seadistada fikseerima ka kaugenevaid sõidukeid, sh. mootorrattaid. Ära ei tasuks unustada ka seda, et lisaks kaameratele teevad tööd ka tavapatrullid, kes samuti mõõdavad liiklejate sõidukiirusi.

Võtmesõnad: hoiatustrahv ; tasumine ; olema ; maksmine ; makse

Vastus: Hoiatustrahvi tasumiseks vajalikud andmed ja tingimused on kirjas trahviteatel. Hoiatustrahvi tasumiseks või selle vaidlustamiseks on aega 30 päeva. Hoiatustrahvi tasumata jätmisel suunatakse fikseeritud rikkumine täitemenetlusse.

Võtmesõnad: või ; vaidlustama ; tulema

Vastus: Kui isik, kellele tuli trahviteade soovib trahvi vaidlustada, tuleb tal selleks täita teatega kaasas olev eeltäidetud vaidevorm. Hoiatustrahvi saab vaidlustada juhul, kui sõiduk või numbrimärk on varastatud (vajalik on pädeva asutuse kinnitus varguse kohta) või ilmneb mõni muu kiiruseületamise õigusvastasust välistav asjaolu.

Võtmesõnad: foto ; vastutav ; trahviteade ; ülesvõte ; päevapilt

Vastus: Trahviteatele koopiat fotost ei lisata. Mootorsõiduki eest vastutava isiku taotlusel saadetakse talle koopia fotost, mille abil tegu tuvastati. Taotluse esitamine foto saatmiseks ei peata trahviteate tasumiseks antud 30-päevast tähtaega. Foto saadetakse vastutava kasutaja poolt näidatud elektron- või tavaposti aadressile. Juhul, kui kiiruskaamera poolt salvestatud fotol on sõidukisalongis näha peale sõidukijuhil ka reisijaid, siis hāgustatakse isikandmete kaitsmise eesmärgil kõigi isikute kujutised peale sõidukijuhil.

Võtmesõnad: sõiduk ; ja ; pidama ; veok ; sõiduvahend

Vastus: Vastavalt Liiklusseaduse § 11 peab sõidukiomanik, kui ta annab mootorsõiduki kasutada teisele isikule, säilitama mootorsõiduki kasutamise ajal ja teise isiku poolt mootorsõiduki kasutamise lõppemisest kuue kuu jooksul andmed sõiduki kasutaja kohta. Seega peab sõiduki omanik teadma, kes viimase kuue kuu jooksul tema omandis olevat sõidukit on kasutanud. Vastavalt Liiklusseadusele peab sõiduki omanik säilitama kasutaja ees- ja perekonnanime, aadressi, juhiloa numbri ja sünniaja või isikukoodi. Trahviteatega on kaasas eeltāidetud vaidlustusvorm, kuhu omanik saab kanda sõidukit tegelikult juhtinud isiku andmed ja vaidlustusvormi politseile tagasi saata.

Võtmesõnad: saatma ; hoiatustrahv ; vastutav

Vastus: Kui kiiruskaamera fikseerib liisingus oleva auto kiiruseületamise, saadetakse hoiatustrahv sõiduki vastutavale kasutajale. Rendiautode puhul saadetakse hoiatustrahvi teade rendifirmale kätub edasi juba rendilepingus kokkulepitud korras.

Võtmesõnad: ja ; politsei ; piirivalveamet ; sandarmeeria ; korravalve

Vastus: Politsei- ja Piirivalveameti kodulehel on lahti seletatud mitmed kiiruskaameraid puudutavad teemad ning pūütud anda vastuseid korduma kippuvatele küsimustele. See info täieneb vastavalt vajadusele pidevalt. Samuti saab kiiruskaameratega seotud probleemide ja küsimustega pūrduda Politsei- ja Piirivalveameti infotelefoni poole numbril 612 3000.

Võtmesõnad: olema ; väikeettevõte ; tööalane

Vastus: Koolitusosak on sihtotstarbeline toetus tööalase täiendkoolituse ostmiseks, mille eesmärgiks on tõsta mikro- ja väikeettevõtete konkurentsivõimet läbi parema ligipāasu koolitusteenusele.

Võtmesõnad: ja ; väikeettevõte ; teenusepakkuja

Vastus: Koolitusosaku abil saavad FIED, mikro- ja väikeettevõtted hankida koolitusteenust riiklikelt koolitusasutustelt ja koolitusorganisatsioonidelt, kes on kantud teenusepakkujate nimekirja.

