Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6

Keel ja arvuti

Tartu 2006

Keel ja arvuti

Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6

Keel ja arvuti

Toimetajad Mare Koit, Renate Pajusalu, Haldur Õim

Tartu 2006

Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6
Keel ja arvuti
Toimetajad Mare Koit, Renate Pajusalu, Haldur Õim
Tartu 2006

Raamatu väljaandmist on toetanud ETF grant 5534

ISSN 1406–619X
ISBN-10 9949–11–309–1
ISBN-13 978–9949–11–309–5

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr 734/2005

Sisukord

Keeled ja arvuti: muid seoseid

Lisad

Saateks

Aastal 2000 ilmus TÜ üldkeeleteaduse õppetooli toimetiste sarja esimese numbrina kogumik "Arvutilingvistikalt inimesele" (toimetaja Tiit Hennoste). Siinset kogumikku võib käsitada selle põhimõttelise jätkuna: selle peamine taotlus on anda läbilõige arvuti ja inimese – keelekasutaja, keeleuurija – kokkupuutealast tänases Eestis. Teise eesmärgina on peetud silmas seda, et artiklid peaksid kirjeldama tehtavaid töid ja uuritavaid probleeme nii, et need oleksid arusaadavad ka väiksema professionaalse ettevalmistusega inimestele ja kogumik oleks kasutatav õppematerjalina ülikooli(de) vastavatel kursustel.

Siinsel kogumikul on veel üks lisaeesmärk. Eesti arvutilingvistid ja keeletehnoloogid on arvukalt avaldanud oma töö tulemusi võõrkeeltes (eelkõige inglise keeles), eesti keeles (nt Keeles ja Kirjanduses), aga suhteliselt vähe ja mittesüstemaatiliselt. Käesolev kogumik püüab seda lünka täita.

Arvutilingvistika ja keeletehnoloogia arengut mõjutanud sündmused

Siinse ja eelmise kogumiku ilmumise vahele jääb mitmeid eesti arvutilingvistika arengule põhimõttelise tähtsusega sündmusi, mida taustana on mõttekas lühidalt kirjeldada.

Kõige olulisem dokument, millele järgnev areng on tuginenud, on "Eesti keele arendamise strateegia 2004–2010", mille töötas välja Eesti keelenõukogu ja kinnitas valitsus 5. augustil 2004. See ilmus väikese brošüürina 2004. aasta sügisel (Eesti keele arendamise strateegia 2004–2010. Haridus- ja Teadusministeerium. Eesti keelenõukogu. Tartu 2004). Selle dokumendi tähtsus on olnud suur, seejuures ka arvutilingvistikale ja keeletehnoloogiale. Muuhulgas sisaldab see peatüki "Eesti keele keeletehnoloogiline tugi", mis näeb ette rea tegevusi eesti keele funktsioneerimisvõime tagamiseks rahvuskeelena ka tuleviku infoühiskonnas.

Keelestrateegia ühe ettevalmistava lisana ilmus raamat "Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Eesti Keele Sihtasutus, Tallinn 2003" (autorid Kadri Muischnek, Heili Orav, Heiki-Jaan Kaalep ja Haldur Õim). Selle trükise lõpust võib leida ühe kõige

täielikuma eesti keele alase arvutilingvistika ja keeletehnoloogia bibliograafia tollase seisuga.

Teiseks sammuks, mis (osaliselt) tulenes eelmainitust, oli riikliku programmi "Eesti keele keeletehnoloogiline tugi 2006–2010" väljatöötamine. Keeletehnoloogia (mida ei ole eriti üritatud eristada arvutilingvistikast) oli omaette moodulina sees juba varem käivitatud riiklikus programmis "Eesti keel ja rahvuslik mälu", kuid tööde mahukuse ja suhteliselt suure maksumuse tõttu leiti olevat otstarbekas käivitada omaette riiklik programm. Selliselt ongi alates 2006. aastast programmil oma finantseering ja omad projektide valiku põhimõtted.

Ja kolmandaks on oluliselt muutunud õppimis- ja õpetamispool. Tartu Ülikooli matemaatika-informaatikateaduskonnas asutati 2001. aastal keeletehnoloogia professuur ja õppetool (professor Mare Koit). Arvutilingvistikat/keeletehnoloogiat on nüüd võimalik õppida Tartu Ülikoolis nii bakalaureuse-, magistri- kui doktoriõppes.

2005. aastal hakkas Tartu Ülikooli juures tööle doktorikool, mille ametlik nimetus on Keeleteaduse ja -tehnoloogia doktorikool. Kooli rahastatakse suures osas Euroopa Liidu programmide raames, nii nagu see on ka teiste analoogiliste doktorikoolide puhul. Ühelt poolt näitab kooli nimetus, et meil on keeleteadust ja keeletehnoloogiat/arvutilingvistikat soovitud õpetada (ja uurida) koos. Teisest küljest märgitagu, et siin ei ole kõne all ainult eesti keel: ühe poole moodustab (üld)keeleteadus ja teise poole keeletehnoloogia ja arvutilingvistika. Kui suure osa neist uurimustest moodustavad eesti keele alased uurimused, sõltub tegijatest. Aga praegu ei ole mingit kahtlust, et doktorikool on Eesti keeleteadusele ja arvutilingvistikale ning keeletehnoloogiale positiivset mõju avaldanud. Suurem osa siinse kogumiku autoreid on näiteks doktorandid ja ühtlasi doktorikooli teadurid.

Lisaks sellele tasub märkida, et 28. oktoobril 2004 allkirjastati Tartus Põhjamaade keeletehnoloogia tippspetsialistide külaskäigu ajal leping Põhjamaade Keeletehnoloogia Kraadiõppekooli (Nordic Graduate School of Language Technology ehk NGSLT[1]) ja Tartu Ülikooli vahel. Leping annab Tartu ülikooli arvutilingvistika ja keeletehnoloogia üliõpilastele õiguse osa võtta kraadiõppekooli kursus-

[1] http://www.ngslt.org/

test, kusjuures kraadiõppekool katab kõik kulud. See on ainulaadne võimalus saada osa eriala parimate spetsialistide loengutest. Seda võimalust on juba kasutanud mitmed meie üliõpilased. Siiani on enamus kursusi toimunud kas Rootsis, Soomes või Taanis, kuid väärib märkimist, et ühe sellise kursuse korraldas 2005.a veebruarikuus ka Tartu Ülikool (Sheffieldi ülikooli professori Yorick Wilksi kursus dialoogi automaattöötlusest Tartu Ülikoolis).

Eespoolkirjeldatu oli taustaks, et lugeja saaks kogumiku artikleid teatud konteksti paigutada.

Kogumiku alajaotustest ja artiklitest

Vahepealse aja jooksul on selgemalt välja joonistunud ka eesti arvutilingvistika uurimistööde kesksed suunad. Seda on mõeldud kajastama kogumiku (sisukorra) alajaotused: eesti keele arvutigrammatika, eesti keele arvutisemantika, suuline eesti keel ja dialoog arvutiga (ehk pragmaatika), lisaks mõned artiklid, mis nende alajaotuste sisse ei mahu, – aga see ei vähenda nende olulisust. See "muu" on lihtsalt osutus, et meie arvutilingvistika ei kata süstemaatiliselt kogu valdkonda, mis üldiste arusaamade kohaselt siia kuulub.

Arvutigrammatika alla on paigutatud (eesti keele) morfoloogia ja süntaksi alased tööd. Heli Uibo jätkab eesti keele morfoloogia modelleerimise uute võimaluste katsetamisega. Heli Uibo ja Helen Nigoli kirjutis puude pankadest osutab uuele olulisele suunale arvutisüntaksis: omapärase nimega termin *puude pank* viitab süntaktiliselt märgendatud korpusele, kus iga lause süntaktiline märgend on formaalses mõttes puu (ehk hargmik). Meenutame, et varem tegelesid eesti keele automaatse süntaksianalüüsi arendajad lineaarse (mitte hierarhilise) nn kitsenduste grammatika esitusviisiga.

Kitsenduste grammatika esitusviisi on katsetanud ja tulemusi kirjeldanud Kaili Müürisep, Helen Nigol ja Heli Uibo eesti kõnekeelele rakendatult. Peamine põhjus, miks siin puude panka (esialgu) ei üritata teha, on see, et kõnekeele morfoloogia ja süntaks erinevad oluliselt kirjakeele omadest ja neid ei tunta eesti keele osas piisavalt, et vajalikke märgendustöid automaatselt teha.

Nende vahele jääb Kadri Muischneki artikkel eesti keele verbikesksete püsiühendite nominaalsetest komponentidest. Mitmesõnalised väljendid nagu *torkas silma* või *ripub juuksekarva küljes* on

automaatse süntaksianalüüsi jaoks omaette probleem, sest neid tuleb käsitleda kui terviküksusi.

Arvutisemantika katab märksa mitmekülgsemat valdkonda kui arvutigrammatika, ehkki artikleid ei ole palju. Kogumikus on esindatud eelkõige sõnasemantika-alased tööd.

Heili Orav ja Kadri Vider küsivad, missugust leksikoni – missugust semantilist infot sõnade kohta – vajaks süsteem, mis püüab keelest (tekstist) aru saada.

Kadri Kerneri, Kadri Videri ja Neeme Kahuski artikkel sõnatähenduste automaatsest ühestamisest käsitleb "vana" teemat selles mõttes, et tähenduste ühestamise probleem (st meetodite leidmine otsustamaks selle üle, mis tähenduses esineb mitmetähenduslik sõna nagu *pank* mingis konkreetses kasutuses) oli samal viisil aktuaalne ka viis aastat tagasi. Praegu on selle aktuaalsus pigem kasvanud, aga niisuguseid arenguid, nagu on nt automaatses süntaktilises analüüsis, ei ole võimalik täheldada.

Semantika ja pragmaatika on üldse keelevaldkonnad, mille arvutirakendused alles ootavad oma läbimurret. Inimese ja arvuti loomuliku suhtlemise tagamiseks on nad samas võtmevaldkonnad. Eelöeldut illustreerivad ühelt poolt Liisi Piitsi artikkel sõnade *mees* ja *naine* kasutamise semantilisest taustast ja kasutustest ning teiselt poolt Kaili Müürisepa artikkel sisukokkuvõtteid tegevast programmist EstSum – need lähenevad keele semantikale üpris erinevatest aspektidest.

Rubriiki "Suuline eesti keel ja dialoog arvutiga" on koondatud kaht tüüpi artiklid, mis oma sisu poolest on aga vahetult seotud. Erinevus on eesmärkides, seob aga käsitletav keelematerjal: see on suuline (eesti) keel. Suuline keelekasutus erineb kirjutatud keelest (mis on eelmise kahe alajaotuse artiklites kirjeldatud uurimuste objektiks) paljude parameetrite osas. Erinevused on nii morfoloogias, süntaksis, semantikas kui pragmaatikas. Olulisimad erinevused tulenevad sellest, et inimeste omavahelise suhtlemise algne ja praegugi valdav vorm on kõne. Kiri ja kirjalikud tekstid on ses mõttes hiline areng. Suulises suhtluses mängivad olulist rolli nähtused, mida kirjakeele reegleid järgiv keelekasutus tüüpiliselt ei kajasta ning tihti ei võimaldagi kajastada: intonatsioon, pausid, aga ka poolelijäetud laused, kordused, üksteisele vahele- ja pealerääkimised jne. Ülevaate neist

annab Tiit Hennoste sissejuhatav artikkel. Arvutilingvistika (keeletehnoloogia) seisukohalt on siin esindatud kaks erinevat temaatikat.

Ühes käsitletakse suulise dialoogi keelekasutuse eripärasusi ja nende fikseerimist-märgendamist arvutikorpuses. Siia kuuluvad Andriela Rääbise ja Riina Kasterpalu artiklid institutsionaalsete dialoogide algusrituaalidest, Krista Strandsoni artikkel kahepeale tehtud parandustest dialoogides ja Liina Eskori artikkel suhtlusstrateegiatest.

Teise rühma moodustavad artiklid, kus arvutil modelleeritakse inimestevahelist loomulikku suhtlust ning eesmärk on jõuda selleni, et inimene võiks suhelda arvutiga samal viisil nagu teise inimesega, s.o kõne abil: näiteks küsida mingeid andmeid, koostada reisiplaani jne (niisugune dialoog ei pruugi sugugi osutuda lihtsaks küsimus–vastus-paariks, sest vastaja – antud juhul arvuti – võib esitada täpsustavaid vastuküsimusi vms).

Olga Gerassimenko ja Maret Valdisoo võrdlevad inimese ja arvuti kui infoandja strateegiaid, kui info küsijaks on inimene. Margus Treumuth ja Tarmo Truu tegelevad küsimusega, millise info abil võiks arvuti tuvastada kuupäevi või eristada samamoodi vormistatud, kuid erinevat vastust ootavaid kas-küsimusi. Mark Fišeli ja Taavet Kikase artikkel aga proovib lahendada laiemalt dialoogiaktide automaatse tuvastamise probleemi.

Niisuguse suhtluse modelleerimise aluseks on eelmise rühma uurimistööde tulemused. Näiteks peab arvuti ära tundma suhtluspartneri (inimese) pöördumises sisalduvad kõneaktid (küsimus, korraldus, ettepanek jne), oskama siduda iga konkreetse pöördumise dialoogi eelnevate osadega, lõpuks ka oskama modelleerida inimpartneri mõttekäike, mida too eksplitsiitselt ei väljenda (sest nii on ta harjunud suhtlema teiste inimestega). See problemaatika moodustabki arvutipragmaatika tuuma.

Viimases alajaotuses on, nagu öeldud, artiklid, mis ei paigutu loomulikul viisil eelmiste alajaotuste alla.

Andres Loopmanni, Kati Seina ja Ülle Viksi artikkel kirjeldab Eesti Keele Instituudis arendatavat sõnastike haldussüsteemi ja sellisena võib selle liigitada arvutileksikoloogia/-leksikograafia alla.

Anni Oja artikkel käsitleb eripärasusi, mis iseloomustavad internetis – jututubades, kommentaarides jm – kasutatavat eesti keelt. Ehkki omal eripärasel viisil, demonstreerib internet mõju, mida võib

keelele ja keelekasutusele osutada põhimõtteliselt uus, aga massiliselt kasutusele tulev suhtlusmeedium. Spetsialistid hakkasid "keeletehnoloogilisest revolutsioonist" rääkima juba möödunud sajandi 90. aastate alguses, kui osutati, et arvuti lülitamine keelekasutusse ja suhtlusprotsessi võib avaldada keelele sama suurt mõju kui omal ajal kirja leiutamine ja massiline kasutuselevõtt.

Lõpuks, Kaarel Kaljuranna artikkel osutab – küll inglise keele najal – teisele arvutite ulatusliku kasutamise mõjule: kuna niipea ei ole loota, et arvuti saaks hakkama täiesti vaba keelekasutusega (nt sisulises infootsingus, masintõlkes), siis on hakatud välja töötama piiratud keelekasutusvorme nende tekstide koostamiseks, millest on ette teada, et nendega hakatakse töötama arvuti vahendusel. Need piirangud on seatud täiesti teadlikult ja määravad, missugust sõnavara, missuguseid konstruktsioone jne tohib kasutada. Eesti keele osas ei ole selles suunas midagi tõsist tehtud, kuid nt inglise keele osas on tehtud ära märkimisväärne töö.

Lõpuks märgime, et kogumik ei kajasta Eestis tehtavat arvutilingvistika ja keeletehnoloogia alal tehtavat tööd täielikult. Nii ei kajastu siin kitsamas mõttes kõnetehnoloogia alal tehtav – kõnesüntees ja kõnetuvastus, millega tegeldakse eelkõige TTÜ Küberneetika Instituudis ja ka Eesti Keele Instituudis. Kuid kajastatud valdkondadest, eriti kui võtta kõrvale 2000. a. ilmunud kogumik, annab see päris tervikliku pildi.

Mare Koit
Renate Pajusalu
Haldur Õim

Eesti keele morfoloogia modelleerimisest lõplike muundurite abil

Heli Uibo
Tartu Ülikool

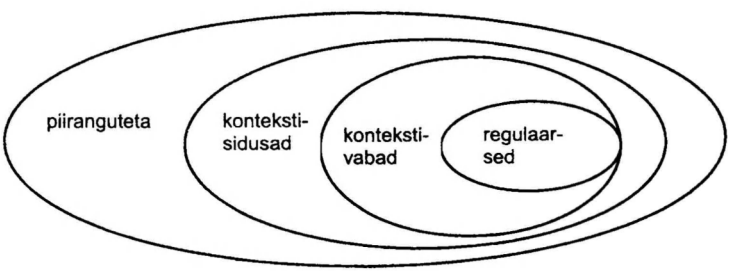
1. Sissejuhatus

Loomulike keelte morfoloogia modelleerimisel on maailmas viimase 15–20 aasta jooksul vaieldamatult edukaim olnud lõpliku olekute hulgaga seadmete – lõplike automaatide (*finite-state automata*) ja lõplike muundurite (*finite-state transducers*) põhine lähenemine. Formaalsete keelte teooriast on teada, et lõplike automaatidega on genereeritavad ja äratuntavad regulaarsed ehk 3. tüüpi keeled. N. Chomsky esitas ja tõestas oma legendaarses raamatus "Syntactic structures", mida peetakse arvutilingvistika ajaloos üheks tähelepanuväärsemaks teoseks, väite: "English is not a finite state language." (Chomsky 1957). Samuti oletas ta, et inglise keel ei ole kontekstivaba. Seega joonisel 1 esitatud diagrammil ei saa loomulikke keeli genereerivad grammatikad paikneda kahes väiksemas alamhulgas – regulaarsete ega kontekstivabade grammatikate hulgas.

Tõsi küll, Chomsky esitas oma väite süntaksi kohta. Väidet, et inglise keel ei ole regulaarne, saab tõestada (teoreetiliselt piiramatute) rekursiivsete protsessidega süntaksis:

a) kõrvallausete lisamine

I saw a dog, who chased a cat, who ate a rat, who ...



Joonis 1. Chomsky formaalsete grammatikate hierarhia

b) vabade lauselaiendite (põhiliselt määruste) lisamine

S → NP (AdvP)* VP (AdvP)*

Kummalgi juhul ei piisa lausete genereerimiseks reeglitest kujul A → xB ja A → x, mis on ainsad regulaarsetes grammatikates lubatavad reeglite kujud (A, B – mitteterminaalsed e süvaesituse sümbolid, x – terminaalne ehk pindesituse sümbol). Seetõttu on 1960. aastatest süntaksi modelleerimisel kasutatud võimsamaid formalisme, näiteks fraasistruktuuri- ja unifikatsioonigrammatikaid (GPSG, HPSG, LFG).

Kuid ka morfoloogias (generatiivses fonoloogias) kasutati esialgu samasuguse võimsusega grammatikaid – kontekstisidusaid ümberkirjutusreegleid (Chomsky, Halle 1968), mida tuli rakendada kindlas järjekorras, et teisendada abstraktne fonoloogiline esitus läbi mitmete vaheesituste pindesituseks (sõnavormiks).

Generatiivse fonoloogia reeglite üldkuju on

x → y / z _ w (loe: x asemele y kontekstis z_w),

kus x, y, z ja w on suvalise keerukusega tunnusstruktuurid.

Praktiliselt kasutatavate kontekstisidusate grammatikate kirjutamine osutus isegi palju uuritud keelte, nagu inglise keel jaoks väga raskeks ülesandeks, kuna keeruline on defineerida reeglite järjestust ning transformatsioonidel tekib palju vaheesitusi. Lisaks pole transformatsioonireeglid pööratavad – genereerivat grammatikat ei saa mingil standardsel viisil teisendada analüüsigrammatikaks.

Regulaarsed keeled, neid genereerivad regulaarsed grammatikad ja neid aktsepteerivad lõplikud automaadid "avastati uuesti" ja neid hakati keeletehnoloogias kasutama umbes 25 aastat tagasi.

Lõplik automaat (ingl *finite-state automaton)* on defineeritud kui viisik

A = (Σ , Q, q_1, F, d),

kus

1) Q = {q_1, ..., q_n} on **lõplik** olekute hulk;
2) Σ on sisendtähestik;
3) d: Q x Σ * -> 2^Q on olekuteisendusfunktsioon, s.o eeskiri, mis määrab, kuidas automaat iga konkreetse oleku ja sümbolipaari korral käitub (sellise eeskirja saab kirja panna näiteks olekuteisendustabelina);
4) q_1 on automaadi algolek;

5) $F \subseteq Q$ on lõppolekute hulk. Need on olekud, milles automaat oma töö lõpetab. Lõppolekuid võib olla mitu, kuid iga olek ei tarvitse olla lõppolek.

Lõpliku automaadi laiendust, mis lisaks sümbolite lugemisele saab neid ka muuta, nimetatakse lõplikuks muunduriks (ingl *finite-state transducer*).

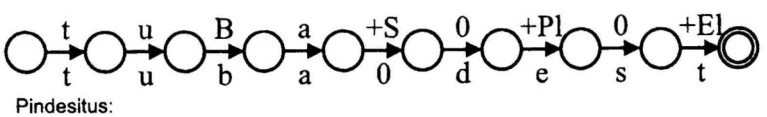
Eriti suur edu on lõplikel automaatidel põhinevatel meetoditel olnud loomuliku keele morfoloogia kirjeldamisel. Lõplike automaatide ja muundurite kasutatavus arvutimorfoloogias põhineb järgmistel tulemustel:

1) D. Johnson, 1972: Fonoloogilised ümberkirjutusreeglid ei ole sisuliselt kontekstisidusad, vaid neid saab kirjeldada lõplike muunduritena.

2) Schützenberger, 1961: Kui kaks lõplikku muundurit rakendada järjestikku, siis leidub üks lõplik muundur, mis on nende kahe lõpliku muunduri kompositsioon.

Kui kompositsiooni üldistada n muundurile, saame läbi ilma vaheesitusteta – süvaesitus teisendatakse pindesituseks üheainsa lõpliku muunduri abil. 1980. a avastasid selle tulemuse taas R. Kaplan ja M. Kay (Kaplan, Kay 1981).

Lõplikel muunduritel põhinev arvutimorfoloogia toetub veendumusele, et seos keele sõnavormide ja nende algvormide ehk lemmade vahel on kirjeldatav regulaarse relatsioonina. Regulaarse relatsiooni saab kirja panna regulaaravaldisena. Regulaaravaldise saab kompileerida lõplikuks muunduriks, mis realiseerib selle relatsiooni arvutuslikult. Muunduris seab mistahes tee algolekust lõppolekusse omavahel vastavusse mingi sõnavormi ehk pindesituse (*surface form*) ja tema lemma + morfoloogilise info (*lexical form*) ehk sõnastikuesituse (joonis 2).



Joonis 2. Muundur

Lõplikele muunduritele ja automaatidele on loomuliku keele töötlusel leitud mitmeid rakendusi (Beesley, Karttunen 2003).

• Sõnastik (sõnade loend) esitatud lõpliku automaadina – pakkimismeetod.

• Kakskeelne sõnastik kui lõplik muundur (nn leksikaalne muundur, ingl *lexical transducer*).

• Morfoloogiline muundur – leksikaalne muundur, mis võib olla kombineeritud reeglistik-muunduritega, näiteks Koskenniemi kahetasemelised reeglid, ingl *two-level rules* (Koskenniemi 1983) või Karttuneni asendusreeglid, ingl *replace rules* (Karttunen 1997). Morfoloogiline muundur on leksikaalse muunduri ja reeglistik-muunduri kompositsioon. Iga tee algolekust lõppolekusse kujutab sisuliselt sõnavormi tema algvormiks ehk lemmaks (millele võib vastavalt eesmärgile kuuluda morfoloogiline, süntaktiline, semantiline vms info).

Morfoloogiat võib pidada lõplike muundurite edukaimaks rakendusvaldkonnaks loomuliku keele töötlusel.

Morfoloogiline analüüs on morfoloogilises muunduris realiseeritav nii, et käiakse läbi leksikaalses muunduris leiduvad teed, mis algavad algolekust ja lõpevad lõppolekus. Kui leitakse tee, milles kaarte alumised märgendid annavad kokku analüüsitava sõna, siis väljastatakse vastavate kaarte ülemiste märgendite konkatenatsioon (lemma + morfoloogiline info). Kui ükski tee ei ole edukas (muundur ei lõpeta tööd lõppolekus), siis ei kuulu sõnavorm muunduri poolt kirjeldatud keelde.

Morfoloogiline süntees tähendab, et käiakse läbi leksikaalses muunduris leiduvad teed, mis algavad algolekust ja lõpevad lõppolekus. Kui leitakse tee, milles kaarte ülemised märgendid annavad kokku etteantud lemma + morfoloogilised märgendid, siis väljastatakse vastavate alumiste märgendite konkatenatsioon (sõnavorm). Kui ükski tee ei ole edukas (muundur ei aktsepteeri antud lemmat + morfoloogiliste märgendite järjendit), siis on kaks võimalust – kas see sõna puudub sõnastikust või ei ole etteantud grammatiline info korrektne.

Kuna lõplike automaatide ja muundurite arvutuskeerukus on väike, töötavad neil põhinevad keeletöötlusprogrammid kiiresti. Sellises keeletarkvaras on keelekirjeldus lahutatud reeglite ja sõnastike kompilaatoritest, mis võimaldab keelekirjeldust hõlpsasti modifitsee-

rida. Nagu eespool kirjeldatud morfoloogilise analüüsi ja sünteesi algoritmidest järeldada võib, on lõplikud muundurid juba idee poolest kahesuunalised, st ühte ja sama muundurit saab kasutada nii sama tasandi (sõna, lause jne) analüüsi kui ka sünteesi tegemiseks.

Lõplikel muunduritel põhinevad meetodid, sealhulgas kahetasemeline morfoloogiamudel, on osutunud edukaks sõnamuutmispõhimõtete poolest väga erinevate keelte (inglise, saksa, rootsi, prantsuse, hispaania, taani, norra, soome, vene, türgi, araabia, aimara, suahiili jt) morfoloogianähtuste formaliseerimisel, seega on alust arvata, et tegemist on tõepoolest universaalse mudeliga, mis võimaldab täielikult kirjeldada ka sedavõrd keerulise morfoloogiaga keele, nagu seda on eesti keel.

Lõplikel automaatidel põhineva keeletöötlusega on suuri kogemusi firma *Xerox* uurimiskeskustel Grenoble'is (Prantsusmaal) ja Palo Altos (Californias). Seal on loodud reeglite ja sõnastike kompilaatorid, morfoloogilisi analüsaatoreid ja süntesaatoreid, morfoloogilisi ühestajaid ja pindsüntaktilisi analüsaatoreid. Lisaks on nende baasil välja töötatud mitmeid rakendusi, nagu õigekirjakontrollija, infootsija, terminite ekstraheerija, autori- ja tõlkevahendid[1].

2. Morfoloogilised protsessid ja nende modelleerimine

Morfoloogias tuleb modelleerida järgmisi protsesse:

1) morfotaktika (kuidas kombineeritakse morfeemidest sõnavormid):
   a) ees- ja järelliidete lisamine, liitsõnamoodustus – konkatenatiivsed protsessid;
   b) reduplikatsioon (terve tüve kordus), infiksatsioon (liidete vahelelisamine), interdigitatsioon (täishäälikute lisamine konsonantide vahele, nt araabia keeles) – mittekonkatenatiivsed protsessid.
2) fonoloogilised/ortograafilised alternatsioonid
   - assimilatsioon (*hind* : *hinna*)
   - lisandumine (*jooksma* : *jooksev*)
   - kadu (*number* : *numbri*)
   - geminatsioon (*tuba* : *tuppa*)

[1] http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html

On tõestatud, et kõik loetletud morfoloogianähtused on kirjeldatavad regulaaravaldiste abil, sealhulgas isegi mittekonkatenatiivsed morfotaktilised protsessid (Beesley, Karttunen 2000).

Eesti keeles esinevad järgmised sõnavormide moodustamisviisid:
1) aglutinatsioon – morfeemide konkatenatsioon, kusjuures morfeemid on üksteisest selgesti eristatavad:
a) käänamine: *tuba + de + st = tubadest*;
b) pööramine: *ela + ksi + me = elaksime*;
c) sõnatuletus: *kiire + sti = kiiresti*;
d) liitsõnamoodustus: *piiri + valve + väe + osa = piirivalveväeosa*;
2) fleksioon – ühe ja sama tähendusega morfeem muudab erinevates grammatilistes vormides oma kuju, nt *tuba : toa*;
3) supletiivsus – ühe sõna eri vormidel on hoopis erinevad tüved, mis varem on olnud tähenduselt lähedased sõnad, nt *minema : lähen, hea : parem, üks : esimene, kaks : teine*;
4) analüütilisus – abisõnad ja mitmesõnalised vormid:
a) verbi liitajad, nt *oli tehtud, on käinud*;
b) ahelverbid, nt *hakkab olema, paneb põlema*;
c) ühendverbid, nt *alla kirjutama*;
d) väljendverbid, nt *jalga laskma*;
e) kaassõnafraasid, nt *laua peal, metsa sees, kodu poole, minu järel;*
5) reduplikatsioon ehk tüvekordus esineb mõnedes kirjeldavates määr- ja omadussõnades (*kilin-kolin, kimpsud-kompsud, siiruviiruline, kiira-käära, pilla-palla*).

3. Kahetasemeline morfoloogiamudel

Lõplikel automaatidel ja muunduritel põhineb **kahetasemeline morfoloogiamudel**, mille esitas Kimmo Koskenniemi Helsingi Ülikoolist oma väitekirjas (Koskenniemi 1983). Mudel koosneb sõnastikest ja reeglitest, mis mõlemad on modelleeritud lõplike muunduritena. Konkreetse keele morfoloogiline muundur kujutab endast sõnastikmuunduri ja reeglistik-muunduri kompositsiooni:

$$\text{MorphFST} = \text{LexiconFST} \circ \text{RuleFST}$$

ning seda saab kasutada nii morfoloogilise analüüsi kui ka sünteesi tegemiseks.

Mudeli kahetasemelisus tähendab seda, et sõnastikus säilitatakse morfeemide nn süvakujusid, millest reeglite ja sõnastikevaheliste viitade abil saab moodustada kõik tegelikkuses esinevad sõnavormid (näide 1).

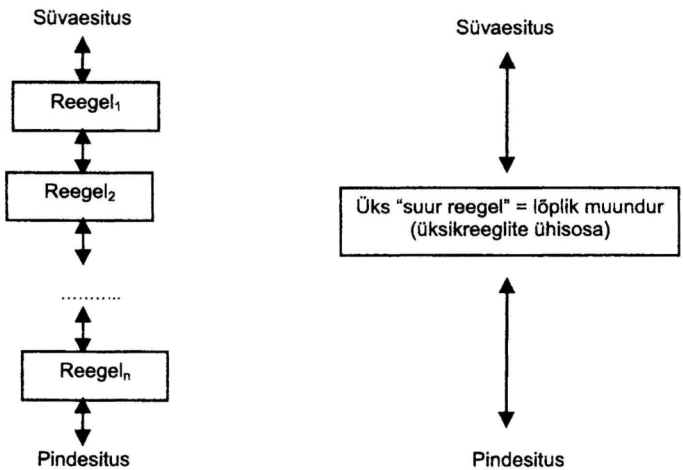
(1) Sõnavormi *mõtetes* süva- ja pindesitus
süvaesitus: m õ t T e $ + t e + s #
pindesitus: m õ t 0 e 0 0 t e 0 s 0

Kahetasemeline morfoloogiamudel on olemuselt keelest sõltumatu. Erinevate keelte puhul võib sõnastike ja reeglite koostamine olla aga lihtsam või keerulisem, sõltuvalt keele morfoloogilisest süsteemist. Mudel sobib hästi aglutineerivate keelte morfoloogia kirjeldamiseks (need on keeled, mille sõnavormid moodustatakse morfeemide konkatenatsiooni teel). Morfotaktika reeglid defineeritakse sel juhul sõnastike süsteemis sõnastikevaheliste viitade abil. Flekteerivate keeltega (kus käänamisel, pööramisel ja tuletamisel etendab olulist osa tüve teisenemine) tekib rohkem raskusi, kuid sõnatüvede muutusi on tavaliselt võimalik kirjeldada kahetasemeliste reeglite abil. Lihtne on kirjeldada tüvemuutusi, mis hõlmavad ühte häälikut (näiteks astmevaheldus, hääliku lisandumine või kadu). Kui tüvemuutus on keerulisem, hõlmab mitut häälikut (nt *idu : eo*), tuleb sellele läheneda analüütiliselt – töödelda iga häälikumuutus omaette reegliga. Eesti keel on nii aglutineeriv kui flekteeriv. Näiteks sõnavorm *hammastega* moodustatakse ühelt poolt morfeemidest *hammas* + *te* + *ga* (aglutineeriv), ning teiselt poolt, kuna tegemist on astmevaheldusliku sõnaga, määravad reeglid kindlaks, et tüvekuju on just *hammas* ja mitte *hamba* (flekteeriv).

Reeglid seavad sõnavormide sõnastiku- ja pindesitused omavahel vastavusse. Iga reegel on realiseeritud lõpliku automaadina. Kogu reeglistik kujutab endast üksikreeglite ühisosa (arvutatakse automaatide ühisosa ehk intersektsioon), mis tähendab, et kõik reeglid peavad olema üheaegselt rahuldatud. Lõplike muundurite ühisosa ei ole arvutatav üldjuhul, vaid ainult tingimusel, et sõned on ühepikkused. Selle tingimuse täidetus on kahetasemelises mudelis saavutatud, asendades tühisümbolid reaalsete nullidega (0).

Järjestatud ümberkirjutusreeglitel põhineva generatiivse fonoloogia üheks puuduseks on see, et tuletusprotsessi ajal kättesaadavad fonoloogilised tunnused ei suuda kirjeldada kogu selle taga olevat morfoloogiat (Karlsson 1974). Seevastu kahetasemelistele reeglitele

on kättesaadavad nii fonoloogilised kui morfoloogilised tunnused, mis on tegelikult mõlemad aluseks sõnade klassifitseerimisel muuttüüpidesse. Võrreldes kahetasemelise mudeli reeglistikku järjestatud ümberkirjutusreeglitest koosneva grammatikaga on põhierinevus selles, et kahetasemelises mudelis on täpselt kaks lingvistilise tähendusega kuju – sõnastiku- ja pindkuju (Kaplan, Kay 1994), kuna vahepealsed sõned, mida saadetakse ühest ümberkirjutusreeglist teise, on elimineeritud. Võrdlevalt on need kaks lähenemist toodud joonisel 3.



Joonis 3. Järjestatud fonoloogilised ümberkirjutusreeglid *versus* kahetasemelised reeglid

Kahetasemelise reegli üldkuju (Koskenniemi 1983) on

CP op LC _ RC,

kus op ∈ {⇐, ⇒, ⇔}, CP on sümbolipaar, LC – vasakpoolne kontekst ja RC – parempoolne kontekst. CP, LC ja RC on üldjuhul hulgad, mitte konkreetsed sümbolid või sümbolijärjendid. Nii on näites 2 (vokaali madaldumise reegel) kasutatud kõigi eesti keele vokaalide hulka Vok = {a, e, i, o, u, õ, ä, ö, ü}.

(2) Reegel "Vokaali madaldumine"

```
KorgeVok : MadalVok <=> Algus _ LV: (%=:) [a|e|i:|u:](l)
%$: ;
```

```
                    Algus Vok (%.:) LV: (%=:) _ %$:;
where KorgeVok in (u ü i)
      MadalVok in (o ö e)
matched ;
```

Sõnastikes on paralleelselt vaatluse all morfoloogiline info ja sõnavormi sõnastikuesitus (näide 3). Sõnastike ja reeglite kompositsioon annab tulemuse, et lemmale ja morfoloogilisele infole seatakse vastavusse sõnavormi pindesitus ehk kirjapilt.

(3) Väljavõte eesti keele kahetasemelise morfoloogia sõnastike süsteemist: nimisõnatüvede sõnastik ja käändelõppude sõnastik

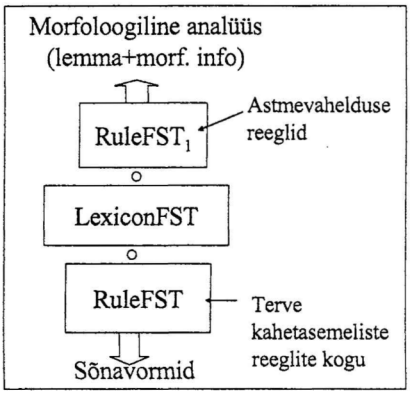
```
LEXICON Substantive
 hammas:hamBa 07_S-0;
 hein 23_A;
 huvi:huv=i 17_Adt;
 idee:ide. 26;
 jalg:jalG 22_A;
 jõgi:jõG=i 18_Adt;
 kala 17;
LEXICON Cases_1
 +ill:+sse GI;
 +in:+s GI;
 +el:+st GI;
 ....
 +ab:+ta GI;
 +kom:+ga GI;
```

4. Eesti keele kahetasemeline morfoloogia

Lõplikel muunduritel põhinev eesti keele morfoloogiamudel on koostatud lähtudes kahetasemelise morfoloogiamudeli põhimõtetest: morfoloogiakirjeldus koosneb sõnastike võrgustikust ja hulgast kahetasemelistest reeglitest, kusjuures astmevahelduse reeglite alamhulka rakendatakse ka eraldi (joonis 4). Sellise lähenemise põhjustest on juttu alajaotuses 5.4.

4.1. Reeglid

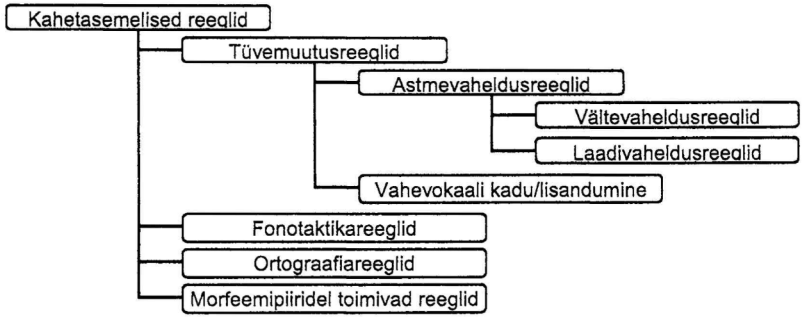
Kahetasemeliste reeglitena on formaliseeritud fonotaktika (nt *lumi* : *lumd** → *lund*), astmevahelduse (nii laadi- kui vältevaheldus, nt *kägu : käo*, *hüpata : hüppan*), morfofonoloogilise distributsiooni (*seis* + *da* → *seista*) ja ortograafiareeglid (*kristall* + *ne* → *kristalne*); morfotaktika reeglid on aga sätestatud sõnastikes.



Joonis 4. Eesti keele kahetasemelise morfoloogiamudeli arhitektuur

Tüvelõpumuutuste kirjeldamine jaguneb sõnastike ja reeglite vahel, kuna rohkem kui ühte grafeemi hõlmavat teisendust on äärmiselt ebamugav reeglitega läbi viia – see nõuab reeglite omavahelist koordinatsiooni. Kuna iga reegel vaatleb korraga vaid ühte süva- ja pindsümboli vastavust, on loomulik kirjeldada reeglitega ühte häälikut puudutavad häälikumuutused. Kahetasemelised reeglid ei suuda korraga teisendada ühte tervet segmenti teiseks ning teisenduse läbiviimine mitme reegli abil nõuab reeglite omavahelist koordinatsiooni. Seepärast kirjeldatakse "ebaloomulikud", fonoloogiliselt ja morfoloogiliselt põhjendamatud häälikumuutused sõnastike süsteemis. Reeglitega saab kirjeldada nii fonoloogiast kui morfoloogiast tingitud häälikumuutusi, kuna kontekstides võivad esineda nii pind- kui süvaesituse sümbolid. Fonotaktilise häälikumuutuse korral piisab enamasti pindkontekstist, morfo(fono)loogiline muutus aga nõuab ka süvaesituse kaasamist konteksti.

Reeglite koostamisel osutus kasulikuks see, et reeglite kontekstides saab kasutada sõnaosade piire märkivaid tunnuseid. Näiteks astmevahelduse reeglid vaatlevad ainult sõnatüve sisehäälikuid ning tüve lõppu paigutatavat nõrga astme tunnust, suur osa reeglitest käivitub just tüve ja tunnuse/lõpu piiril või liitsõnapiiril (nt reegel: "umbisikulise tegumoe tunnus *d*→*t* pärast *s*-i või *h*-ga lõppevat tüve"). Ülevaate tervest reeglite kogust annab joonis 5 ning mõned reeglid (kokku on eesti keele kahetasemelise morfoloogia reeglistikus 45 reeglit) on toodud näites 4.