Võtmesõnad: üks ; toetusesaaja ; toetus ; tugi ; kaasabi

Vastus: Toetuse maksimaalne suurus on 15 000 krooni toetusesaaja kohta ning taotleda saab ainult üks kord 12 kuu jooksul.

Lisa 3. Väljavõtte teise katse käigus saadud võtmesõnade komplektidest

Võtmesõnad: tegema ; kiirus ; tempo ; ajaline omadus

Vastus: 90 km/h alas on kaamerad seadistatud fikseerima kiiruspiirangu rikkumisi alates 97 km/h. Sellest kiirusest lahutatakse maha kaamerate mõõtemääramatus 4 km/h, mistõttu määratakse esimene hoiatustrahv sõidukiiruse ületamise eest 3 km/h. Iga lubatud sõidukiirust ületatud kilomeetri kohta tuleb tasuda 50 krooni, seega on minimaalseks hoiatustrahvi summaks 150 krooni. Hoiatustrahvi ülempiir on 3000 krooni – kui kiiruseületamisest tulenev trahv ületaks seda määra, suunatakse rikkumine üldmenetlusse. Üldmenetluse tulemusel tehtud trahviotsus on karistusõigusliku jõuga.

Võtmesõnad: kohta ; eraldi ; iga ; vanus ; omadus

Vastus: Kui sama sõiduki kiiruseületuse fikseerib järjestikku mitu kaamerat, tuleb sõiduki omanikul või vastutaval kasutajal tasuda iga fikseeritud kiiruseületuse eest. Sellisel juhul tuleb rikkujale postkasti ka iga fikseeritud kiiruseületuse kohta eraldi trahviteade.

Võtmesõnad: kas ; tuletõrje ; kiirabiauto ; auto

Vastus: Kiiruskaamera fikseerib kiiruseületamise olenemata sellest, kas tegemist on tavasõiduki, politseisõiduki, kiirabiauto või tuletõrjega. Trahviteade saadetakse välja võrdsetel alustel kõigile, kes on kiirust ületanud. Kui tegemist oli alarmsõiduga ning sõidukil töötasid vilkurid, tuleb sõiduki vastutaval kasutajal täita hoiatustrahvi teatega kaasa tulnud vaidlustusvorm ning tõendada kiiruseületamise õiguspärane põhjus.

Võtmesõnad: kiiruseületamine ; mootorratas ; fikseerima ; taga ; olema

Vastus: Praegu on kiiruskaamerad seadistatud mõõtma ainult lähenevate sõidukite kiirusi ning ei fikseeri endast kaugenevaid sõidukeid. Kuna mootorrattastel on numbrimärk reeglina taga, siis mootorrattaste kiiruseületamisi ei fikseerita. Vajadusel saab kaameraid seadistada fikseerima ka kaugenevaid sõidukeid, sh. mootorrattaid. Ära ei tasuks unustada ka seda, et lisaks kaameratele teevad tööd ka tavapatrullid, kes samuti mõõdavad liiklejate sõidukiirusi.

Võtmesõnad: hoiatustrahv

Vastus: Hoiatustrahvi tasumiseks vajalikud andmed ja tingimused on kirjas trahviteatel. Hoiatustrahvi tasumiseks või selle vaidlustamiseks on aega 30 päeva. Hoiatustrahvi tasumata jätmisel suunatakse fikseeritud rikkumine täitemenetlusse.

Võtmesõnad: hoiatustrahv ; vaidlustama ; saama

Vastus: Kui isik, kellele tuli trahviteade soovib trahvi vaidlustada, tuleb tal selleks täita teatega kaasas olev eeltäidetud vaidevorm. Hoiatustrahvi saab vaidlustada juhul, kui sõiduk või numbrimärk on varastatud (vajalik on pädeva asutuse kinnitus varguse kohta) või ilmneb mõni muu kiiruseületamise õigusvastasust välistav asjaolu.

Võtmesõnad: nägema ; foto ; ülesvõte ; päevapilt ; pilt

Vastus: Trahviteatele koopiat fotost ei lisata. Mootorsõiduki eest vastutava isiku taotlusel saadetakse talle koopia fotost, mille abil tegu tuvastati. Taotluse esitamine foto saatmiseks ei peata trahviteate tasumiseks antud 30-päevast tähtaega. Foto saadetakse vastutava kasutaja poolt näidatud elektron- või tavaposti aadressile. Juhul, kui kiiruskaamera poolt salvestatud fotol on sõidukisalongis näha peale sõidukijuhil ka reisijaid, siis hädustatakse isikuandmete kaitsmise eesmärgil kõigi isikute kujutised peale sõidukijuhil.