Joonis 5. Eesti keele kahetasemelise morfoloogia reeglite kogu

(4) Reeglite näiteid

a) välte- ja laadivaheldus
"AV 3 - t kadu"
T:0 <=>Vok [t|h] _ Vok (S:) %$:; !tt-t (ro**tt**-ro**t**i) ja ht-h (eh**t**ima-ehib)
Algus Vok: _ e (l) %$:; !ü**t**lema-öelda, jä**t**ma-jäetakse
"AV 7 - assimilatsioon d-n" !hi**nd**-hi**nn**a
D:n <=> n _ (Vok) (S:) %$:;

"AV 11 - g kadu"
G:0 <=> Vok _ (%=:) (Vok|h) %$: ;
[a | i | õ | u][l | r] _ (Vok) %$: ; ! jal**g**-jala, kir**g**-kire
[e | ä | ü] [l | r] _ [i | u] %$: ; ! pür**g**ima-pürin
a %.: _ i: %$: ; !saa**g**ida-saen

"AV 17: g-j" !mär**g**-mär**j**a, hül**j**es-hül**g**e
G:j <=>[e | ä | ü][l | r] _ [a | e](S:) %$: ;

b) vahevokaali kadu või lisandumine
"Vahevokaali kadu" ! Kui l,m,n,r,v-lõpulisele sõnatüvele lisandub tüvevokaal, siis !vahevokaal kaob, nt. tü**t**ar-tütre, suhku**r**-suhkru.
VaheVok:0<=>_ [Liq|m|n|v|s](%+:)TyveVok;

c) fonotaktika
"m → n t|d ees" ! (lu**m**d→lu**n**d)
m:n <=> _ %+: [d | t] ;
"j → i sõna lõpus" ! (kir**j**→kir**i**, pur**j**→pur**i**)
j:i <=> _ [.#.|%&:] ;

"vokaalide teisenemine järgsilbis o→u, ö→e, ä→e" ! **soo**-soh**u**, **töö**-töh**e**, p**ea**-päh**e**
%.:V2 <=> V1 %.: h _ Piir;
where V1 in (a o u ö ä)
V2 in (a u u e e)
matched;

```
d) morfeemipiiridel toimuvad muutused
"i → e enne i-ga algavat formatiivi"! kauni+im=kauneim
 i:e <=> Kons _ (S:) %+: i ;

e) ortograafia
"Konsonantühendi reegel "  ! kukkru->kukru, kristallne->kristalne.
  K1:0 <=> Vok _ :K1 (%$:)(%+:)(VaheVok:0) [Kons-K1];
        where K1 in Kons;

"Liitsõnapiirile poolituskriips kolme ühesuguse hääliku korral"  !plekk-katus,
jää-äär
  %&:%- <=> :A1 (:0) :A1 (:0) _ :A1; where A1 in (a e f h i k l m n o p r s sh t
u õ ä ö ü);
```

4.2. Sõnastikud

Sõnastike abil kirjeldatakse eesti keele kahetasemelises morfoloogias järgmised protsessid:

- käänamine;
- pööramine;
- omadussõnade võrdlemine;
- sõnatuletus;
- liitsõnamoodustus;
- tüvelõpumuutused ne-se, 0-da, 0-me jne;
- tüvevokaali valik a, e, i, u.

Eesti keele kahetasemelise morfoloogia sõnastike võrk on üles ehitatud vastavalt Ülle Viksi morfoloogilisele klassifikatsioonile (Viks 1994). Viksi morfoloogiline klassifikatsioon võeti aluseks ühelt poolt selle ülevaatlikkuse ja optimaalsuse tõttu ja teisalt selleks, et lihtsustada tüvedesõnastiku automaatset täiendamist. (Eesti Keele Instituudis on kirjutatud tarkvara, mille abil tuvastada sõna muuttüüp vastavalt tema häälikkoostisele.) Sõnastikes on nüüdseks kirjeldatud kõikide noomeni- ja verbitüüpide paradigmad. Nii noomenite käänamine, verbide pööramine, adjektiivide võrdlusastmete moodustamine kui sõnatuletus ja liitsõnamoodustus on realiseeritud jätkusõnastike abil. Veidi lihtsustatud ülevaate sõnastike süsteemi struktuurist annab joonis 6, milles orienteeritud kaared vastavad sõnastikke ühendavatele jätkuviitadele.

Sisu järgi võib eristada järgmisi sõnastikke:

1) tüvikusõnastikud;
2) tüvelõpumuutuste sõnastikud;
3) tunnuste ja lõppude sõnastikud;

4) hargnemissõnastikud.

Sõnatüüpide sõnastikud (noomenitüübid 01–26 ja verbitüübid 27–38) sisaldavad endas hulka jätkuviitadega ühendatud sõnastikke, millest esimene grupp tegeleb tüvevariantide moodustusega, teine tüvevariantide paigutusega paradigmas, kolmas põhivormide ja nende analoogiavormide moodustamisega. Selline struktuur on põhimõtteliselt üle võetud "Väikesest vormisõnastikust" (Viks 1992).

Erinevatele grammatilistele tähendustele vastavad morfoloogilised märgendid on teisendatavad T. Puolakaineni morfoloogilises ühestajas (Puolakainen 2001) kasutatavateks märgenditeks, et kahetasemelisel morfoloogiamudelil põhineva morfoloogilise analüsaatori väljund sobiks nimetatud morfoloogilisele ühestajale sisendiks.

5. Eesti keele kahetasemeline morfoloogia: probleemid ja lahendused

Järgnevalt käsitleme mõningaid eesti keele kahetasemelise morfoloogiakirjelduse koostamisel tekkinud probleeme ja nende võimalikke või juba realiseeritud lahendusi.

5.1. Tüvemuutused

Läbiv lahendus on "mitu ühes", st kõik tüvevariandid on reeglite abil tuletatavad ühest ja samast tüvikusõnastiku kirjest, kasutades leksikaalseid sümboleid (morfofoneeme), millele reeglid seavad sõltuvalt kontekstist vastavusse erinevaid foneeme pindesituses (näide 5). Sel viisil kirjeldatakse muutusi tüve sisehäälikutes (astmevaheldus) ja fonoloogiliselt tingitud tüvelõpumuutusi.

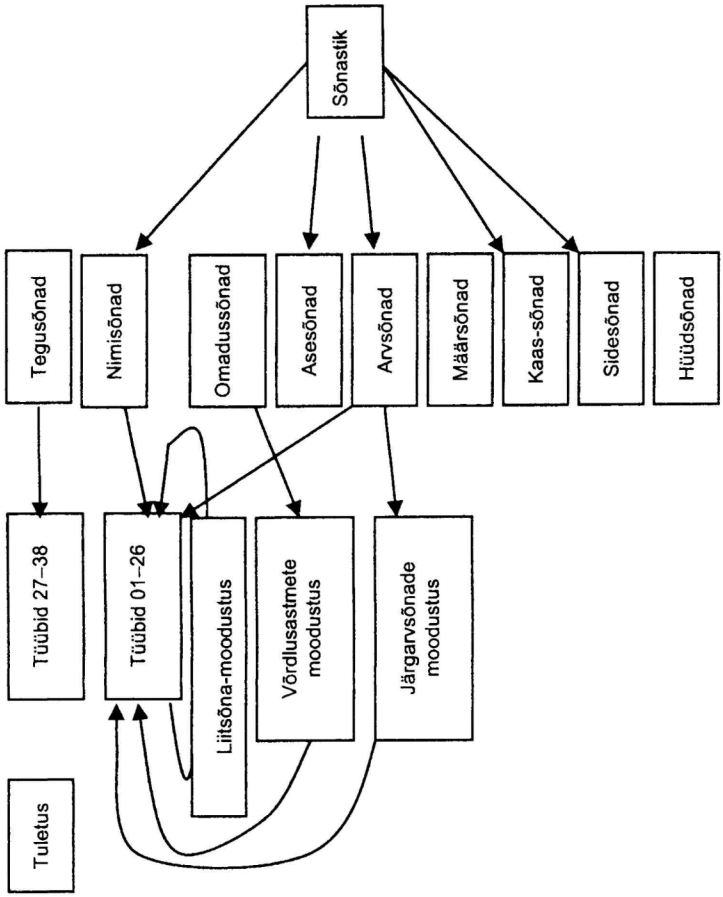
(5) Sõna *jõgi* tüve sõnastikuesitus ja sellest reeglite abil tuletatavad tüvevariandid

jõG=i → jõgi, jõe, jõge, jõkke

Ebaloomulikumaid tüvelõpumuutusi, nt ***hobune*** *:* ***hobuse*** *:* ***hobust*** käsitletakse jätkusõnastike abil (näide 6).

(6) Tüvelõpumuutuste kirjeldamine sõnastike abil

```
LEXICON Substantive
hobu 10_NE-SE-S;
```



Joonis 6. Eesti keele kahetasemelise morfoloogia sõnastike süsteem

```
LEXICON 10_NE-SE-S !soolane
:ne An_SgN;
:se An_SgG;
:s An_SgP_t;
:s An_PlG_te;
:se An_PlP_id;
```

5.2. Morfoloogiliste tunnuste ja lõppude lisamine

Sõnade käänamine ja pööramine on kirjeldatud kolme tasandi sõnastike abil:

1. tüvikusõnastikud, millest lähevad viited muuttüüpide sõnastikele (sõnastik Substantive näites 6);

2. iga morfoloogilise muuttüübi jaoks sõnastik, milles kirjeldatakse tüvevariantide paigutus paradigmas (sõnastik 10_NE-SE-S näites 6);

3. grammatiliste tunnuste ja lõppude lisamine ning nende vastavusse seadmine morfoloogilise informatsiooniga (grammatiliste kategooriate väärtuste kombinatsiooniga) – analoogiarühmade sõnastikud An_xxx näites 7, käände- ja pöördelõppude sõnastikud, nt Cases_1 ja Cases_2.

(7) Käänamist kirjeldavad jätkusõnastikud

```
LEXICON An_SgN
+Sg+N:0 GI;
Compound;

LEXICON An_SgG
+Sg+G:0 GI;
+Sg:0 Cases_1;
+Pl+N:+d GI;
Compound;

LEXICON An_SgP_t
+Sg+P:+t GI;
LEXICON An_PlG_te
+Pl+G:+te GI;
+Pl:+te Cases_1; !mitmuse käänded
LEXICON An_PlP_id
+Pl+P:+id GI;
+Pl:+i Cases_2; ! tüvemitmuse käänded
LEXICON Cases_1  ! käändelõpud illatiivist komitatiivini
+Ill:+sse GI;
+In:+s GI;
...
+Kom:+ga GI;

LEXICON Cases_2  !käändelõpud illatiivist translatiivini
+Ill:+sse GI;
...
```

5.3. Morfoloogilised muuttüübid

Muuttüüpide puhul tuleb teha valikuid, mida kirjeldada sõnastike ja mida reeglite abil. Praegune lahendus on selline, et sõnastike võrgustik põhineb Viksi tüübisüsteemil (Viks 1992). Regulaarsed tüvemuutused, fonotaktika, ortograafia ja morfoloogilise distributsiooni reeglid on kirjeldatud kahetasemeliste reeglitena. Süsteem on veidi tasakaalust väljas – sõnastikel on kogu mudelis suhteliselt suur osa kanda. Seepärast võib küsida: kas reegleid ei peaks rohkem olema? Siin aga tuleb teha kompromiss arvutusliku efektiivsuse ja inimesele loetavuse vahel. Lisaks võimaldab Viksi tüübisüsteem tüvikusõnastike automaatset täiendamist, kasutades EKI-s välja töötatud automaatse tüübituvastuse moodulit.

5.4. Sõnatuletus

Absoluutselt produktiivne tuletus on modelleeritud sõnastike süsteemis ja seejuures tekkinud probleemid on nüüdseks lahendatud. Esialgu ei olnud selge, millisest sõnastikust võetakse tuletatud tüvega sõna puhul algvorm ja sõnaliik ning kuidas käsitleda astmevaheldust like sõnade tuletamist (juur võib olla tugevas astmes, aga tuletatud sõna nõrgaastmelise tüvega). Probleem oli produktiivse tuletusega nõrgeneva tüvega verbides: esialgses lahenduses oli eraldi verbituletiste sõnastik, mistõttu iga verbitüvi esines sõnastikes 2-3 korda (näide 8).

(8) Verbituletuse modelleerimine tüvede dubleerimisega

```
LEXICON Verb
lugema+V : luGe 28;

LEXICON 28 ! Tüvevariantide moodustamine
TP_28at;
: $ TP_28an;

! Tüvevariantide paigutus paradigmas
LEXICON TP_28at ! luge+...
An_ma;
Am_mata;
An_v;
An_sin;
An_sime;
An_da;
An_ge;
Ja_mine;
```

```
LEXICON TP_28an ! loe+...
 An_b;
 An_me;
 An_tud;
 An_takse;

! Põhivormid ja nende analoogiavormid.
LEXICON An_ma
 ma+V+sup+ill :                ma        GI;
 ma+V+quot+pres+ps :  vat      GI;

LEXICON An_v
 ma+V+partic+pres+ps : v       GI;
 v+A+pos+sg+nom+partic :       v         02_A;
 ...

LEXICON An_takse
 ma+V+indic+pres+imps+af :  takse  GI;

! Produktiivne tuletus
LEXICON Verb-Deriv
  loe Partic/N-N;
  luge Partic/N-T;

LEXICON Partic/N-N
  tav+A :                  tav       A_02_A;
  tav+S :                  tav       Axx;
  tud+A+Sg+N :             +tud      #;
  tu+S :                   tu        01;

LEXICON Partic/N-T
  v+A :        v           A_02_A;
  nud+A :      nud         #;
  nu+S :       nu          01;
                           Ja_mine;
LEXICON Ja_mine
  ja+S :       +ja         01;
  mine :       +m          12_nE-SE-S;
  mata+A :     +mata       #;
```

Kirjeldatud probleemile leitud lahenduse idee seisneb selles, et ka lemmadele (ehk sõnastik-muunduri vasakule või ülemisele poolele) rakendatakse astmevahelduse reegleid (joonis 4). Milliseid muudatusi toob see kaasa sõnastike süsteemis, seda on illustreeritud näites 9. Tulemusena ei ole iga verbi jaoks vaja enam kolme, vaid ainult ühte tüvekirjet. Eesti keele morfoloogiline muundur on siis lühidalt kirjapandav valemina, mis on samaväärne joonisega 4:

$$((\text{LexiconFST})^{-1} \circ \text{RuleFST}_1)^{-1} \circ \text{RuleFST},$$

kus LexiconFST on sõnastik-muundur, RuleFST on reeglistik-muundur (kõikide kahetasemeliste reeglite ühisosa) ja RuleFST$_1$ on astmevaheldusreeglite ühisosa. Operatsioon ∘ tähistab muundurite kompositsiooni ja operatsioon $^{-1}$ (tavaliselt tähistatakse matemaatikas niimoodi pöördelemendi võtmist) muunduri ümberpööramist.

(9) Verbituletuse modelleerimine: uus lahendus

```
LEXICON Verb
 luGe 28;

! pöördeliste vormide moodustamine nagu näites 8; vahele jäetud

! produktiivne tuletus
LEXICON 28_deriv
 ja+S                :       +ja     Szz;
mine+S       :       +mine   Sqq;
v+A          :       +v      Aww;
$tav+A       :       @+tav   Aww;
+nud+G       :       +nud    #;
$tud+G       :       $+tud   #;
nu+S                :       +nu     Scc;
$tu+S        :       $+tu    Sdd;

LEXICON Substantive
 Scc;
 ...
LEXICON Adjective
Aww;
```

Väga keeruline on aga osaliselt produktiivse tuletuse modelleerimine. Näiteks esineb olukordi, kus ühest sõnaliigist teise siirdumisel kasutatakse erinevaid tuletusliiteid ja ei ole võimalik anda reeglit, millisel juhul kasutada üht, millisel teist, millisel mõlemaid (näide 10).

(10) Osaliselt produktiivne tuletus: määrsõnade tuletamine omadussõnadest

Omadussõna	Sellest tuletatav(ad) määrsõna(d)
kiire	kiiresti, kiirelt
halb	halvasti, *halvalt
rikas	*rikkasti, rikkalt

Kas panna kõik tuletatud sõnad omaette kirjetena sõnastikku või markeerida kuidagi sõnad, millele saab teatavaid tuletusprotsesse rakendada? Millisel juhul lugeda tuletusprotsessi piisavalt produktiivseks? Milliseid formaalseid tunnuseid saab kasutada, kui tahame sõnatuletust reeglite abil modelleerida? Kõik need küsimused on hetkel vastuseta.

5.5. Liitsõnamoodustus

Vastavalt eesti keele õigekirjareeglitele ei ole liitsõnaosiste piirid enamasti markeeritud. Liitsõnade analüüsimine on eesti keele puhul üks keerulisemaid probleeme, kuna ühelt poolt on liitsõnamoodustus suhteliselt vaba, aga teiselt poolt, semantiliselt mõistlike liitsõnade genereerimine ja äratundmine on arvutile väga raske.

Loodud sõnastike süsteemis on lahendus esialgu väga üldine (Uibo 2002): omavahel saab kombineerida kõiki nimisõnu, osa määrsõnu saab olla liitsõna eelkomponendiks ning omadussõnad ja verbide partitsiipidest tuletatud nimi- ja omadussõnad võivad olla liitsõna järelkomponendiks.

Sõnastike süsteemis realiseeritud liitsõnamoodustuse reeglid on järgmised:

a) (nimisõna (sg nom | sg gen)) + nimisõna
   raudteejaam, jututuba, laevatrepikäsipuu
b) määrsõna + verbituletis
   ümbertehtud, ülejooksja, mahajääja
c) (nimisõna sg nom | X-sõna) + omadussõna
   tulikuum, hirmkallis, ebakindel, uhiuus

Sellised reeglid genereerivad liitsõnu selgelt rohkem kui vaja. Näiteks on morfoloogilise muunduri jaoks aktsepteeritavad liitsõnad nagu **rõõmunimivartest, *möödahüljanud, *uksesööjais, *kuhisilmade, *veauned, *õppimismerede.*

Sõltub morfoloogiakomponendi rakendusest, kuivõrd segav selline ülegenereerimine on. Kui infootsingu puhul ei ole see väga suureks probleemiks, siis õigekirjakontrollija puhul küll. Ühelt poolt on sõnamoodustuse protsessid eesti keeles vabad – igaüks võib kokku panna uue liitsõna, mida keegi varem pole kasutanud, ning speller ei peaks selliseid uudismoodustisi vigadeks märkima. Seetõttu ei ole hea loetleda sõnastikus lõplik hulk võimalikke tuletisi ja liitsõnu. Teisalt, kui morfoloogiakirjeldus lubab väga produktiivset sõnamoodustust, tekivad teist laadi probleemid. Keskmiselt pooled eestikeelses tekstis leiduvatest sõnavormidest on homonüümsed (kahe või enama morfoloogilise tõlgendusega). Juhtub ka seda, et tuletis või liitsõna on homonüümne mingi lihtsõnavormiga. Kui sõnatuletuse ja liitsõnamoodustuse protsesse mitte piirata, siis tekib selliseid homonüüme veelgi rohkem, kusjuures mõnel juhul on värdtuletis või -liit-

sõna juhuslikult homonüümne mingi vigase lihtsõnavormiga, mistõttu vigane sõnavorm jääb spelleril avastamata. Vaatleme mõningaid näiteid.

**naljakass* = *nalja+kass* S Sg Nom – võimalik, kuid veider liitsõna, esinemise tõenäosus nullilähedane. Arvatavasti taheti kirjutada sõna *naljakas*.

**kaustatud* = *kaustatu* S Pl Nom (mõeldi sõna *kasutatud*, aga kaks tähte läksid vahetusse)

Arvutil kiiresti kirjutades on sellised vead väga tavalised. Siin oleks tõenäoliselt abiks tekstikorpusest saadud sõnasagedused. Sõnavormi **kaustatud* sagedus on arvatavasti 0, aga sõnavormil *kasutatud* üsna kõrge.

Kuidas sõnamoodustuse produktiivsust optimeerida, on veel avatud probleem. Igal juhul on ülegenereerimise vältimiseks vaja täpsustavaid reegleid. Seejuures on liitsõnamoodustusele kaht tüüpi kitsendusi:

1. morfoloogilised – eelkomponendi kääne (sg nom, sg gen, harva ka pl gen), mida ka praegu liitsõnamoodustuse modelleerimisel kasutatakse;
2. semantilised – arvestada liitsõna komponentide semantikat.

Milliseid semantilisi tunnuseid saaks liitsõnamoodustuse kitsendamiseks kasutada, ei ole veel selge. Teine võimalik lahendus on kasutada suure korpuse analüüsimisel saadud statistikat. See aga nõuab põhimõtteliselt erinevat lähenemist – kasutada tavaliste lõplike muundurite asemel kaalutud lõplikke muundureid (ingl *weighted finite state transducers*).

5.6. Sõnastikuesituses kasutatavad lisatunnused

Astmevahelduslike sõnade puhul on nõrk aste markeeritud sümboli $ abil tüve lõpus (nagu Koskenniemi 1983). Reeglid kasutavad selle markeri olemasolu kontrolli oma kontekstitingimustes.

II ja III välde ei ole kirjapildis eristatavad (välja arvatud sulghäälikutel). Välte märkimine sõnastikuesituses võiks olla kasulik tekst–kõne sünteesil ja sõna muutetüübi automaatsel tuvastamisel.

5.7. Universaalsus *vs* keeletehnoloogilistele rakendustele orienteeritus

Kas eesti keele morfoloogia kirjeldus peaks olema võimalikult universaalne või orienteeritud konkreetsetele keeletehnoloogilistele rakendustele?

Näiteks automaatpoolitus eeldab, et liitsõnapiir on markeeritud. Õigekirjakontroll on väga tundlik ülegenereerimise suhtes, kuid infootsing ei ole, samal ajal on viimase jaoks vajalik morfoloogiline ühestamine.

Hetkel on keelekirjeldus eesti keele kahetasemelises morfoloogiamudelis universaalne.

6. Kokkuvõte ja tulevikuväljavaated

Katsed rakendada lõplikel muunduritel põhinevaid meetodeid eesti keele morfoloogia modelleerimisel on viinud järeldustele, et eesti keele seisukohast on kahetasemelise morfoloogiamudeli tugevad küljed järgmised:

1) kahetasemeline esitus on kasulik tüvemuutuste kirjeldamisel, eriti seetõttu, et tänapäeva eesti keeles ei sõltu muutetüüp enam sageli sõnatüve häälikkujust;
2) sõnastike süsteem ja morfeemipiiridel toimivad reeglid kirjeldavad loomulikul viisil kogu eesti keele morfotaktika;
3) jätkusõnastike abil on mugav kirjeldada keerulisemaid tüvelõpumuutusi.

Põhiprobleem morfoloogia modelleerimisel on erinevate morfoloogianähtuste produktiivsus. Kui nähtus on absoluutselt produktiivne, siis on seda lihtne formaliseerida – kas reeglina või osana sõnastike võrgustikust. Erandid põhjustavad alati probleeme ja viivad keele kirjeldamisel sageli kohmakate lahendusteni.

Lahendamist vajavad ülesanded on hetkel järgmised:

1) tüvikusõnastike mahu (automaatne) suurendamine;

2) ülegenereerimise vältimine liitsõnamoodustusel;

3) eesti keele kahetasemelise morfoloogia praktikas rakendamine. Huvipakkuv idee, mida realiseerida, on hägus infootsing (mis leiab võtmesõna üles ka siis, kui see on veidi vigaselt kirjutatud).

4) Tundmatute sõnade analüüs. On tehtud esialgne katse panna sõnastikku lisaks konkreetsetele tüvedele ka regulaaravaldisi, mis

defineerivad teatava häälikulise struktuuriga tüvesid (sisuliselt on see tüve häälikulisel kujul põhinev tüübituvastus).

Näiteks võiks nimisõnade sõnastikus esineda kirje:

<(C) V (V) C (C) V (C V)> 01; (tüüp "aasta")

Selline lähenemine on edukas siis, kui häälikkuju määrab üheselt sõna muuttüübi (alati aga ei määra). Võib arvata, et sõnade äratundmise saagis suureneb, kuid täpsus väheneb (mustriga sobib ka hulgaliselt olematuid sõnu). Et analüüsid ei korduks, on ilmselt vaja arvutada põhisõnastiku ühend tundmatuid sõnu analüüsiva sõnastikuga.

Kirjandus

Beesley, Kenneth, Karttunen, Lauri 2000. Finite-State Non-Concatenative Morphotactics. – Proceedings of SIGPHON-2000. 5th Workshop of the ACL Special Interest Group in Computational Phonology, Centre Universitaire, Luxembourg, 1–12.

Beesley, Kenneth, Karttunen, Lauri 2003. Finite State Morphology. CSLI Studies in Computational Linguistics. Stanford, USA: CSLI Publications.

Chomsky, Noam 1957. Syntactic Structures. Mouton.

Chomsky, Noam, Halle, Morris 1968. The sound pattern of English. New York: Harper and Row.

Johnson, Douglas 1972. Formal Aspects of Phonological Description. The Hague: Mouton.

Kaplan, Ronald, Kay, Martin 1981. Phonological rules and finite-state transducers. – Linguistic Society of America Meeting Handbook, Fifty-Sixth Annual Meeting, New York.

Karlsson, Fred 1974. Phonology, Morphology and Morphophonemics. – Gothenburg Papers in Theoretical Linguistics. Göteborg.

Karttunen, Lauri 1995. The Replace Operator. – Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. ACL-95, Boston, Massachusetts, 16–23.

Karttunen, Lauri 2001. A short history of two-level morphology. – ESSLLI 2001 Special event "20 years of finite state morphology", Helsinki.

Karttunen, Lauri, Kaplan, Ronald M., Zaenen, Annie 1992. Two-level morphology with composition. – COLING 1992.

Koskenniemi, Kimmo 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production.

Helsinki: University of Helsinki, Dept of General Linguistics, Publications No. 11.

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestaja. Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu.

Schützenberger, Marcel Paul 1961. A remark on finite transducers. – Information and Control, 4(2–3), 185–196.

Uibo, Heli 2002. Experimental Two-Level Morphology of Estonian. – LREC 2002. Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain. Proceedings, Vol III., 1012–1015.

Viks, Ülle 1992. A Concise Morphological Dictionary of Estonian I: Introduction & Grammar. Tallinn.

Viks, Ülle 1994. Klassifikatoorne morfoloogia. (Dissertationes philologiae estonicae Universitatis Tartuensis). Tartu.

Puude pangad meil ja mujal

Heli Uibo, Helen Nigol
Tartu Ülikool

1. Sissejuhatus

Korpusepõhised meetodid on viimastel aastakümnetel loomuliku keele töötluses saavutanud üha enam populaarsust ja edu. Reeglipõhised meetodid on küll lingvistiliselt usaldusväärsemad, kuid reeglite koostamine toimub käsitsi ning on seetõttu töömahukas. Samal ajal statistiliste meetodite kasutamisel "avastab" seaduspärasused arvuti (seda nimetatakse masinõppimiseks, ingl *machine learning*). Seejuures on aga eelduseks, et on olemas piisav hulk õppimismaterjali, st vastavas keeles tekste.

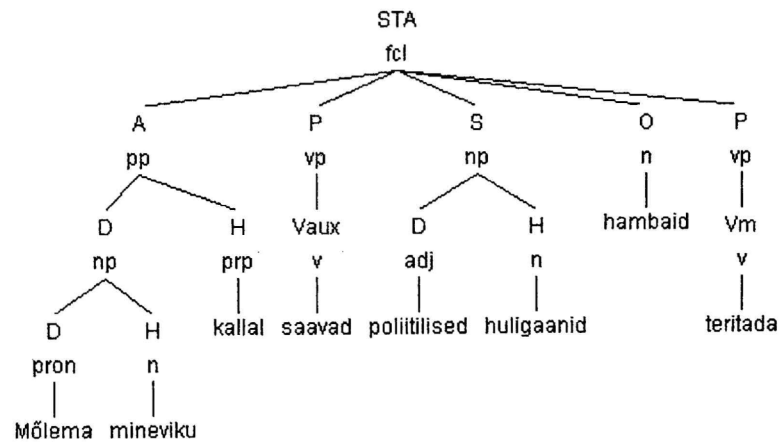
Masinõppimise meetodeid on kahte tüüpi: juhendatavad (ingl *supervised*) ja juhendamata (ingl *unsupervised*) meetodid. Juhendamata meetodid koguvad statistilisi andmeid lingvistilise märgenduseta tekstidest (nn toortekstidest, ingl *raw text*), juhendatavad meetodid eeldavad aga suurte lingvistiliselt märgendatud korpuste olemasolu. Kuigi viimasel ajal on demonstreeritud esimesi edukaid pelgalt juhendamata meetodite kasutamisel põhinevaid eksperimente (nt Klein, Manning 2004), on masinõppimisel parimaid tulemusi saavutatud siiski juhendatavate meetoditega või juhendamata ja juhendatavate meetodite kombineerimisel.

Selleks, et rakendada statistilisi meetodeid automaatsel süntaksianalüüsil, on vaja süntaktiliselt märgendatud korpusi ehk **puude panku** (ingl *treebank*). Mõiste puude pank tuleneb sellest, et süntaktiliselt analüüsitud lause on graafiliselt esitatav puuna[1] (vt näidet joonisel 1). Kuid kokkuleppeliselt loetakse puude pankadeks igasuguseid korpusi, mis sisaldavad lingvistilist informatsiooni sügavamatel keeleanalüüsi tasanditel kui sõnaliigi märgendus – lause moodustajastruktuur, moodustajate süntaktilised funktsioonid, semantiline ja diskursuse analüüs (Nivre jt 2005). Lisaks masinõppimisele saab puude panku kasutada ka morfoloogiliste ja süntaktiliste analüsaato-

[1] Puu graafiteooria mõistes – sidus tsükliteta graaf.

rite ning neil põhinevate keeletehnoloogia rakenduste testimisel, keeleõppetarkvaras ja keeleteaduslikus uurimistöös.

Puude panga loomine on töömahukas ülesanne, kuna isegi juhul, kui puude panka luuakse poolautomaatselt, st esmalt tehakse teksti automaatne morfoloogiline ja süntaktiline analüüs, peab märgendus olema käsitsi kontrollitud ja parandatud, soovitavalt mitme eksperdi poolt. On olemas tarkvaralisi vahendeid puude pankade loomiseks ja kasutamiseks, mis hõlbustavad lingvistide tööd. Puude panga loomise algfaasis on olulisemad otsused puude panga märgendusskeemi ja esitusformaadi valik. Selleks, et oleks võimalik olemasolevat tarkvara kasutada, peab esitusformaat olema vastavate programmide poolt aktsepteeritav või neisse hõlpsasti konverteeritav. Ideaaljuhul peaks puude panga disain (aluseks olev korpus, märgendusskeem, esitusformaat) olema motiveeritud eesmärgiga, milleks seda puude panka luuakse, näiteks kas keeleteaduslikuks uurimistööks või keeletehnoloogiatoodete testimiseks. Praktikas osutuvad sageli määravaks hoopis tekstide ning tarkvara kättesaadavus (Nivre jt 2005).



Joonis 1. Visualiseeritud lause (*Mõlema mineviku kallal saavad poliitilised huligaanid hambaid teritada.*) **eesti keele puude pangast Arborest**

Eesti keele korpuste süntaktilise märgendamisega tehti Tartu Ülikoolis algust 1990. aastate teisel poolel, kuna oli vaja test- ja treeningkorpust kitsenduste grammatika põhisele süntaksianalüsaatorile

(Roosmaa jt 2003). Olemasolev süntaksianalüsaator teeb pindmist süntaktilist analüüsi, määrates igale sõnavormile lauses süntaktilise funktsiooni (alus, sihitis, määrus, öeldis või selle osa, nimisõnaline või omadussõnaline eestäiend jne). Samasugune märgendus on ka kitsenduste grammatika korpuses. Seega ei saa kitsenduste grammatikat pidada veel päris puude pangaks – vähim grupeeritav ühik on seal osalause, seejuures märgendus ei väljenda osalausete omavahelisi seoseid (mis on pealause ja mis kõrvallause, juhul kui on tegemist põimlausega), kuid seda on võimalik rikastada süvasüntaktilise infoga ning tuletada sellest eesti keele puude pank, millega on ka algust tehtud.

Käesolevas artiklis antakse ülevaade tuntumatest teistele keeltele tehtud puude pankadest ning esimestest sammudest eesti keele puude panga loomisel. Kirjeldatakse eesti keele puude panga loomisprotsessi ja märgendamisskeemi ning tuuakse näiteid eesti keele puude panga kasutusaladest – nendeks on hetkel lingvistiline uurimistöö ja eesti keele grammatika õppimine veebipõhiste keelemängude abil. Viimasena käsitletakse mitmekeelseid puude panku ja nende rakendamist näidetepõhisel masintõlkel.

2. Ülevaade mujal tehtust

Esimesi lingvistiliselt märgendatud korpusi hakati looma juba 1960. aastate teisel poolel eesmärgiga parandada otsingu- ja päringuvõimalusi ning korpuste automaattöötlust (Abeillé 2003: 14). Esimene puude pangale sarnane korpus loodi 1970. aastatel Göteborgi ülikoolis A. Ellegårdi (1978) juhtimisel. *Gothenburg Corpus*, nagu seda kutsuti, koosnes *Brown*i *Corpus*e 128 000-sõnalisest osast, mis analüüsiti käsitsi (Leech 1997: 10). Kuid erinevatel põhjustel leidis see puude pank vähe kasutust enne seda, kui see muudeti praegu laialt tuntud puude pangaks *SUSANNE Corpus* (Sampson 2001: 35). Arvestades tänapäevaseid korpuste mahte, on SUSANNE korpus tõeliselt pisike, sisaldades kõigest 130 000 ingliskeelset sõna. Võrdluseks võib tuua 1990. aastail valminud puude panga *Penn Treebank*, mis sisaldab üle 4,5 miljoni ingliskeelse sõna. Vanematest puude pankadest tuntumad on veel ka *Lancaster-Leeds Treebank*, *Nijmegen Corpus* ja *TOSCA Corpus* (Leech, Eyes 1997: 43). Kui 1960.–1990. aastatel loodud puude panku iseloomustab see, et kõik puude pangad loodi eranditult vaid inglise keelele, siis 1990. aastatel

hakati intensiivselt puude panku looma ka teistele keeltele, nt *Negra Corpus* saksa keelele, *Prague Dependency Treebank* tšehhi keelele, *BulTreeBank* bulgaaria keelele.

2.1. Märgendamisskeemid

Puude panga märgendamisskeemi valik sõltub paljudest teguritest. Olulisimaks küsimuseks on, milline teoreetiline raamistik puude pangale valida. Raamistiku valikut mõjutavad omakorda analüüsitava keele grammatiline iseloom ja selle keele deskriptiivse grammatika traditsioon. Samuti tasub vaadata, mida on tehtud teistele keeltele. Ühtse märgendamisskeemi kasutamine aitaks kaasa nii keelte võrdlevale analüüsile kui ka paralleelpuudepankade koostamisele. Sellisel juhul oleksid ka puude pankade loomiseks ja kasutamiseks mõeldud vahendid korduvkasutatavad (Nivre jt 2005: 100).

Puude pankade arvu järkjärgulise kasvamisega kasvab ka erinevate märgendamisskeemide arv. Eristada võib kolme põhilisemat märgendamisviisi:

1) moodustajastruktuuri märgendamist,
2) funktsionaalse struktuuri märgendamist,
3) semantilist märgendamist.

Täpsustuseks olgu öeldud, et peaaegu kõikide puude pankade märgendamisskeemides on elemente kahest, kui mitte kõigist kolmest märgendamisviisist, st selget piiri nende kolme vahele pole võimalik tõmmata.

Lisaks võib märgendamise jagada veel teooriakeskseks ja teoorianeutraalseks. Kuivõrd võiks üks puude pank olla seotud lingvistiliste teooriatega, on küsimus, mis on olnud vaidluspunktiks nii kaua, kui puude pangad on eksisteerinud. On väidetud, et süntaktiliselt töödeldud korpuste koostamine on tõestanud, et püüdlemine üha detailsema andmeanalüüsi poole muutub aina rohkem teooriast sõltuvaks[2]. Sellepärast põhinevadki paljud puude pankade kirjeldused just teatud lingvistilistel teooriatel. J. Nivre (2003) on samas seisukohal, et puude pank peaks olema võimalikult teoorianeutraalne ning teisendatav võimalikult paljudesse erinevatesse formaatidesse. Näiteks annab Nivre (2003) algoritmi fraasistruktuuripuude teisendamiseks

[2] http://www.bultreebank.org/TLT2002.html

sõltuvuspuudeks. See pole võimalik küll üldjuhul, vaid eeldusel, et fraasistruktuuripuudes on fraasipõhjad märgendatud.

2.1.1. Moodustajastruktuuri märgendamine. Moodustajate märgendamine ehk sulundamine (ingl *bracketing*) sisaldab sõnaliigi ja fraasistruktuuride märgendamist. Moodustajate märgendamine on teooriast sõltumatu ja seetõttu proovib kasutada ka üldiselt aktsepteeritavaid kategooriaid, mis on tuntud pea igas süntaksiteoorias. Sellist märgendamist on kasutatud näiteks puude pangas *Penn Treebank*. Sulundatud lause esitus on süntaksipuude esitusest tunduvalt kompaktsem, nagu on näha jooniselt 2.

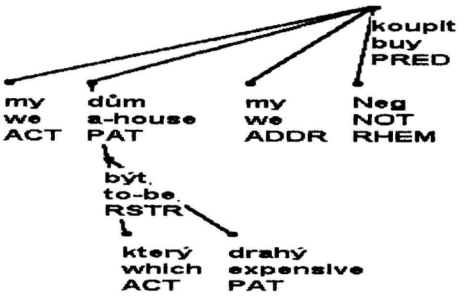
```
( (S
  (NP Martin Marietta Corp.)
      was
      (VP given
              (NP a
                      $ 29.9
                  million Air Force contract
                              (PP for
                          (NP low-altitude navigation
                              and
                              targeting equipment) ) ) ) )
).
```

Joonis 2. Sulundatud lause (*Ettevõttele* Martin Marietta *anti 29,9 miljoni krooni väärtuses õhujõudude leping madala kõrguse navigatsiooniks ja sihtimisvarustuseks*) **puude pangas *Penn Treebank* (Taylor jt 2003: 7)**

2.1.2. Funktsionaalse struktuuri märgendamine. Viimastel aastatel on väga populaarseks saanud funktsionaalse struktuuri märgendamine. Kõige ehedamad näited sellest on märgendamisskeemid, mis põhinevad sõltuvussüntaksil, mida on kasutatud puude pankades *Prague Dependency Treebank* ja *METU Treebank of Turkish*. Mõlemas puude pangas järgneb sõltuvusstruktuuri märgendamine kohe morfoloogilisele märgendamisele, moodustajastruktuuri kindlaks määramata (näitelause Praha Sõltuvuspuudepangast on joonisel 3).

Sõltuvusgrammatikast on välja kasvanud funktsionaalne sõltuvusgrammatika (ingl *functional dependency grammar*; FDG). FDG põhineb Tesnière'i strukturaalsel süntaksiteoorial, mis rõhutab lähedast vastavust süntaktilise ja semantilise kirjelduse vahel. FDG mudeli kasutamine on näidanud, et see suudab süntaktilisi fenomene erinevat tüüpi keeltes adekvaatselt kirjeldada, näiteks on seda mude-

lit katsetatud juba inglise, hispaania, rootsi, saksa, taani ja soome keele peal. (Järvinen 2003: 94) Ka eesti keele sügavama süntaksianalüüsi saavutamiseks nähakse ühe võimalusena just FDG-d (Roosmaa jt 2003: 209).



Joonis 3. Lause *Dům, který je drahý, si nekoupíme* (*Maja, mis on kallis, me endale ei osta.*) **sõltuvuspuu Praha Sõltuvuspuudepangast**

Trendi funktsionaalsuse poole on näha ka moodustajate põhistes märgendamisskeemides. Näiteks puude panka *Penn Treebank II* (II faas 1993–1995) lisati moodustajate struktuuri märgendamisele ka funktsionaalsed märgendid, mis määravad loogilise subjekti ja verbi loogilise objekti (Taylor jt 2003: 8). Funktsionaalne märgendamisviis on kasutusel ka saksa keele puude pangas *TIGER Treebank*, kus moodustajate ja sõltuvuste märgendamine on integreeritud graafi, mille sõlmede märgendid esitavad fraasikategooriaid ja servamärgendid süntaktilisi funktsioone (Brants jt 2002).

2.1.3. Semantiline märgendamine. Juba kaua on oldud seisukohal, et süntaktilised struktuurid üksinda ei suuda anda piisavalt informatsiooni inimkeele automaatseks mõistmiseks. Puude panku, kus on märgendamisskeemi lisatud ka semantiline märgendamine, on vähe. Esile võib tõsta puude pangast *Penn Treebank* välja kasvanud puude panka *Propositional Bank*, kus märgendamisse on uue etapina lisatud predikaadi iga argumendi tähistamine sobiva semantilise märgendiga identifitseerimaks selle rolli predikaadi suhtes, nt subjekt, objekt jne (Kingsbury, Palmer 2003: 105).

Puude pangas *Italian Syntactic-Semantic Treebank* kasutatakse semantiliseks märgendamiseks leksikaalsemantilist andmebaasi *ItalWordNet*. Leksikaalsemantilisel märgendamisetapil varustatakse iga

märgendamisüksus asjakohase tähendusega vastavalt *ItalWordNeti* tähenduseristustele. Kui konteksti sobib rohkem kui üks *ItalWordNeti* tähendus, siis välditakse juhuslike tähenduste määramist, võttes appi teised, juba määratud tähendused (Montemagni jt 2003: 190–200).

2.2. Vahendid käsitsi ja automaatseks märgendamiseks

Esimestes puude panga projektides, nt SUSANNE korpus, märgendati korpused üleni käsitsi. Nüüdseks on see töö küll osaliselt automatiseeritud, kuid inimesepoolset lisakontrolli peetakse endiselt vajalikuks. Käsitsi märgendaja kasutajaliidestest on enim kasutust leidnud algselt *Negra* korpuse jaoks koostatud programm *Annotate*[3], mis on tõhus vahend tekstikorpuste poolautomaatseks märgendamiseks. Praegusel hetkel võib programmi *Annotate* pidada üheks kõige enam kasutatavaks vahendiks teiste omasuguste seas, kuigi peab mainima, et MySQL andmebaasi haldamine, kus hoitakse redigeeritavaid puude panku ning informatsiooni puude panga kasutajate ja nende õiguste kohta, on üsna keeruline ja vaevanõudev, samuti ei ole programmi *Annotate* kasutajaliides paljude kasutajate meelest piisavalt intuitiivne (Volk jt 2005).

Populaarseimad puude pankade kasutamisvahendid on graafilise kasutajaliidesega programm *TIGERSearch*[4] ja päringukeel *tgrep*.[5]

Programm *TIGERSearch* on spetsiaalselt välja töötatud selleks, et kätte saada informatsiooni puude pankadest. *TIGERSearchi* abiga saab leida uuritava sõna leksikaalseid omadusi, nt millistes kollokatsioonides sõna esineb. Samuti saab leida soovitud lauseid. *TIGERSearchi* plussiks on ka see, et ta toetab erinevaid puude pankade formaate.

Programm *tgrep* on sarnane Unixi käsureaprogrammiga *grep*, kuid on kohandatud just erinevate puustruktuuride otsimiseks, millele viitab ka *t* nime algul. Käsureaprogrammi *tgrep* abil saab otsida nii üksikuid sõnu, fraase kui ka lauseid, samuti domineerimis- ja eelnevussuhteid suluesituses puude pangast, nagu nt puude pank *Penn Treebank*.