Võtmesõnad: teine ; olema ; saatma ; sõiduk ; omanik

Vastus: Vastavalt Liiklusseaduse § 11 peab sõidukiomanik, kui ta annab mootorsõiduki kasutada teisele isikule, säilitama mootorsõiduki kasutamise ajal ja teise isiku poolt mootorsõiduki kasutamise lõppemisest kuue kuu jooksul andmed sõiduki kasutaja kohta. Seega peab sõiduki omanik teadma, kes viimase kuue kuu jooksul tema omandis olevat sõidukit on kasutanud. Vastavalt Liiklusseadusele peab sõiduki omanik säilitama kasutaja ees- ja perekonnanime, aadressi, juhiloa numbri ja sünniaja või isikukoodi. Trahviteatega on kaasas eeläidetud vaidlustusvorm, kuhu omanik saab kanda sõidukit tegelikult juhtinud isiku andmed ja vaidlustusvormi politseile tagasi saata.

Võtmesõnad: olev ; hoiatustrahv ; liising ; auto ; essiiv

Vastus: Kui kiiruskaamera fikseerib liisingus oleva auto kiiruseületamise, saadetakse hoiatustrahv sõiduki vastutavale kasutajale. Rendiautode puhul saadetakse hoiatustrahvi teade rendifirmale käitub edasi juba rendilepingus kokkulepitud korras.

Võtmesõnad: saama ; kiiruskaamera

Vastus: Politsei- ja Piirivalveameti kodulehel on lahti seletatud mitmed kiiruskaameraid puudutavad teemad ning püütud anda vastuseid korduma kippuvatele küsimustele. See info täieneb vastavalt vajadusele pidevalt. Samuti saab kiiruskaameratega seotud probleemide ja küsimustega pöörduda Politsei- ja Piirivalveameti infotelefoni poole numbril 612 3000.

Võtmesõnad: olema ; mis ; koolitusosak

Vastus: Koolitusosak on sihtotstarbeline toetus töölase täiendkoolituse ostmiseks, mille eesmärgiks on tõsta mikro- ja väikeettevõtete konkurentsivõimet läbi parema ligipääsu koolitusteenusele.

Võtmesõnad: saama

Vastus: Koolitusosaku abil saavad FIED, mikro- ja väikeettevõtted hankida koolitusteenust riiklikelt koolitusasutustelt ja koolitusorganisatsioonidelt, kes on kantud teenusepakkujate nimekirja.

Võtmesõnad: olema ; toetus ; tugi ; kaasabi ; kaasaaitamine

Vastus: Toetuse maksimaalne suurus on 15 000 krooni toetusesaaja kohta ning taotleda saab ainult üks kord 12 kuu jooksul.

Lisa 4. Väljavõtte kolmanda katse käigus saadud võtmesõnade komplektidest

Võtmesõnad: tegema ; kiirus ; tempo ; ajaline omadus

Vastus: 90 km/h alas on kaamerad seadistatud fikseerima kiiruspiirangu rikkumisi alates 97 km/h. Sellest kiirusest lahutatakse maha kaamerate mõõtemääramatus 4 km/h, mistõttu määratakse esimene hoiustrahv sõidukiiruse ületamise eest 3 km/h. Iga lubatud sõidukiirust ületatud kilomeetri kohta tuleb tasuda 50 krooni, seega on minimaalseks hoiustrahvi summaks 150 krooni. Hoiustrahvi ülempiir on 3000 krooni – kui kiiruseületamisest tulenev trahv ületaks seda määra, suunatakse rikkumine üldmenetlusse. Üldmenetluse tulemusel tehtud trahviotsus on karistusõigusliku jõuga.

Võtmesõnad: tuletõrje ; kiirabiauto ; auto

Vastus: Kiiruskaamera fikseerib kiiruseületamise olenemata sellest, kas tegemist on tavasõiduki, politseisõiduki, kiirabiauto või tuletõrjega. Trahviteade saadetakse välja võrdsetel alustel kõigile, kes on kiirust ületanud. Kui tegemist oli alarmsõiduga ning sõidukil töötasid vilkurid, tuleb sõiduki vastutaval kasutajal täita hoiustrahvi teatega kaasa tulnud vaidlustusvorm ning tõendada kiiruseületamise õiguspärane põhjus.

Võtmesõnad: numbrimärk ; fikseerima ; mootorratas ; kiiruseületamine

Vastus: Praegu on kiiruskaamerad seadistatud mõõtma ainult lähenevate sõidukite kiirusi ning ei fikseeri endast kaugenevaid sõidukeid. Kuna mootorratasatel on numbrimärk reeglina taga, siis mootorrataste kiiruseületamisi ei fikseerita. Vajadusel saab kaameraid seadistada fikseerima ka kaugenevaid sõidukeid, sh. mootorrataid. Ära ei tasuks unustada ka seda, et lisaks kaameratele teevad tööd ka tavapatrullid, kes samuti mõõdavad liiklejate sõidukiirusi.