---

[3] http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html
[4] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
[5] http://mccawley.cogsci.uiuc.edu/corpora/tgrepdocs.html

3. Eesti keele puude pank

Eesti keele puude panga loomise peale hakati mõtlema 2002. aastal. Selleks ajaks oli olemas eesti keele kitsenduste grammatika korpus suurusega 100 000 sõna, milles on märgendatud iga sõna süntaktiline funktsioon.

Eesti keele kitsenduste grammatika korpust luuakse poolautomaatselt. Esmalt lastakse tekstid läbi morfoloogilise analüsaatori (Kaalep 1996), mille väljund ühestatakse käsitsi. Edasi läbivad morfoloogiliselt analüüsitud tekstid kitsenduste grammatikal põhineva süntaktilise analüüsi programmi, mille täpsus[6] on üle 98% ja saagis[7] üle 80% (Müürisep jt 2003). Näide programmi väljundist on toodud joonisel 4. 100% korrektselt märgendatud teksti puhul on iga sõnavormi järel üks ja õige süntaktiline märgend. Sellist tulemust ei ole automaatanalüüsil põhimõtteliselt võimalik saavutada ning tekstid tuleb viimaks üle vaadata eesti keele süntaksi ekspertidel, kes viivad läbi lõpliku süntaktilise ühestamise. Lisaks puudustele, mida oleks võimalik edasise treenimise ja testimisega süntaksianalüsaatorist kõrvaldada, leidub eesti keeles keelekonstruktsioone, mis formaalselt on täpselt ühesugused – vaadeldavas sõnade järjendis langeb morfoloogiline informatsioon kokku, kuid erineva semantika tõttu esineb sõna erinevates süntaktilistes funktsioonides ning seetõttu ei tule programm süntaktilise ühestamisega toime (vrd *Ostis poest* (ADVL) *saia.* ja *Ostis kullast* (NN>) *kella.*). Peale selle on koguni olemuslikult mitmeseid lauseid, kus ka inimene satub süntaktilise märgendi valikul raskustesse.

Iga süntaksianalüsaatorist läbikäinud tekst antakse kahele erinevale inimesele paralleelselt korrastamiseks, et vältida subjektiivsest keeletunnetusest tulenevaid vigu. Lahendada tuleb olukorrad, kus süntaksianalüüsi programm

a) on määranud ühele sõnavormile mitu süntaktilist märgendit;
b) pole määranud mitte ühtegi süntaktilist märgendit;
c) on määranud vale süntaktilise märgendi.

---

[6] Täpsus (ingl k *precision*) – leitud õigete analüüside osakaal kõikide leitud analüüside hulgas.

[7] Saagis (ingl k *recall*, eesti keeles on varem kasutatud ka termineid katvus ja korrektsus) – leitud õigete analüüside osakaal kõikide õigete (inimese poolt tehtud) analüüside hulgas.

```
Olin
    ole+in //_V_ aux indic impf ps1 sg ps af #cap #FinV #Intr // **CLB
@+FCV
neid
    see+d //_P_ dem pl part // @NN>
sõnu
    sõna+u //_S_ com pl part // @OBJ
vanast
    vana+st //_A_ pos sg el // @AN>
kooliraamatust
    kooli_raamat+st //_S_ com sg el // @ADVL @<NN
lugenud
    luge+nud //_V_ main partic past ps #NGP-P // @-FMV
$.
   $. //_Z_ Fst //
```

Joonis 4. Näide süntaksianalüüsi programmi väljundist

Tulemusena peab iga sõnavormi morfoloogilisele analüüsile järgnema üks ja õige süntaktiline märgend. Joonisel 4 antud näite puhul peaks siis valima õige tõlgenduse sõnavormile *kooliraamatutest* – kas ADVL (määrus) või <NN (nimisõnaline järeltäiend).

Töö hõlbustamiseks on loodud süntaktilise märgendamise kasutajaliides (http://kadri.pirn.net). Kasutades nimetatud tarkvara, ei pea süntaktilist märgendust kontrolliv ja parandav lingvist märgendeid *Delete*-klahviga kustutama ega klaviatuurilt sisse tippima, vaid kõike seda saab teha hiirega menüüst valides. Nii hoitakse ära palju vigu.

Kuna kitsenduste grammatika korpuses on esitatud vaid lausete pindsüntaktiline analüüs, siis kaalusime lause süvastruktuuri esitamiseks mõne teise formalismi valimist. 2003. aasta aprillis ühines rühm eesti keele arvutisüntaksiga tegelevaid inimesi Tartu ülikoolist teaduskoostöövõrgustikuga *Nordic Treebank Network*[8]. Paljud selles koostöövõrgustikus osalevad töörühmad polnud enne projektiga liitumist veel teinud oma valikut planeeritava "rahvusliku" puude panga märgendamisskeemi osas. Erandiks oli siin projekt VISL Lõuna-Taani Ülikoolist, kellel oli oma märgendamisskeem, tarkvaralised vahendid selles esitatud korpuste kasutamiseks (veebipõhine päringuvahend, puude visualiseerimisvahend ning mitmed *online*-mängud keele süntaksi ja morfoloogia tundmaõppimiseks) ning kogemused selle märgendamisskeemi rakendamisel umbes 20 keelele. VISL märgendamisskeem (Bick 2003) kombineerib fraasistruktuuri- ja

[8] http://w3.msi.vxu.se/~nivre/research/nt.html

sõltuvuspuu head omadused. Igal tipul on lausepuus kaks märgendit – vormimärgend osalause-, fraasi- ja sõnaliikide tähistamiseks (nt fcl = finiitne osalause, NP = noomenifraas, VP = verbifraas, n = noomen, adj = adjektiiv jne) ja funktsioonimärgend sõnade süntaktiliste funktsioonide märkimiseks (nt osalausetasandil S = alus e subjekt, O = sihitis e objekt, P = öeldis e predikaat ja fraasisisesed funktsioonid, H = fraasipõhi ja D = fraasipõhja laiend).

Eesti keele puude panka *Arborest*[9] luuakse poolautomaatselt, lähtudes kitsenduste grammatika märgendusest, mis määrab sõnade süntaktilised funktsioonid lauses. Lisaks toetuvad CG→PSG (kitsenduste grammatikast fraasistruktuuri tuletavad) reeglid kirjavahemärkidele, leksikaalsele ja morfoloogilisele infole. Meetodit on varem kasutatud ka teiste keelte puhul (Bick 2003). Puude panga Arborest maht on hetkel 2500 lauset, neist 149 on käsitsi kontrollitud. Automaatselt genereeritud puude esialgne hindamine näitas, et 40% puudest oli õige struktuuriga (Bick jt 2004). Õigesti analüüsitud lausestruktuurid olid üldistatult järgmised:

1) lihtlaused, kus esinesid alus, öeldis (ja sihitis) mistahes järjekorras ning võib-olla ka määrusi ja predikatiive mistahes positsioonidel;
2) tüüpi (1) osalausetest moodustatud rindlaused (osalausete vahel rinnastav sidesõna *ja*, *ning*, *või*, *ehk* või koma);
3) põimlaused, kus kõrvallause oli määruse või sihitise funktsioonis.

Peamised vigu põhjustavad konstruktsioonid olid vahetult nimisõnafraasi ees esinev määrsõna, mis ekslikult loeti omadussõna laiendiks, lauselühendiga laused, keerulise struktuuriga nimisõnafraasid ja põimlaused, millel oli rohkem kui üks sõltuv kõrvallause (kitsenduste grammatika märgendus ei anna informatsiooni kõrvallausete hierarhia kohta).

Parandamaks pindsüntaktiliselt struktuurilt süvastruktuurile automaatset üleminekut, täiendati CG->PSG reeglite leksikaalset infot ning toodi sisse hulk vahemärgendeid. Eriti suur vajadus oli adverbiaali märgendi lahutamise järele mitmeks alaliigiks: lausetasandi vaba adverbiaal (fA), fraasisisene määruslik täiend (DA), rektsiooniadverbiaal ehk verbi seotud laiend, mis ei esine sihitise käänetes (Av), ja subjektiadverbiaal, mis on sisuliselt nagu öeldistäide, aga

[9] http://corp.hum.sdu.dk/arborest.html

mis ei vasta küsimustele *kes*? *mis*? ega *missugune*?, vaid on määrus, näidates aega, kohta, viisi jm (As).

Lisaks puude pangale *Arborest* on eesti keele jaoks loodud ka käsitsi märgendatud eksperimentaalne eesti keele e-õpet toetav puude pank[10], mis esialgu koosneb 100 lausest. Puude pank on ühildatud projekti VISL[11] (*Visual Interactive Syntax Learning*) käigus loodud interaktiivsete õppemängudega, mida saab üle veebi tasuta mängida. Andmebaasis olevaid lauseid saab vaadata nii teksti- kui ka graafilises formaadis. Puude pangas olevad laused on oma struktuurselt keerukuselt jaotatud kümnesse klassi. Mängu alustamisel on võimalik valida lauseid sobiva keerukusega klassist. Lahendatavateks ülesanneteks on sõnaliikide äratundmine (mängud "Shooting gallery", "Labyrinth", "Wordfall") ning süntaktiliste funktsioonide määramine (mäng "Space rescue").

Järgmiseks sammuks on eesti keele e-õppe praktikasse, s.o kooli viimine. Seni on mänge demonstreeritud TÜ teaduskeskuse "Ahhaa" keelenurgas.

4. Mitmekeelsed puude pangad

Kui ükskeelseid puude panku on praeguseks loodud juba mitmeid, siis paralleelpuudepanku, st süntaktiliselt märgendatud ja joondatud paralleelkorpusi, on teada suhteliselt vähe. Võib mainida Pennsylvania Ülikoolis statistilise masintõlke otstarbel loodavat inglise–korea paralleelpuudepanka, mis sisaldab üle 5000 lause (Han jt 2002). Üks suhteliselt väikesemahuline, kuid kümmet keelt hõlmav puude pank on Sofie paralleelpuudepank[12], millele pandi alus tänu koostöövõrgustikule *Nordic Treebank Network*. Sofie puude panga algmaterjaliks on kaks esimest peatükki Jostein Gaarderi romaanist "Sofie maailm". Paralleelpuudepankadest on kõige rohkem kasu siis, kui neis esinevad paralleelsed tekstid on võimalikult täpsed tõlked originaalist. Ilukirjanduse puhul ei saa seda muidugi alati tagada. Oluline on ka, et pangad oleksid esitatud ühes ja samas või kergesti üksteiseks teisendatavates formaatides ning märgendusskeemid toetuksid ühisele teoreetilisele alusele. Need nõuded peaksid olema täidetud, et pa-

[10] http://beta.visl.sdu.dk/visl/et
[11] http://beta.visl.sdu.dk
[12] http://omilia.uio.no/sofie

ralleelpuudepanka oleks hõlbus kasutada, nt viia selle baasil läbi võrdlevaid keeleuuringuid ning (piisavalt suure paralleelpuudepanga korral) treenida korpusepõhiseid meetodeid kasutavaid rakendusi, millest üks olulisemaid on näidetepõhine masintõlge (ingl *Example Based Machine Translation*). Sofie puude panga märgendamisskeemiks valiti *Nordic Treebank Networki* liikmete konsensusel märgendamisskeem VISL. Puude kodeerimise standardiks valiti TIGER XML formaat, mis on välja töötatud Stuttgardi Ülikoolis NEGRA puude panga projekti raames. Selline valik võimaldab paralleelpuudepanga ükskeelsete osade töötlemiseks kasutada redigeerimisvahendit *Annotate* ning päringu- ja visualiseerimisvahendit *TigerSearch*.

Nagu üldiselt paralleelkorpuste puhul, on ka paralleelpuudepankade jaoks oluline küsimus joondamine. Lausete joondamiseks on loodud üsna häid vahendeid, nt *Vanilla aligner*[13], kuid paralleelpuudepangad on seda väärtuslikumad, mida täpsemalt on lausest väiksemad struktuursed üksused, eeskätt fraasid, omavahel joondatud. Sõnade joondamine (ingl *word alignment*) on 1990. aastatel saanud populaarseks uurimisalaks seoses statistikapõhise masintõlke arenguga. Samas on fraaside joondamine parem kui sõnade joondamine, kuna tekstis esineb alati palju mitmesõnalisi üksusi, mida peaks tõlkima tervikuna (tõlkides sõna-sõnalt ja pannes tõlked kokku, saame hoopis erineva tähenduse), näiteks ühend- ja väljendverbid. Eriti oluline on aga vaadelda suuremaid üksusi, näiteks fraase, keelte puhul, mis on teineteisest väga erineva süntaktilise struktuuriga.

M. Volk ja Y. Samuelsson Stockholmi Ülikoolist tegid katse süntaktiliselt märgendatud paralleeltekste automaatselt fraasitasandil joondada (materjaliks esimene peatükk, 220 lauset Sofie paralleelpuudepanga saksa- ja rootsikeelsest osast), kasutades alusena J. Tiedemanni sõnade automaatse joondamise programmi (Tiedemann 2003) ning projekteerides seda fraasitasandile. Automaatsel fraaside joondamisel oli saagis 65% ja täpsus 79% (Volk, Samuelsson 2004).

Krista Liin (Liin 2005) tegi katse joondada automaatselt nimisõnafraase eesti ja saksa keeles, kasutades samuti Sofie paralleelpuudepanka, kuid väiksemat osa sellest (53 lauset). Selle eksperimendi juures kasutati nimisõnafraasipõhjade tõlkimist veebisõnasti-

[13] http://nl.ijs.si/telri/Vanilla/doc/ljubljana/

ku abil ning fraaside asukohainfot lauses. Automaatsel nimisõnafraaside joondamisel oli saagis 53% ja täpsus 84%.

Näidetepõhine masintõlge kasutab tõlkimisel katkeid inimese poolt tõlgitud tekstist, seega on tema töö tulemusel eeldusi sarnaneda inimtõlkega rohkem kui seda on reeglipõhisel masintõlkesüsteemil, millel on parimal juhul küll kõrge lingvistiline kompetents, kuid puudub ettekujutus sellest, kuidas inimesed tegelikult tõlgivad.

5. Kokkuvõte

Puude pangad on tänapäeval kiiresti arenev valdkond. Kahtlemata on selles valdkonnas eelis nn suurtel keeltel, mis on paremini varustatud keeleressurssidega (suured korpused, automaatanalüüsi vahendid jne). Samal ajal saavad väikesed keeled suurtel keeltel tehtud uurimistöö tulemusi ära kasutada – üle võtta puude pankade poolautomaatse loomise tehnikaid ja meetodeid, puude pankade loomise ja kasutamise tarkvaralisi vahendeid (eeldab märgenduse vastavust teatavale standardile), kanda märgendust paralleelpuudepangas ühelt keelelt teisele üle jne. Puude panga loomist ei saa täielikult automatiseerida, kuid poolautomaatne märgendamine kergendab tööd tunduvalt. Seda teed on mindud ka eesti keele puude panga *Arborest* loomisel, mida genereeritakse kitsenduste grammatika korpusest (mida samuti luuakse poolautomaatselt). Kitsenduste grammatikast süntaktilisele süvastruktuurile ülemineku reeglid ning abimärgendite loend on täiendamisel ja loodetavasti võimaldab see edaspidi eesti keele puude panga jõudsamat kasvu.

Näidetepõhist masintõlget, mida nähakse paralleelpuudepankade ühe kõige olulisema kasutusalana, peetakse hetkel kolmest põhilisest masintõlke meetodist (reeglipõhine, statistikapõhine ja näidetepõhine) kõige perspektiivikamaks. Selle meetodi edukaima rakendamise eelduseks on fraasitasandil joondatud paralleelpuudepankade olemasolu, mistõttu on puude pankade kiire kasvatamise kõrval suureks väljakutseks fraaside automaatse joondamise meetodite väljatöötamine erinevate keeltepaaride vahel.

Kirjandus

Abeillé, Anne (ed) 2003. Building and Using Parsed Corpora. Text, Speech and Language Technology, Vol 20. Dordrecht: Kluwer.

Bick, Eckhard, Uibo, Heli, Müürisep, Kaili 2004. Arborest – a VISL-style treebank derived from an Estonian Constraint Grammar corpus. – Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). Ed by S. Kübler, J. Nivre, E. Hinrichs, H. Wunsch. Tübingen, 1–14.

Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang, Smith, George 2002. The TIGER Treebank. – Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT 2002), Sozopol, Bulgaria, 24–42.

Ellegård, Alvar 1978. The syntactic structure of English texts: A computer-based study of four kinds of text in the Brown University Corpus. – Gothenburg Studies in English 43. Göteborg.

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi, Palmer, Martha 2002. Penn Korean Treebank: Development and Evaluation. – Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation. The Korean Society for Language and Information.

Järvinen, Timo 2003. Multi-layered Annotation Scheme for Treebank Annotation. – Proceedings of the Second Workshop on Treebanks and Linguistic Theories, Växjö University Press, 93–104.

Kingsbury, Paul, Palmer, Martha 2003. PropBank: the Next Level of TreeBank. – Proceedings of the Second Workshop on Treebanks and Linguistic Theories, Växjö University Press, 105–116.

Klein, Dan, Manning, Christopher 2004. Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency. – Proceedings of the 42nd Annual Meeting of the ACL.

Leech, Geoffrey 1997. Introducing Corpus Annotation. – Corpus Annotation. Ed by R. Garside, G. Leech, A. McEnery. London/New York: Longman, 1–18.

Leech, Geoffrey, Eyes, Elizabeth 1997. Syntactic Annotation: Treebanks. – Corpus Annotation. Ed by R. Garside, G. Leech, A. McEnery. London/New York: Longman, 34–52.

Liin, Krista 2005. Eesti–saksa paralleelpuudepanga paralleelistamine fraasitasandil. Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut.

Montemagni, Simonetta, Barsotti, Francesco, Battista, Marco, Calzolari, Nicoletta, Corazzari, Ornella, Lenci, Alessandro, Zampolli, Antonio

2003. Building the Italian Syntactic-Semantic Treebank. – Abeillé 2003, 189–210.

Müürisep, Kaili, Puolakainen, Tiina, Muischnek, Kadri, Koit, Mare, Roosmaa, Tiit, Uibo, Heli 2003. A New Language for Constraint Grammar: Estonian. – Proceedings of International Conference "Recent Advances in Natural Language Processing". Borovets, Bulgaria, 304–310.

Nivre, Joakim, de Smedt, Koenraad, Volk, Martin 2005. Treebanking in Northern Europe: A White Paper. – Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004. Ed by H.Holmboe. Copenhagen: Museum Tusculanums Forlag, 97–112.

Roosmaa, Tiit, Koit, Mare, Muischnek, Kadri, Müürisep, Kaili, Puolakainen, Tiina, Uibo, Heli 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? – Keel ja Kirjandus 3, 192–209.

Sampson, Geoffrey 2001. Empirical Linguistics. CONTINUUM, London, New York.

Taylor, Ann, Marcus, Mitchell P., Santorini, Beatrice 2003. The Penn Treebank: an overview. – Abeillé 2003, 5–22.

Tiedemann, Jörg 2003. Recycling Translations. – Extraction of Lexical Data from Parallel Corpora and Their Application in Natural Language Processing. Acta Universtitatis Upsaliensis, Uppsala University.

Volk, Martin, Gustafson-Capková, Sofia, Hagstrand, David, Uibo, Heli 2005. Teaching Treebanking. – Nordisk Sprogteknologi 2004. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004. Ed by H. Holmboe. Copenhagen: Museum Tusculanums Forlag, 143–159.

Volk, Martin, Samuelsson, Yvonne 2004. Bootstrapping Parallel Treebanks. – Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC) at COLING. Geneva, 63–69.

Eesti keele verbikesksete püsiühendite nominaalsetest komponentidest

Kadri Muischnek
Tartu Ülikool

1. Sissejuhatus

Loomuliku keele automaattöötlusel vajab lahendamist ka mitmesõnaliste üksuste ehk püsiühendite probleem. Näiteks analüüsib eesti keele kitsenduste grammatika süntaksianalüsaator ilma püsiühendite tuvastamiseta lauses *Jüri sai oma veast aru* substantiivi partitiivivormi *aru* verbi *saama* objektiks (vrd *Jüri sai poest hea raamatu*), kuigi semantilisel tasandil on tegevusobjektiks substantiivi *viga* elatiivivorm *veast*. Võib-olla ei maksagi süntaksianalüsaatorilt oodata tegevusobjektide tuvastamist, kuid antud näites sõnavormi *aru* objektiks analüüsimine viib edasise tõlgenduse eksiteele. Seda viga aitab vältida ühendi *aru saama* äratundmine teatud tüüpi mitmesõnalise üksusena.

Mitmesõnalised üksused võivad oma süntaktiliselt struktuurilt olla nii noomenifraasid – *Egiptuse nuhtlus, löök allapoole vööd*, adverbifraasid – *läbi ja lõhki, maani täis,* adpositsioonifraasid – (kellegi) *käe läbi, metsa poole* kui ka verbi ja tema seotud laiendi püsivad ühendid – *jalga laskma, läbi saama, kõnet pidama.*

Tartu Ülikooli arvutilingvistika uurimisrühmas on koostamisel verbikesksete püsiühendite andmebaas (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html; vt Kaalep, Muischnek 2003). Verbikesksete püsiühendite all on mõeldud verbide püsivaid ühendeid nii adverbide (ühendverbid) kui ka noomenitega, kusjuures verbi *olema* ühendid on (vähemalt praegusel etapil) vaatluse alt välja jäetud. Noomenist ja verbist koosnevat öeldisena toimida võivat püsiühendit nimetatakse eesti keele kirjeldustes väljendverbiks. Kuid kuna väljendverbi iseloomustab idiomaatilisus (EKG II: 20, Rätsep 1978: 25), aga mitte kõiki verbikesksete püsiühendite andmebaasis olevaid noomeni(fraasi) või adpositsioonifraasi ja verbi ühendeid ei saa idiomaatilistena kirjeldada, siis nimetataksegi neid koos ühendverbidega verbikesksteks püsiühenditeks. Selliseid keelendeid on eelnevalt kirjeldatud artiklis (Muischnek 2004b), kus neile on lähenetud verbikeskselt,

käesolevas kirjutises keskendutakse nimetatud püsiühendite nominaalsetele komponentidele.

1.1. Kasutatud materjal

Nagu öeldud, põhineb käesolev kirjutis peamiselt verbikesksete püsiühendite andmebaasil. Selleks, et teha lemmade ja käänete statistikat, ühestati andmebaasi väljendid morfoloogiliselt, st igale sõnavormile lisati tema algvorm ja morfoloogiline analüüs – sõnaliigi märgend ja käändsõnade puhul käände- ja arvukategooria. Kõik selle sõnavormi võimalikud analüüsid lisati automaatselt kasutades programmi Estmorf (selle kohta vt Kaalep, Vaino 2000), kuid konkreetses ühendis ainsa õige analüüsi väljavalimine – ühestamine – toimus käsitsi. Käsitsi ühestamise käigus muudeti ka mõningaid Estmorfi märgendeid. Tähtsaima muutusena asendati osa partitiivi analüüse märgendiga "objekti kääne", mida on kasutatud siis, kui püsiühendi nominaalne komponent saab tekstis muutuda vastavalt objekti käändevahelduse reeglitele. Nimelt esitatakse sõnaraamatutes verbist ja tema objektist koosnevate ühendite nominaalsed komponendid alati partitiivis, nt ühend, mis andmebaasis (ja selle aluseks olnud sõnastikes) on esitatud kujul *joont alla tõmbama* võib tekstis esineda ka kujul *tõmbab joone alla* või *joon tõmmatakse alla*.

Estmorf annab ainsuse lühikese illatiivi vormis olevale noomenile aditiivi märgendi. Käsitsi ühestamisel võeti kasutusele ka mitmuse aditiivi analüüs, nt ühendites *paneb käed raudu* või *jääb jalgu*.

Andmebaasi on ühestatud maksimaalselt "käändsõnasõbralikult", st kui oli vähegi võimalik sõnet noomenivormiks (mitte muutumatuks sõnaks) analüüsida, siis seda ka tehti. Nii on sõnavorm *kätte* ühendis *kätte saama* analüüsitud noomeni *käsi* vormiks, kuid sõnavorm *peale* ühendis *peale käima* on siiski saanud adverbi analüüsi.

Erinevalt andmebaasi üheks aluseks olnud "Fraseoloogiasõnastikust" (Õim 1993) on ühe tähendusega väljendi erinevad leksikaalsed variandid andmebaasis esitatud eri kirjetena, kuna andmebaasi esmane eesmärk on nende väljendite automaattöötlus. Nii on "Fraseoloogiasõnastiku" eessõnas esitatud järgmine põhimõte: "Kui komponentide asendamisel ei muutu püsiväljendi struktuur ja asendatavad komponendid on sünonüümid või sama semantilise välja sõnad, siis kahest osisest koosnev väljend loetakse variandiks ning esitatakse sama märksõna all (nt *kukalt kratsima, kukalt sügama*)." (Õim 1993: 7). Käes-

oleva artikli aluseks olevas püsiühendite andmebaasis aga on väljendid *kukalt kratsima* ja *kukalt sügama* eraldi kirjetes. Enamgi veel, eraldi kirjetes on andmebaasis ka üksteisest vaid morfosüntaktilise kuju poolest erinevad väljendid, näiteks *südamele hakkama* ja *südame peale hakkama*. Ent selliste paralleelvormide hulk ei ole nii suur, et mõjutaks oluliselt lemmade ja käänete statistikat.

Käesolevas artiklis jäävad lähema vaatluse alt välja ühendverbid, st ainult (afiksaal)adverbist ja verbist koosnevad püsiühendid, püsiühendite verbide sagedusloendis on aga ka nendega arvestatud.

Artiklis esitatud näitelaused pärinevad Eesti Ekspressi korpusest (http://test.cl.ut.ee/korpused/segakorpus/ekspress/).

2. Verbikesksete püsiühendite struktuur ja komponendid

Enamik verbikesksetest püsiühenditest on kahekomponendilised, koosnedes kas noomenist ja verbist (*ajab vihale, vaagub hinge, teeb kindlaks, jätab külmaks, sõidab nelja*), noomenifraasist ja verbist (*pääseb kerge nahaga, leiab märja haua, mängib kahe otsaga mängu, koorib kümme nahka*) või adpositsioonifraasist ja verbist (*viskab hinge alla, ripub juuksekarva küljes, jääb kahe tule vahele*). Kolmekomponendilised püsiühendid koosnevad verbist ja selle kahest argumendist, milleks võivad olla noomenid või noomenifraasid, adverbid ja adpositsioonifraasid, nt *ajab hirmu nahka, läheb riski peale välja, saab kindla pinna jalge alla, kütab kõrvad kuumaks, tuleb puhtalt välja* jms. Andmebaasis esineb väga vähesel määral ka neljakomponendilisi püsiühendeid, mis koosnevad verbist ja selle kolmest argumendist, nt *paneb endale köie kaela*.

2.1. Sagedasemad verbid

Nagu kirjeldatud artiklis (Muischnek 2004b), moodustavad verbidest enim püsiühendeid sellised üldise tähendusega polüseemsed verbid nagu *tegema, minema, panema, võtma, ajama, saama, andma, lööma, jääma* ja *pidama*. Marja Nenonen (2002: 85) kirjeldab soome keele idioome järgmiselt: "suomen kielen verbilliset idiomit koostuvat tyypillisesti yleisista, tutuista perusverbeista sekä taivutetuista nomineista." Soome keele idioomide sagedasemad verbid on tema materjali põhjal *olla, otta, saada, mennä, pitää, vetää, tulla, tehdä, käydä, panna, lähteä, antaa, pistää, heittää, päästää* ja *jääda*, mis,

välja arvatud *antaa, vetää, heittää* ja *pistää*, kuuluvad ka soome keele 30 sagedasema verbi hulka (Nenonen 2002: 85). Ka siin loetletud eesti verbidest kuuluvad kõik, v.a *ajama* ja *lööma,* "Eesti kirjakeele sagedussõnastiku" (Kaalep, Muischnek 2002) 30 sagedasema verbi hulka. Kõige sagedasemad püsiühendeid moodustavad verbid, v.a verb *lööma*, kuuluvad ka eesti keele tuumverbide hulka. Tuumverbe on defineeritud kui selliseid tegusõnu, mida kasutatakse grammatilistes funktsioonides ja/või mis väljendavad üldisi mõisteid (täpsemalt vt Tragel 2002, Pajusalu jt 2004).

2.2. Sagedasemad noomenid

Püsiühendites kõige sagedamini esinevad noomenid ei ole (kirjalikus) keeles kõige sagedasemad. Soome keele idioomides esinevate käändsõnade kohta on Marja Nenonen väitnud: "Verbillisissä idiomeissa käytetään paljon ruuminosannimiä, kommunikaatioon liittyviä sanoja ja muihin yleisen tason käsitteisiin viittaavia sanoja." (Nenonen 2002: 114). Kõige sagedasemad noomenid soome keele idioomides (mitte ainult verbikesksetes idioomides) on Nenose järgi *silmä, pää, mieli,suu, naama, aika, asia, korva, turpa, sana, niska.* Eesti keele verbikesksetes püsiühendites kõige sagedamini esinevad noomenid on esitatud tabelis 1. Põgusalt on käsitletud verbikesksete püsiühendite komponente ka preprindis (Muischnek 2004a), kuid nimetatud käsitluses olid vaatluse all ainult kahekomponendilised verbikesksed püsiühendid, siin tabelis on aga esitatud kõigis verbikesksetes püsiühendites sageli esinevad noomenid.

Tabel 1. 20 enim verbikeskseid püsiühendeid moodustavat noomenit

Noomen (sulgudes on esinemiste arv püsiühendite andmebaasis)	**Sagedus kirjakeeles** (jrk nr noomenite sagedusloendis)	**Polüseemia**
käsi (228)	14.	6
silm (203)	22.	4
pea (187)	20.	9
suu (122)	112.	7
süda (114)	133.	5
jalg (113)	70.	4
nahk (84)	475.	5
tee (77)	31.	7

Noomen (sulgudes on esinemiste arv püsiühendite andmebaasis)	**Sagedus kirjakeeles** (jrk nr noomenite sagedusloendis)	**Polüseemia**
nina (73)	363.	5
kõrv (70)	207.	3
ots (69)	250.	10
kael (69)	356.	4
hing (69)	121.	6
sõna (66)	10.	8
hammas (66)	310.	3
selg (61)	89.	5
nägu (58)	43.	5
keel (56)	46.	7
tuli (53)	105.	8
elu (52)	25.	9

Arv veerus "Sagedus kirjakeeles" näitab järjekorranumbrit "Eesti kirjakeele sagedussõnastiku" (Kaalep, Muischnek 2002) põhjal tehtud noomenite sagedusloendis. Lemma *pea* sageduses sisaldub sagedussõnaraamatus ka adverbi *pea* (nt *üsna pea*) sagedus. Arv lahtris "Polüraamatus" (EKSS). Püsiühendite nominaalsed komponendid on vähem polüseemsed kui sageli püsiühendeid moodustavad verbid – "Eesti kirjakeele seletussõnaraamatus" on verbil *tegema* 18 tähendust, verbil *minema* 13, *panema* – 7, *ajama* – 16, *saama* – 12, *andma* – 10, *lööma* – 8, *jääma* – 3, *pidama* – 17.

Nendest kahekümnest sõnast neliteist (= 70%) on somaatilised (kui sõna *hing* mitte somaatiliseks lugeda). Sõnad, mis esinevad püsiühendite nominaalsete komponentidena, ei ole sagedusloendi absoluutse tipu sõnad, aga siiski kuulub 10 neist 100 sagedasema käändsõna hulka. Kõik või vähemalt enamik neist noomenitest väljendavad põhitasandi mõisteid. Võrdluseks: "Eesti kirjakeele sagedussõnastikus" on 20 sagedasemat noomenit sageduse järjekorras järgmised: *aasta, mees, aeg, inimene, sõna, kord, pool, naine, käsi, päev, kroon, laps, asi, pea, riik, silm, töö, elu, raha, linn, tee*. Püsiühendites esinevad neist sageli *aeg* (26x), *sõna* (66x), *käsi* (228x), *pea* (187x), *silm* (203x) ja *elu* (52x).

Sagedasimad omadussõnad püsiühendites on *hea* (34 esinemist), nt ühendites *arvab heaks, laseb hea maitsta, seisab heas kirjas*; *tühi* (30), nt *ajab tühja juttu, läks tühja*; *õige* (26), nt *keegi mõis-*

tetakse õigeks, näitab õiget nägu; *külm* (21), nt miski *jätab külmaks, säilitab külma vere*; *kuum* (21), nt kellelgi *köetakse kere kuumaks*; *halb* (21), nt midagi *pannakse halvaks*; *tuline* (21) ja *puhas* (20).

Tabelist 1 ja püsiühendites sageli esinevate verbide loendist võib järeldada, et tüüpiline eesti keele verbikeskne püsiühend koosneb sagedasest polüseemsest verbist, mis kuulub ka tuumverbide hulka, ja morfoloogiliselt lihtsast somaatilisest või muid "inimlähedasi" mõisteid väljendavast substantiivist.

Somaatiliste sõnade sagedust idiomaatilistes ühendites on kirjeldatud tüpoloogiliselt sagedase nähtusena. Nii väidab František Čermak: "...it seems that there is no escaping the field of somatic idioms once any attempt is made at a more general description of idioms in any language." (... paistab, et niipea, kui püütakse üldisemalt kirjeldada mingi keele idioome, pole pääsu somaatiliste idioomidega seotud temaatikast) (Čermak 1995: 109).

Minoji Akimoto võrdleb somaatiliste sõnade esinemist verbikesksetes idioomides inglise, jaapani, prantsuse ja saksa keeles. Tema andmetel (Akimoto 1994: 461) moodustavad vaadeldud keelte (v.a jaapani keele) idioomides somaatiliste sõnade esikolmiku *käsi*, *pea* ja *silm* – nagu eesti ja soome (Nenonen 2002: 115) keelteski.

Feliks Vakk (1970: 10) juhib tähelepanu somaatilise fraseoloogia ebaühtlasele jaotumisele kehaosade vahel, kusjuures kõige rohkem metafoore seostub olulisemate kehaosade nagu pea, silmade, käte, jalgade ja südamega. F. Vakk eristab somaatiliste fraseologismide hulgas ka väljendusliigutusi kirjeldavaid ühendeid (*nina norgu laskma, pead püsti hoidma*), mille puhul „võib vahetut seost keelelise väljendi ja vastava lähtepildi vahel jälgida igal sammul" (Vakk 1970: 259).

Püsiühendite sagedasemad nominaalsed komponendid on tuntud ka keelenditena, mis tüpoloogiliselt sageli grammatikaliseeruma kalduvad. Bernd Heine ja Tania Kuteva "Grammatikalisatsioonileksikonis" (2002) leiame grammatikaliseerumisallikate (*source*) loendist ca 15 somaatilist sõna (tähendust). Eesti (ja soome) somaatiliste sõnade grammatikaliseerumist on lähemalt käsitlenud nt Krista Ojutkangas (2000).

2.3. Püsiühendite ainukordsed komponendid

Püsiühendite ainukordsed komponendid ehk idiomaatilised isolaadid on sellised sõnavormid, mille esinemiskontekst on äärmiselt piiratud: neid kasutatakse ainult teatud kindlate püsiühendite koosseisus, nt *annab/palub/saab andeks*, *veab kihla*, *lööb lulli* või *luuslanki*, *annab/leiab/saab mahti*, *annab/saab peksa*, *paneb/pistab/pääseb plehku*, *teeb putket*, *annab rooki*, *laseb sihku*, *annab/saab/teeb säru*, *saab/teeb tuupi*, *paneb tähele*, *teeb vehkat*.

Verbikeskseteüsiühendite ainukordsed komponendid esinevad ainult koos ühe või maksimaalselt paari-kolme verbiga. Ülle Viksi "Väikeses vormisõnastikus" (Viks 1992: 15) nimetatakse neid "sõnaühendi seotud komponentideks" ja sõnastikuosas ei saa nad endale sõnaliigi märgendit. Nendel sõnadel on olemas ainult esitatud sõnavormid, nii et tegu peaks olema muutumatute sõnadega. Ometi on eesti keelt emakeelena kõneleja võimeline enamiku selliste sõnavormide käändekategooria intuitiivselt ära tundma. Verbi juurde kuuluva sõna käändevormilise, põhiliselt partitiivse tõlgenduse võib põhjustada püsiühendi transitiivne verb, kuid ei saa eitada, et need sõnavormid sarnanevad oma morfofonoloogilise kuju poolest käändevormidele, põhiliselt partitiivile (nt *kihla, lulli, luuslanki, mahti, putket, vehkat*[1]), ka aditiivile (*plehku*) või translatiivile (*andeks*).

Arvo Krikmann on sarnase probleemiga kohtunud eesti keele löömist ja peksmist märkivaid väljendeid vaadeldes (Krikmann 2004: 144–145). Kirjeldades seoses verbiga *andma* esinevaid sõnavorme *peksa*[2], *kolki*, *parki* ja *rooki*, ütleb ta nende sõnaliigilisuse ja lauseliikmelisuse kohta, et see on "mingit hämarat "adverbiaalset" laadi" (2004: 144). Samas juhib ta tähelepanu sellele, et selline fonotaktiline mall – kahesilbiline kolmandavälteline sõna – seostub verbidega *andma* ja *saama* (nii peksmise-teemalistes fraseologismides

[1] Mägiste (1983) seostab vormi *vehkat* verbiga *vehkima* ja oletab, et vorm on tuletatud vene infinitiivi lõpu *-at* mõjul või eeskujul, kuid võib siiski arvata, et tänapäeva keelekasutaja seostab lõppu *-t* pigem partitiiviga.

[2] *peksa*-vormi kohta kinnitab A. Krikmann, et seda ei saa mingil juhul kujutleda taanduvat substantiivi partitiivile, kuna *peksa* on normaalse da-infinitiivse murdevormina domineeriv kogu läänepoolses Eestis ja paralleelselt *peksta*-vormidega ka mitmetel kesk- ja idamurde aladel.

kui ka väljaspool neid) nii sageli, et "on omandanud teatava püsimalli kvaliteedi" (2004: 144). Tundub, et eesti keele materjali põhjal ei saa nõustuda Wolfgang Fleischeri väitega (1982: 45), nagu võiks (saksa keeles) idiomaatilise isolaadina (Fleischeril: unikaalse komponendina) esinev substantiiv oma substantiivsuse täielikult kaotada.

Marja Nenonen ja Jussi Niemi (1999: 159) ütlevad soome keele samalaadset nähtust kirjeldades, et [idiomaatilised isolaadid on] "noun-like case-inflected complements that do not appear outside of the idioms at all" (substantiivilaadsed käändevormis laiendid, mis ei esine üldse väljaspool idioome). Nad tõstatavad küsimuse, kas selliseid idiomaatilisi isolaate nagu *mönkään* ühendis *meni mönkään* 'ebaõnnestus' tajutakse pigem morfoloogiliste isolaatidena, (*cranberry morphs*) või tavaliste noomenitena. Artiklis kirjeldatud tajukatsete (*lexical decision experiments*) tulemusena väidavadki autorid, et soome keele kõnelejate jaoks on nendel sõnavormidel olemas nominatiivivormid ja üldse on idiomaatilised isolaadid talletatud mentaalses leksikonis sarnaselt tavaliste noomenitega.

Wiedemanni sõnaraamatule tuginedes võib väita, et suur osa eesti keele sellistest sõnavormidest on varasema täisparadigmaga noomeni püsiühendi koosseisus säilinud jäänukvormid. Näiteks on nimetatud sõnaraamatus olemas märksõnad

lull G.lulli 'Vergnügen im Freien', (vrd tänapäeva lööb lulli)
maht G. mahi, mahu 'Macht, Gewalt, Vermögen, Freiheit' (vrd tänapäeva saab mahti)
plehk G. plehu, davon plehku panema (tõlge puudub)
putk G. putku 'Flucht' (vrd paneb/pääseb putku)
tūp G. tūbi 'Stoss, Schub, Schlag' tūpi sāma

Mõnda tänapäeval ainult ühe püsiühendi koosseisus esinevat sõnavormi sai varem Wiedemanni andmetel kombineerida mitme verbiga, nt lisaks väljendile *tähele panema* on märksõna *täht* juures esitatud ka ühend *tähele saama 'bekannt werden'*, märksõna *soik* juurest leiame lisaks väljendile *soiku jääma* ka *on soigus* ning *putk* juures ka *on putkus*.

Rosamund Moon on kirjeldanud idiomaatilisi isolaate (*cranberry collocations*) inglise keeles ja temagi näitab, et paljud tänapäeva inglise keeles ainult teatud püsiühendites esinevad sõnavormid on keele varasemal arenguperioodil olnud laiemas kasutuses (Moon 1998: 78–80). Sama väidab Fleischer saksa keele kohta (1982: 46).

Idiomaatilistele isolaatidele lähedased on püsiühendite koosseisus esinevad noomenivormid, millel on teoreetiliselt olemas ja tänapäeva keelekasutust kajastavates sõnaraamatutes esitatud ka nominatiivivorm, aga tegelikult kasutatakse tänapäeval ainult teatud käändevormi ühenduses teatud verbidega. Selline sõnavorm võib olla seega tulevane (või tegelikus keelekasutuses juba praegune) idiomaatiline isolaat. Nii on noomenist *lokk: loku* tänapäeva keeles kasutusel ainult partitiivivorm väljendverbi *lööb lokku* koosseisus (vt ka Muischnek 2004b: 586–587).

Teine võimalus sellise sõnavormi saamiseks on laenamine (vt ka Fleischer 1982: 45, Moon 1998: 78). Näiteks on malai keelest laenatud eesti keelde ühendi *jookseb amokki* nominaalne komponent. Sellel sõnavormil on "Eesti Kirjakeele Seletussõnaraamatu" järgi kaks nominatiivi vormi – kõnekeelne *amokk* ja korrektsem *amok*, ometi on vähemalt üldkeeles kasutusel peamiselt ainsuse partitiivi vorm ühenduses verbiga *jooksma*. "Eesti Ekspressi" korpuses esineb see tüvi vaid liitsõnas ja väljendverbis:

(1)
Ta jookseb amokki ühest telefoniautomaadist teise ning valib 002.