Võtmesõnad: hoiustrahv

Vastus: Hoiustrahvi tasumiseks vajalikud andmed ja tingimused on kirjas trahviteatel. Hoiustrahvi tasumiseks või selle vaidlustamiseks on aega 30 päeva. Hoiustrahvi tasumata jätmisel suunatakse fikseeritud rikkumine täitemenetlusse.

Võtmesõnad: hoiustrahv ; vaidlustama ; saama

Vastus: Kui isik, kellele tuli trahviteade soovib trahvi vaidlustada, tuleb tal selleks täita teatega kaasas olev eeltäidetud vaidevorm. Hoiustrahvi saab vaidlustada juhul, kui sõiduk või numbrimärk on varastatud (vajalik on pädeva asutuse kinnitus varguse kohta) või ilmneb mõni muu kiiruseületamise õigusvastasust välistav asjaolu.

Võtmesõnad: nägema ; foto ; ülesvõtte ; päevapilt ; pilt

Vastus: Trahviteatele koopiat fotost ei lisata. Mootorsõiduki eest vastutava isiku taotlusel saadetakse talle koopia fotost, mille abil tegu tuvastati. Taotluse esitamine foto saatmiseks ei peata trahviteate tasumiseks antud 30-päevast tähtaega. Foto saadetakse vastutava kasutaja poolt näidatud elektron- või tavaposti aadressile. Juhul, kui kiiruskaamera poolt salvestatud fotol on sõidukisalongis näha peale

sõidukijuhi ka reisijaid, siis hāgustatakse isikuandmete kaitsmise eesmärgil kõigi isikute kujutised peale sõidukijuhi.

Võtmesõnad: saatma ; sõiduk ; omanik ; veok ; sõiduvahend

Vastus: Vastavalt Liiklusseaduse § 11 peab sõidukiomanik, kui ta annab mootorsõiduki kasutada teisele isikule, säilitama mootorsõiduki kasutamise ajal ja teise isiku poolt mootorsõiduki kasutamise lõppemisest kuue kuu jooksul andmed sõiduki kasutaja kohta. Seega peab sõiduki omanik teadma, kes viimase kuue kuu jooksul tema omandis olevat sõidukit on kasutanud. Vastavalt Liiklusseadusele peab sõiduki omanik säilitama kasutaja ees- ja perekonnanime, aadressi, juhiloa numbrit ja sünniaja või isikukoodi. Trahviteatega on kaasas eeltäidetud vaidlustusvorm, kuhu omanik saab kanda sõidukit tegelikult juhtinud isiku andmed ja vaidlustusvormi politseile tagasi saata.

Võtmesõnad: olev ; hoiustrahv ; liising ; auto ; essiiv

Vastus: Kui kiiruskaamera fikseerib liisingus oleva auto kiiruseületamise, saadetakse hoiustrahv sõiduki vastutavale kasutajale. Rendiautode puhul saadetakse hoiustrahvi teade rendifirmale käitub edasi juba rendilepingus kokkulepitud korras.

Võtmesõnad: saada ; kiiruskaamera

Vastus: Politsei- ja Piirivalveameti kodulehel on lahti seletatud mitmed kiiruskaameraid puudutavad teemad ning pūūitud anda vastuseid korduma kippuvatele küsimustele. See info täieneb vastavalt vajadusele pidevalt. Samuti saab kiiruskaameratega seotud probleemide ja küsimustega pūūrduda Politsei- ja Piirivalveameti infotelefoni poole numbril 612 3000.

Võtmesõnad: koolitusosak

Vastus: Koolitusosak on sihtotstarbeline toetus tööalase täiendkoolituse ostmiseks, mille eesmärgiks on tõsta mikro- ja väikeettevõtete konkurentsivõimet läbi parema ligipääsu koolitusteenusele.

Võtmesõnad: saada

Vastus: Koolitusosaku abil saavad FIEd, mikro- ja väikeettevõtted hankida koolitusteenust riiklikelt koolitusasutustelt ja koolitusorganisatsioonidelt, kes on kantud teenusepakujate nimekirja.

Võtmesõnad: toetus ; tugi ; kaasabi ; kaasaaitamine ; abistamine

Vastus: Toetuse maksimaalne suurus on 15 000 krooni toetusesaaja kohta ning taotlelda saab ainult üks kord 12 kuu jooksul.