(2)
Enne kui selgitused-seletused inimesteni jõuavad on kunst oma amokijooksus jälle kaugenenud.

2.4. Adpositsioonifraasid püsiühendites

Tabel 2 esitab püsiühendites sageli esinevad adpositsioonid.

Kolmandas veerus on võrdluseks esitatud adpositsioonide esinemine morfoloogiliselt ühestatud korpuses, mitte "Eesti kirjakeele sagedussõnastikus", sest viimases pole eristatud muutumatute sõnade adpositsioonilist ja adverbilist kasutust. 10 sagedasemat adpositsiooni morfoloogiliselt ühestatud korpuses on *pärast, kohta, eest, üle, poolt, vastu, jooksul, enne, kuni, peale*. Kattuvad nendes kahes sagedusloendis ainult kaassõnad *üle, vastu* ja *peale*. Eesti keele tuumsõnade loendi koostajad (Pajusalu jt 2004: 12) on tuumsõnade hulka arvanud sõnad *alla, peale, üle, vastu, läbi* ja *eest* (tegemata vahet adverbilise ja adpositsioonilise kasutuse vahel), põhjuseks nende sõnade sagedus, polüseemsus ja tähenduse skemaatilisus. Nii et adpositsioonidest esinevad püsiühendites pigem kognitiivselt olulised kui väga sagedased.

Tabel 2. Kümme sagedasemat adpositsiooni verbikeskstes püsiühendites

Adpositsioon	Sagedus püsiühendite andmebaasis	Jrk nr adpositsioonide sagedusloendis (koostatud morfoloogiliselt ühestatud korpuse põhjal)
alla	96	24.
peale	92	10.
vastu	68	6.
mööda	33	31.
üle	30	4.
all	28	11.
läbi	20	26.
ette	20	38.
peal	19	45.
vahele	18	58.

Sageduselt esineb andmebaasis kõige rohkem prepositsioonifraasi *vastu muhku*, kuid seitse kaheksast esinemisest on tegelikult sama teema variatsioonid, st väljendid on enam-vähem samatähenduslikud – *vastu muhku antakse, käratatakse, pannakse, tõmmatakse, valatakse, virutatakse* ja *äiatakse*. Sama situatsiooni teisest vaatenurgast esitab väljend *saab vastu muhku*. Adpositsioonifraasid, mis moodustavad mitmeid erineva tähendusega väljendeid, on näiteks *südame peale* (*langeb südame peale, jääb südame peale, hakkab või käib südame peale*) või *läbi sõrmede (laseb läbi sõrmede, läheb läbi sõrmede, vaatab* või *vahib läbi sõrmede*). Rõhuv enamik adpositsioonifraase on siiski monofunktsionaalsed – kannavad ühte tähendust ja esinevad koos nendes ühendites sama tähendust väljendavate verbidega, nt *kerkib, kipub, tuleb, tungib keele peale*. Adpositsioonifraasiga püsiühendil võib olla tähenduslikult lähedane vaste, kus kaassõnafraasiga väljendatud tähendust kannab käändsõnavorm, nt *ajab hirmu naha vahele ~ ajab hirmu nahka, hakkab südame peale ~ hakkab südamele.*

3. Nominaalsete komponentide morfoloogiast

Verbikesksete püsiühendite nominaalsete komponentide hulgas domineerivad lühikesed ja lihtsad omasõnad – tuletusliidetega sõnu ja liitsõnu on nende hulgas vähe, tabelis 1 esitatud kahekümne noomeni hulgas on vaid ühe- ja kahesilbilised lihttüvelised sõnad.

Nimisõna tuletusliidetest on sagedasim eesti keele tavalisemaid substantiivisufikseid *-us*: *võtab arutusele, avaldab austust, toob avalikkuse ette, satub ärevusse.* Absoluutse produktiivsusega *mine*-sufiks on püsiühendite esikomponentide hulgas *us*-sufiksist kümneid kordi harvem ja tema abil tuletatud sõna on enamasti leksikaliseerunud: keegi *saadetakse asumisele* või idiomatiseerunud: keegi *saab vastu vahtimist.*

Sagedasimad omadussõnaliited on *-ne*, nt ühendis *jääb paikseks* ja *-line*, nt ühendis *näitab oma tõelist nägu.* Lisaks leidub vähesel määral *v*-partitsiipe omadussõna positsioonis, nagu näiteks väljendis *ulatab abistava käe* ja keskvõrdes omadussõnu, näiteks *läheb paremate jahimaadele.*

Liitsõnu sisaldab umbes 7% verbikesksetest püsiühenditest. Samas moodustavad liitsõnalised lemmad kõigist verbikesksetes püsiühendites esinevate noomenite lemmade hulgast umbes 25%, s.t kui väga sagedased käändsõnad nagu näiteks *käsi* või *süda* osalevad paljudes erinevates püsiühendites, siis enamik liitsõnu esineb vaid ühes püsiühendis. Liitsõnade hulgas leidub ka selliseid, mida võiks pidada idiomaatiliste isolaatide hulka kuuluvateks (vt osa 2.3). Need on sellised liitsõnad, mida kohtab vaid püsiühendite koosseisus seoses ühe või paari kindla verbiga. Näiteks ei kohta me noomenit *sõnasaba* mujal kui väljendites *sai sõnasabast kinni* ja *hakkas/haaras/võttis sõnasabast kinni,* liitsõna *õnnesärk* mujal kui ühendis *sündis õnnesärgis*; *eluküünalt* saab ainult *kustutada.* Mõned sellistest liitsõnadest moodustavad tugiverbiühenditele sarnaseid konstruktsioone – nad väljendavad mingit tegevust ja nendega kombineeruv verb ei lisa ühendi tähendusele eriti midagi, nt *seasörk* väljendis *laseb seasörki* või *kraapjalg* ühendis *teeb kraapjalga.*

3.1. Käändekategooria püsiühendites

Tabel 3 esitab käänete sagedused püsiühendite andmebaasis (selle morfoloogilise ühestamise põhimõtete kohta vt osa 1.2), tabelis 4 on võrdluseks käänete sagedused morfoloogiliselt ühestatud korpuses (http://test.cl.ut.ee/korpused/morfkorpus/). Käänete sageduse arvutamisel võeti arvesse noomenid, s.t substantiivid, adjektiivid, pronoomenid ja sõnadega kirjutatud numeraalid. Arvutuste tegemise hetkel sisaldas morfoloogiliselt ühestatud korpus ainult kirjaliku keele tekste. Nende tabelite omavahelisest võrdlusest ei saa aga teha eriti kau-

geleulatuvaid järeldusi, esitab ju üks neist leksikoni ja teine korpuse andmeid.

Tabel 3. Verbikesksete püsiühendite nominaalsete komponentide käänete sagedused püsiühendite andmebaasi põhjal

Käänded (sageduse järjekorras)	**Sõne** (token)[3]		**Sõnavorm** (type)		**Token /type**
	Arv	**%**	**Arv**	**%**	
Partitiiv	2691	29,6	1120	30,6	2,4
Muutub vastavalt objekti käändevahelduse reeglitele	1614	17,7	617	16,8	2,6
Aditiiv (lühike illatiiv)	1162	12,8	319	8,7	3,6
Genitiiv	950	10,4	321	8,8	3,0
Translatiiv	793	8,7	377	10,3	2,1
Allatiiv	549	6,0	224	6,1	2,5
Elatiiv	349	3,8	130	3,5	2,7
Illatiiv	262	2,9	117	3,2	2,2
Komitatiiv	225	2,5	110	3,0	2,0
Inessiiv	189	2,1	121	3,3	1,6
Nominatiiv	140	1,5	77	2,1	1,8
Adessiiv	111	1,2	88	2,4	1,3
Ablatiiv	34	0,4	19	0,5	1,8
Terminatiiv	22	0,2	17	0,5	1,3
Essiiv	6	0,1	6	0,2	1,0
Kokku	9097	100	3663	100	2,5

Tabel 4. Käänete sagedused morfoloogiliselt ühestatud korpuses

Käänded (sageduse järjekorras)	**Sõne** (token)		**Sõnavorm** (type)		**Token/ type**
	Arv	**%**	**Arv**	**%**	
Nominatiiv	91777	32,8	21354	28,8	4,3
Genitiiv	80015	28,6	16852	22,7	4,7
Partitiiv	36507	13,0	10754	14,5	3,4
Adessiiv	15020	5,4	3402	4,6	4,4
Inessiiv	13502	4,8	4126	5,6	3,3
Allatiiv	10165	3,6	3453	4,7	2,9
Elatiiv	9698	3,5	4319	5,8	2,2

[3] Sõne on sõnavormi iga konkreetne esinemisjuhtum, jooksev sõna.

Käänded (sageduse järjekorras)	**Sõne** (token)		**Sõnavorm** (type)		**Token/ type**
	Arv	**%**	**Arv**	**%**	
Translatiiv	7868	2,8	2847	3,8	2,8
Komitatiiv	7116	2,5	3281	4,4	2,2
Aditiiv	2744	1,0	912	1,2	3,0
Illatiiv	1863	0,7	901	1,2	2,1
Ablatiiv	1435	0,5	744	1,0	1,9
Essiiv	1040	0,4	722	1,0	1,4
Terminatiiv	672	0,2	356	0,5	1,9
Abessiiv	402	0,1	232	0,3	1,7
Kokku	279824	100	74255	100	3,8

Abessiiv, mis esines ainult ühes püsiühendis – *peavarjuta jääma* –, on tabelist 3 välja jäetud.

Tabelitest 3 ja 4 näeme, et morfoloogiliselt ühestatud korpuses on sagedasim kääne nominatiiv, millel on tihedalt kannul genitiiv ja siis järgneb umbes kaks korda vähem esinev partitiiv. Püsiühendite andmebaasis on nominatiiv alles 11. kohal, genitiiv neljandal ja partitiiv esimesel. Loomulikult on põhjuseks asjaolu, et nominatiiv on subjekti kääne, aga püsiühendite andmebaasis sisalduvad ainult verbi ja tema seotud laiendite ühendid. Genitiiv on tekstides objekti ja possessiivatribuudi kääne, samuti nõuab genitiivi enamik adpositsioone. Andmebaasis on genitiivis esineda võiv objekt märgendatud käändsõnana, mis muutub vastavalt objekti käändevahelduse reeglitele. Genitiiv verbikesksete püsiühendite andmebaasis on eelkõige adpositsioonifraasi kuuluva noomeni kääne, nt *saab hamba alla*, *vaatab läbi sõrmede* (ca 2/3 genitiivi esinemistest andmebaasis), ka atribuudi kääne, nt *läheb looja karja, jagab helde käega.*

Noomen, mis püsiühendite andmebaasis on märgendatud partitiivis esinevaks, on tekstides ainult partitiivis esinev objekt, s.t selline, mis ei muutu vastavalt objekti käändevahelduse reeglitele. Tekstis partitiivis olev noomen võib olla kas ainult partitiivis esineda võiv objekt, objekt, mis on partitiivis vastavalt objekti käändevahelduse reeglitele või ka kvantori laiend, partitiivsubjekt jms. Partitiivse objekti sagedust verbikesksetes püsiühendites saab seletada objekti määramatuse või mitte-referentsiaalsusega, nt väljendid *mattis hinge* või *leidis aset*, kuigi vahel saab ka mittereferentsiaalne objekti positsioonis olev noomen muutuda vastavalt objekti käändevahelduse

reeglitele (*heitis hinge* (gen), *ei heitnud hinge* (part); *pani pea* (gen) *tööle, ei pannud pead* (part) *tööle*).

Tabelite 3 ja 4 võrdlemisel äratab tähelepanu lühikese illatiivi (aditiivi) ja translatiivi suurem sagedus püsiühendites korpuse andmetega võrreldes. Kuna ka korpuses peaksid noomenid nendes käänetes olema eelkõige verbi laienditeks, siis võiks nende käänete suuremat sagedust pidada just püsiühenditele iseloomulikuks. Lühikesest illatiivist tuleb pikemalt juttu osas 3.2, translatiiv esineb püsiühendites peamiselt muutumis-, muutmis- ja resultatiivkonstruktsioonides (*läheb halliks, teeb püksid märjaks, hirnub* (ennast) *herneks, lööb* (raha, palga jms) *sirgeks*).

Nn kohakäänetest on püsiühendite andmebaasi käändsõnade hulgas korpusega võrreldes suurem osakaal lisaks aditiivile ka illatiivil ja allatiivil, korpusega võrreldes väiksem osakaal aga adessiivil ja inessiivil, mille vähesuse üks põhjusi peitub asjaolus, et verbi *olema* ühendid on verbikesksete püsiühendite andmebaasist välja jäetud.

3.2. Lühike illatiiv ehk aditiiv verbikesksetes püsiühendites

Asjaolu, mis tabelis 3 tähelepanu äratab, on aditiivi (lühikese illatiivi) kõrge kolmas koht püsiühendite nominaalsete komponentide hulgas. Siinkohal peab märkima, et käesolevas artiklis kasutatakse nimetust "aditiiv" lühikese illatiivi märkimiseks pelgalt lühiduse ja selguse mõttes, mitte selleks, et eristada illatiivi ja aditiivi kui kaht eraldi käänet. Püsiühendite andmebaasi morfoloogiliseks analüüsiks kasutatud programmi Estmorf aluseks on Ülle Viksi "Väike vormisõnastik" (Viks 1992). Selles on aditiivile "antud iseseisva muutevormi staatus, lähtudes eelkõige vormilistest kaalutlustest" (Viks 1992: 27).

Illatiiv on (ka kognitiivsete protsesside) suuna või sihi kääne ja sellisena ei oleks tema kohas sagedusloendis kohe objekti käänete järel midagi imelikku. Kuid, nagu tabelist 3 näha, on püsiühendite käänetest sageduselt kolmandal kohal just lühike illatiiv ehk aditiiv ja illatiivi pika vormi leiame alles kaheksandalt kohalt. Nii et vähemalt verbikesksete püsiühendite kohta kehtib küll Cornelius Hasselblatti väide, et "deskriptiivsest ja sünkroonilisest vaatevinklist on eesti keele sisseütlev ebaregulaarne" (Hasselblatt 2000: 803).

Morfoloogiliselt ühestatud korpuses on aditiiv samuti veidi sagedasem kui illatiiv. Kati Kio on oma bakalaureusetöös jõudnud järeldusele, et aditiivi vorm on tegelikus keelekasutuses palju sageda-

sem kui illatiivi vorm (Kio 2002: 92), samas töös on täheldatud ka aditiivi eelistamist illatiivile eriti just püsiühendites (Kio 2002: 55).

Tabeli 3 viimases veerus on toodud sõnede/sõnavormide suhe. Kogu andmebaasi nominaalsete komponentide keskmine on 2,5; suurim on 3,6 – on see suhe aditiivi puhul. See tähendab seda, et sama aditiivis sõnavormi kasutatakse korduvalt erinevates püsiühendites. Ka Kati Kio on oma töös märkinud, et mida sagedasem on sõna, seda rohkem kasutatakse sellest aditiivi vormi (Kio 2002: 54).

Kõige sagedasemad aditiivis käändsõnad verbikesksetes püsiühendites on ka tabelis 1 kõrgel kohal olnud *pea* (58 püsiühendit, sh *paneb aru pähe, võttis* midagi *pähe* jne), *käsi* (56 püsiühendit, sh *võtab ennast kätte,* miski *käib käest kätte,* midagi *võidetakse kätte* jne), *kael* (43 püsiühendit, sh ebameeldivad tööd *määritakse* kellelegi *kaela,* keegi *jääb* kellelegi *ristiks kaela,* keegi *paneb endale silmuse kaela* jne), *silm* (37 püsiühendit, sh kellelegi *aetakse/loobitakse/puistatakse puru silma,* keegi *paistab* või *hakkab silma,* miski *puutub* või *hakkab silma, toob vee silma* jne) jms somaatilisi mõisteid tähistavad sõnad, aga ka sellised adverbistuvad sõnavormid nagu *otsa* (keegi *sõidab* kellelegi *otsa,* keegi *lõpeb otsa*), *kimpu* (*jääb kimpu, ajab kimpu*), *kihva* (*keerab kihva*) või *ühte* (*hoiavad ühte*), mis võib olla ka täiendiks: *ühte väravasse mängima.*

Noomenid, mis esinevad aditiivis rohkem kui pooltes nende osalusel moodustatavates verbikesksetes püsiühendites, on esitatud tabelis 5.

Tabel 5. Püsiühendites sageli aditiivis (lühikeses sisseütlevas) esinevad noomenid

Noomen	Kokku	Aditiivis	
		Arv	%
pihk (pihku)	15	15	100
varn (varna)	6	6	100
lõks (lõksu)	6	6	100
korts (kortsu)	6	6	100
kimp (kimpu)	6	6	100
selts (seltsi)	4	4	100
lett (letti)	4	4	100
latv (latva)	4	4	100
kihv (kihva)	4	4	100

Noomen	Kokku	Aditiivis	
		Arv	%
jooks (jooksu)	4	4	100
toime (toime)	8	7	88
vang (vangi)	6	5	83
ork (orki)	6	5	83
mõrd (mõrda)	6	5	83
õu (õue)	5	4	80
perse (perse, persse)	14	11	79
õng (õnge)	11	8	73
rüpp (rüppe)	7	5	71
käik (käiku)	7	5	71
plats (platsi)	6	4	67
kraav (kraavi)	6	4	67
kodu (koju)	18	12	66
süli (sülle)	14	9	64
haud (hauda)	11	7	64
üks (ühte)	75	47	63
nurk (nurka)	8	5	63
kael (kaela)	69	43	62
põhi (põhja)	50	31	62
maja (majja)	7	4	57
muld (mulda)	11	6	55
põrm (põrmu)	13	7	54
pilv (pilve)	13	7	54
selg (selga)	61	32	52
rida (ritta)	10	5	50
tüli (tülli)	8	4	50

Tabelist 5 näeme, et on kümme sellist noomenit, mis esinevad verbikesksetes püsiühendites ainult aditiivis:

pihk *(naerab (endale) pihku, võtab kellegi kõrvad/karvad/kõri pihku, pistab kellelegi käe pihku, jookseb/jääb kellelegi pihku, puutub/satub kellelegi pihku, võtab pihku, võtab tuka/silmad pihku, pistab viis/viit kellelegi pihku);*
varn *(keegi paneb/riputab/viskab hambad varna jne);*
lõks *(keegi satub/kukub/langeb/jääb lõksu, kedagi veetakse/püütakse lõksu);*
korts *(miski läheb või tõmbub või tõmbab või kisub kortsu, keegi ajab midagi kortsu , keegi tõmbab kulmu kortsu)*

kimp *(keegi jääb või satub, hakkab või annab kimpu, keegi jätab või ajab kellegi kimpu);*
selts *(keegi heidab või lööb kellegagi seltsi, keegi võtab kellegi (oma) seltsi);*
lett *(keegi laob või paneb midagi letti, keegi võtab või tõmbab kellegi letti);*
latv *(miski hakkab/lööb/läheb või tõuseb latva);*
kihv *(midagi keeratakse, käänatakse, pannakse või pistetakse kihva) ja*
jooks *(saadakse, pannakse, pistetakse või minnakse jooksu)*

Nende noomenite aditiivivormid tunduvad olevat adverbistumas, kõige vähem käib see väide ehk noomeni *pihk* kohta. Ka on nende osalusel moodustatud püsiühendid pigem sama väljendi variantideks või kausatiivseteks edasiarendusteks (keegi *langeb lõksu* või keegi *püüab* kellegi *lõksu*).

Kas püsiühendites võivad aditiiv ja illatiiv olla paralleelvormideks? Andmebaasis on olemas nii väljendid *ninna/ninasse kargama* ja *haisu ninna/ninasse saama*, samuti paralleelvormid *tasku/taskusse* ühendites *tasku/taskusse pistma* või *toppima* ja *rooste/roostesse* ühendis *rooste/roostesse minema*, kuid need on harvad erandid. Üldreeglina ei ole andmebaasi aditiivi sisaldavates püsiühendites aditiivi asendamine illatiiviga võimalik. Kuid see ei paista olevat mitte püsiühendite, vaid pigem nendes osalevate noomenite omadus, enamik neist ei saa ka väljaspool püsiühendit esineda illatiivivormis. Väheste erandite hulgast võiks nimetada noomenit *mälu,* mis püsiühendi koosseisus esineb ainult aditiivis (*mällu suruma/sööbima/jääma*), kuid tekstides ka illatiivis:

(3)
Sõda on neid ajaloosündmusi, mis lõikab eredaid pilte inimese *mälusse.*

Püsiühendites esineb vähesel määral ka mitmuse aditiivi vorme, nt *jääb jalgu, kargab karvu kinni, kostab kõrvu, satub paelu,* keegi *pannakse raudu, võtab südame rindu, kargab silmi kinni.* Mitmuse aditiivi on ära märkinud ka Ülle Viks (Viks 1992: 27), kuid "Väikeses vormisõnastikus" pole mitmuse aditiivivorme esitatud. Samuti tunnistab mitmuse illatiivi uusim eesti keele grammatika (Erelt jt 2003), näiteks käändevormide moodustamise tabelites lk 38–39.

3.3. Arvukategooria verbikesksetes püsiühendites

Verbikesksete püsiühendite andmebaasis on ainsuslike ja mitmuslike käändsõnade suhe umbes 6:1. Võrdluseks: morfoloogiliselt ühestatud korpuses on ainsuse–mitmuse suhe umbes 4:1. Nagu osas 2.2 tõ-

deti, esineb verbikesksete püsiühendite nominaalsete komponentidena palju somaatilisi sõnu. Pole siis üllatav see, et kõige sagedamini esinevad mitmuses samuti somaatilised sõnad ja tähistavad nad neid kehaosi, mida inimesel on mitu (*silmad, jalad, käed, kõrvad, hambad, sõrmed, näpud, mokad, huuled, õlad, päkad, varbad*). Ka sõna *lõug* esineb püsiühendites samuti enamasti mitmuses, nt *laiutab lõugu, hoiab lõuad koomal, saab/annab vastu lõugu* jne.

Niemi jt (1998: 299–300) vaatlevad somaatilisi sõnu ja nende arvukategooriat soome keele idioomides, nimelt ühekaupa esinevaid kehaosi tähistavate sõnade idiomaatilist mitmust (nt *mennää nokilleen* 'ninuli kukkuma') ja paariskehaosi tähistavate sõnade idiomaatilist ainsust (*olla pelkkänä korvana* 'pingsalt kuulama'). Oma analüüsi tulemusena jõuavad nad järeldusele, et erinevalt inglise või rootsi keelest on soome keeles võimalik idiomaatilisuse tunnusena kasutada ühekaupa esinevaid kehaosi tähistavate sõnade mitmust. Nad väidavad selle fenomeni olevat väga soomepärase, sest ka lähedases sugulaskeeles, eesti keeles, ei kasutata sellist idiomaatilist mitmust, vähemalt mitte somaatiliste sõnade puhul. Ja kuigi üksikuid vastunäiteid leiab (seesama *kukub ninuli*; *valetab suud-silmad täis, liigutab ajusid*), ei esine eesti keele püsiühendites tõesti süstemaatilist mitmuse kasutamist ainsuse asemel.

Eesti keele püsiühendites kasutatakse mitmust peamiselt seoses paariskehaosadega (*käed, jalad, silmad, kõrvad*) või hulka moodustavate kehaosadega (*hambad, sõrmed, varbad*) (vt tabel 6). Paul Alvre on vaadelnud somaatiliste sõnade arvukategooriat ja selgitanud, et keeleajalooliselt vanem on paaris-kehaosade väljendamine ainsusliku sõnaga, sest selliseid mõisteid peeti ühtseks tervikuks, paariks. Ainsuse kasutamine mitmuse tähenduses on varem olnud üldisem nii paaris-kehaosade kui ka enama-arvuliste kehaosade puhul. Kuid need vormilt ainsuslikud ja sisult mitmuslikud keelendid taanduvad järk-järgult mitmuslike uusmoodustiste ees, nt *jalal seisma* > *jalul seisma, jalale saama* > *jalule saama* (Alvre 1962: 165).

Lisaks paari- või hulgakaupa esinevatele kehaosadele viitavatele sõnadele on tabelis 6 kaks plurale tantum-sõna (*püksid* ja *ohjad*), ka *sõnu* lausutakse tavaliselt korraga rohkem kui üks.

Tabel 6. Mitmuslikke noomeneid püsiühendites

Noomen	Esinemisi püsiühendites	Neist mitmuses	
		Arv	%
silm	204	117	57
jalg	101	67	66
käsi	227	64	28
kõrv	70	53	76
hammas	66	53	80
sõna	66	26	39
püks	25	22	88
lõug	28	22	79
sõrm	37	19	51
sarv	22	19	86
ohi (ohjad)	18	16	89
karv	22	15	68
mokk	27	12	44
kaigas	16	12	75
närv	17	11	65
koib	11	11	100
kaart	14	11	79
õlg	32	10	31
tiib	12	10	83
kont	16	10	63
huul	11	10	91

Kokkuvõte

Kui püsiühendeid moodustavad valdavalt sagedased verbid, siis verbikeskse püsiühendi prototüüpset noomenit saab kirjeldada eelkõige kui tavalist noomenit. Sagedus ja tavalisus on küll omavahel korrelatsioonis olevad, aga mitte päris kattuvad tunnused. Tavalisuse (üks) tunnus võiks olla põhitasandi mõiste väljendamine. Ka põhitasandi mõistete laiast hulgast kasutatakse püsiühendites just neid, mis väljendavad inimlähedasi kategooriaid, ka püsiühendites väljendub keelele üldiselt omane antropotsentriline maailmapilt. Nii saab püsiühendite vaatlemisel kinnitust vana tõde – keeles väljendatakse sageli keerukamat lihtsama ja abstraktsemat konkreetsema kaudu.

Kirjandus

Akimoto, Minoji 2004. A typological approach to idiomaticity. – The 20th Lacus Forum 1993, 459–467.

Alvre, Paul 1962. Kehaosi märkivate nimetuste nuumerusest. – Keel ja Kirjandus 2, 97–104; 3, 160–167.

Čermak, František 1995. Somatic Idioms Revisited. – EUROPHRAS 95 Europäische Phraseologie im Vergleich: Gemeinsames Erbe und kulturelle Vielfalt. (Studien zur Phraseologie und Parömiologie 15) 109–119. http://ucnk.ff.cuni.cz/doc/somatic-idioms.rtf

EKSS = Eesti Kirjakeele Seletussõnaraamat I–VI.

Fleischer, Wolfgang 1982. Phraseologie der deutschen Gegenwartssprache. VEB Bibliographisches Institut Leipzig.

Hasselblatt, Cornelius 2000. Eesti keele ainsuse sisseütlev on lühike. – Keel ja Kirjandus 11, 796–803.

Heine, Bernd, Tania Kuteva 2002. World Lexicon of Grammaticalization. Cambridge: Cambridge University Press.

Kaalep, Heiki-Jaan, Muischnek, Kadri 2003. Püsiühendite leidmine suurtest tekstikorpustest. – Toimiv keel I. Töid rakenduslingvistika alalt. Tallinn: Eesti Keele Sihtasutus, 101–118.

Kaalep, Heiki-Jaan, Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastus.

Kaalep, Heiki-Jaan, Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Tartu, 87–99.

Kio, Kati 2002. Sisseütleva ja suunduva käände probleemistikust. Bakalaureusetöö. Käsikiri TÜ eesti keele õppetoolis.

Krikmann, Arvo 2004. "Sai hea obaduse vastu obadust". Löömist ja peksmist märkivad väljendid eesti keeles. Reetor 3. Tartu: Eesti Kirjandusmuuseum.

Moon, Rosamund 1998. Fixed Expressions and Idioms in English. A Corpus-based Approach. Oxford: Clarendon Press.

Muischnek, Kadri 2004a. Noomeni ja verbi püsiühendid eesti keeles. Lauseliikmeist eesti keeles. Tartu Ülikooli eesti keele õppetooli preprindid 1. Tartu, 50–56.

Muischnek, Kadri 2004b. Verbi ja noomeni püsiühenditest eesti keeles. – Keel ja Kirjandus 8, 574–589.

Mägiste, Julius 1983. Estnisches etymologisches wörterbuch. Helsinki: Finnisch-Ugrische Gesellschaft.

Nenonen, Marja 2002. Idiomit ja leksikko. Lausekeidiomien syntaktisia, semanttisia ja morfologisia piirteitä suomen kielessa. Joensuun Yliopiston humanistisia julkaisuja 29.

Nenonen, Marja, Niemi, Jussi 1999. Morphological Isolates in Idioms: Cranberries or Real Words? – Brain and Language 68, 158–164.

Niemi, Jussi, Nenonen, Marja ja Penttilä, Esä 1998. Number as a Marker of Idiomaticity. – Proceedings of the XVIth Scandinavian Conference of Linguistics. Turku/Abo, November 14–16, 1996. Abo Akademis Tryckeri, Turku, 291–304.

Ojutkangas, Krista 2000. Ruumiinosannimien kieliopillistumisesta suomessa ja virossa. Virittäjä 1, 2–21.

Pajusalu, Renate, Tragel, Ilona, Veismann, Ann, Vija, Maigi 2004. Tuumsõnade semantikat ja pragmaatikat. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 5.

Rätsep, Huno 1978. Eesti keele lihtlause tüübid. Tallinn: Valgus.

Tragel, Ilona 2002. On Estonian Core Verbs. – Papers in Estonian Cognitive Linguistics. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 2. Toim. I. Tragel. Tartu, 145–169.

Viks, Ülle 1992. Väike vormisõnastik I–II Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Wiedemann = Ferdinand Johann 1973. Eesti–Saksa sõnaraamat. Neljas, muutmata trükk teisest, Jakob Hurda redigeeritud väljaandest. Tallinn: Valgus.

Õim, Asta 1993. Fraseoloogiasõnaraamat. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Eesti suulise keele korpuse automaatne pindsüntaktiline analüüs[1]

Kaili Müürisep, Helen Nigol, Heli Uibo
Tartu Ülikool

1. Sissejuhatus

Suulise keele korpuse käsitsi süntaktiline märgendamine on keerukas ja aeganõudev protsess. Et seda lihtsustada, võtsime kasutusele eesti keele kitsenduste grammatika (ESTKG) analüsaatori (Müürisep 2000), mis oli algselt loodud kirjaliku keele tekstide analüüsimiseks.

Eesti suulise keele korpuse loomine algas 1997. a (Hennoste jt 2000). Hetkel (november 2005) on korpus 911 000-sõnaline. Korpuses on nii argi- kui ka avaliku suhtluse tekste, nii spontaanset kui ka ettevalmistatud kõnet, nii dialooge kui ka monolooge. Korpuse litereerimisel kasutatakse konversatsioonianalüüsi transkriptsiooni. Osa tekste on morfoloogiliselt analüüsitud, osa on ka morfoloogiliselt ühestatud ja osas on märgendatud dialoogiaktid (Hennoste jt 2003).

Oma eksperimentides kasutasime morfoloogiliselt ühestatud tekste. Süntaksianalüsaatori kohandamiseks suulisele kõnele tuli muuta osalausepiiride reegleid, parandada mitmeid süntaktilisi kitsendusi ning võtta kasutusele paar uut märgendit.

2. Eesti keele kitsenduste grammatika süntaksianalüsaator

Eesti keele kitsenduste grammatika analüsaator töötati välja aastatel 1996–2001 Tartu ülikoolis (Roosmaa jt 2003). Süntaktilise analüüsi protsess on selles jaotatud kaheks osaks. Morfoloogiline ühestaja tegeleb kontekstiinfo põhjal morfoloogiliselt mitmese analüüsiga sõnavormile õige morfoloogilise kirjelduse väljavalimisega, süntaksianalüsaator leiab sõnavormi süntaktilise funktsiooni lauses. Meie analüsaator põhineb kitsenduste grammatikal (Karlsson jt 1995), mis on loomult reduktsionistlik, s.o analüüsi alguses lisatakse igale sõnavormile kõik võimalikud analüüsivariandid ja seejärel hakatakse konteksti mittesobivaid eemaldama. Eemaldamine toimub vastavalt

[1] Tööd on osaliselt toetanud Eesti Teadusfond (grant nr 5685) ning HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

kitsenduste grammatika reeglitele ehk kitsendustele, millest igaüks esitab mõnda spetsiifilist keelereeglilaadset fakti. Üldisem grammatikareegel kujuneb alles nende koosmõjust. ESTKG-s on hetkel 1118 süntaktiliste märgendite eemaldamise reeglit.

Ideaaljuhul jääb analüüsi lõppedes igale sõnavormile üks süntaktiline märgend. Kui sõnal võib olla lauses mitu funktsiooni, antakse need kõik. Mitme märgendiga jäävad ka sõnad, mida analüsaator pole suutnud lõpuni ühestada. Grammatikareeglid on kirjutatud nii, et pigem jäetakse sõna mitme analüüsiga kui eemaldatakse korrektne märgend.

ESTKG-s märgendatavad süntaktilised funktsioonid vastavad enam-vähem standardses eesti keele grammatikas (Erelt jt 1993) eristatavatele süntaktilistele funktsioonidele.

Öeldise märgendid eristavad finiitset ja infiniitset öeldist ning eraldi märgendid on põhiverbile (@+FMV, @-FMV) ja abi- ning modaalverbidele (@+FCV, @-FCV). Fraasi põhjadest märgendatakse alust, sihitist, öeldistäidet, määrust (vastavalt @SUBJ, @OBJ, @PRD, @ADVL). Laiendite märgendid näitavad põhja leidumise suunda, kuid ei viita ühelegi sõnale konkreetselt. See tähendab, et on eraldi märgendid ees- ja järeltäienditele (@NN>, @<NN jt), eessõna ja tagasõna laienditele (@<P, @P>) ning kvantori ees- ja järellaienditele (@Q>, @<Q). Täienditest eristatakse omadus-, määr-, kaas- ja nimisõnalisi täiendeid ning partitsiipe ja infinitiivseid verbivorme täiendina.

Näide automaatselt analüüsitud tekstist on toodud joonisel 1. Iga sõnavormi all on antud selle tüvi ja lõpp, morfoloogiline kirjeldus kaldkriipsude vahel ning süntaktiline märgend (algab sümboliga @).

Kahjuks ei ole võimalik kõiki sõnu automaatselt ühestada, ligikaudu iga kümnes sõna jääb mitme märgendiga. Ilukirjandusliku teksti analüüsi tulemused on toodud tabelis 1, kus teises veerus on toodud tulemused, kui tekst oli eelnevalt käsitsi morfoloogiliselt ühestatud, ning kolmandas veerus täisautomaatse analüüsi tulemused. Saagis näitab, mitu protsenti sõnadest on õige märgendiga, pööramata tähelepanu sellele, kas sõna on ühene või mitte. Täpsus näitab, mitu protsenti kõigist märgenditest on oma õigel kohal ehk siis leitud korrektsete märgendite arvu suhet kõigi leitud märgendite arvu. Ühesus näitab, mitu protsenti sõnadest on ühese analüüsiga.

Joonis 1. Näide automaatselt süntaktiliselt analüüsitud kirjaliku keele tekstist

```
Oli
   ole+i //_V_ main indic impf ps3 sg ps af #cap #Intr // **CLB @+FMV
päikesepaisteline
   päikese_paiste=line+0 //_A_ pos sg nom #line // @AN>
hommikupoolik
   hommiku_poolik+0 //_S_ com sg nom // @SUBJ
$,
   , //_Z_ Com //
mustad
   must+d //_A_ pos pl nom // **CLB @AN>
laigud
   laik+d //_S_ com pl nom // @SUBJ
aurasid
   aura+sid //_V_ main indic impf ps3 pl ps af #FinV // @+FMV
keset
   keset+0 //_K_ pre #part // @ADVL
määrdunud
   määrdu=nud+0 //_A_ pos #nud partic // @VN>
lund
   lumi+0 //_S_ com sg part // @<P
$.
   . //_Z_ Fst //
```

Tabel 1. Analüüsi tulemused (%)

	Käsitsi ühestatud	Automaatselt ühestatud
Saagis	98,53	96,41
Täpsus	87,57	78,09
Ühesus	89,54	82,70

3. Süntaksianalüsaatori kohandamine suulisele keelele

Et kohandada kirjaliku keele analüsaatorit suulise keele analüüsiks, tuli lisada uusi märgendeid ja reegleid osalausepiiride tuvastamiseks. Samuti tuli muuta mitmeid süntaktilisi kitsendusi.

3.1. Uued märgendid

Uute märgenditena võeti kasutusele @B partikli tähistamiseks ja @T tundmatu sõna märkimiseks.

Suulises kõnes esineb palju partikleid ja neid vaadatakse kui eraldi sõnaliiki. Partiklid on muutumatud omaette tüvega sõnad, mis võivad esineda ka kirjalikus keeles (*siis*, *jah*) või olla kirjakeele sõ-

nade häälduslikud variandid (*sis*, *kule*), osa partikleid on aga häälikuühendid, mis paiknevad foneetiliselt ja fonotaktiliselt häälitsuse piirimail (*öök*, *phtüi*, *mhmh*). Süntaktiliselt võivad need üksused moodustada vestluses terve kõnevooru, seega ka süntaktilise üksuse (*mhmh*). Kui partiklid kuuluvad mingisse intonatsiooniliselt tervikliku pikemasse üksusesse, nt lausesse, siis ei kuulu nad lause grammatilisse struktuuri. Nad ei seostu mingi kindla sõnaklassiga lauses, kuid võivad töötada kogu lause juurde kuuluvate lauselaienditena (Hennoste 2002).

Kõnes esineb sageli grammatiliselt ebakorrektseid või poolikuid lauseid, samuti leidub sõnu, mille lausumine on poole pealt katkestatud ning seetõttu on nad automaatsel morfoloogilisel analüüsil märgendatud kui tundmatud sõnad. Sellistes lausetes on mõnede sõnade süntaktilise funktsiooni määramine ka lingvisti poolt võimatu ning sellised sõnad on treening- ja testkorpuses märgendatud kui tundmatu süntaktilise funktsiooniga sõnad: @T. Näites 1 ei ole võimalik kahte viimast sõna süntaktiliselt analüüsida, sest lause on lõpetamata.

(1)
A: kui (@J) sa (@SUBJ) võtame (@+FMV) mingisuguse (@NN>) asja (@OBJ) kuigi (@J) seda (@T) see (@T)

3.2. Süntaktilise analüüsi aken

Kirjalikus tekstis on üheks analüüsiühikuks ehk süntaktiliseks aknaks, mille ulatuses vaadatakse sõna süntaktilise funktsiooni määramisel konteksti, lause. Kirjaliku keele lause on autori tahtel märgistatud punktuatsioonimärkidega. Suulisele kõnele on iseloomulik, et seda on väga raske lauseteks jagada. Suulises kõnes on lauselõputunnusteks pausid ja intonatsiooni langused või tõusud. Litereerimisel on need ka erisümbolitega märgistatud. Kõnevool jagatakse intonatsioonilisteks üksusteks, mitte grammatilisteks üksusteks. Selliseid põhiüksusi on kaks: a) üksus, mida võib nimetada lausungiks, selle lõpus on selgelt langev intonatsioon, mis osutab lõpetatusele ning mida märgitakse punktiga; b) lausungid jagunevad intonatsiooniliselt osadeks, mille lõpus intonatsioon langeb, kuid vähem kui lausungi lõpus (poollangev intonatsioon). Selline intonatsioon osutab, et tegu on piiriga, kuid üksus ei lõpe. Seda märgitakse komaga. Lisaks kasutatakse tõusva intonatsiooniga lõppeva üksuse lõpus küsimärki.

Transkriptsioonimärgenduse täpsemal uurimisel ilmnes siiski, et lausungiteks jagamine polnud piisavalt täpne. Sageli tekkis intonatsioonilangus ka keset süntaktilist lauset. Näiteks: ma tõstan kartulid ära. ja sousti ka. Seepärast otsustasime käsitleda süntaktilise aknana ühte kõnevooru ning vaadelda punktidega eraldatud üksusi kui koordineeritud lauseosi.

Ilmselt sõltub lause piiri määramine siiski teksti liigist, monoloogide puhul on punktide kasutamine lauselõputunnusena igati õigustatud.

Samas võib tekstist leida ka näiteid, kus kõnevoor jagab lause pooleks (vahele rääkimine). Näites 2 tuleb sõnavorm *auto* esimesel korral märgendada kui tundmatu süntaktilise funktsiooniga sõna, sest vooru sees ei ole võimalik selle funktsiooni määrata.

(2)
A: see oli kõik ee ausatel eesmärkidel et perele auto
B: ei no loomulikult
A: saada. aga lissalt mai saand seda autot

3.3. Osalausepiiride määramise reeglid

Kirjaliku keele osalausepiirid määratakse sidesõnade, kirjavahemärkide ja verbide põhjal. Osalause esimesele sõnale lisatakse märgend CLB. Osalausepiiride määramise põhireegel on järgmine: kui sõnale eelneb kirjavahemärk ja/või sõna ise on sidesõna ning vasakul ja paremal pool seda sõna leidub verbi pöördeline vorm, siis see sõna on osalause esimene sõna. See reegel võib mõnede tingimuste osas varieeruda.

Koma või rinnastavate sidesõnade *ja*, *ning*, *või*, *ega*, *ehk* abil võib eraldada mitte ainult osalauseid, vaid ka koondlause korduvaid liikmeid. Seda, millise eraldajaga just konkreetsel juhul on tegu, on ilma süntaktilist informatsiooni teadmata raske otsustada, eriti veel juhul, kui antud sõna lähemas kontekstis ei leidu verbe. Seepärast lisatakse nendele sõnadele üksnes oletatava osalause tunnus CLB-C.

Kirjaliku eesti keele kitsenduste grammatikas on 47 osalausepiiride määramise reeglit, paljud neist on väga spetsiifiliste juhtude jaoks.

Kõik kirjaliku keele osalausepiiride määramise reeglid tuli ümber vaadata, sest kirjavahemärkide tähendus on suulise keele tekstides erinev. Uutes reeglites kasutati intonatsioonimärke ja partikleid

(*noh*, pausi täitjad *aa* ja *ee* jt). Punkti loetakse osalausepiiri kindlaks lõputunnuseks. Partiklit loetakse eraldajaks, kui kummalgi pool kontekstis leidub finiitseid verbivorme.

Reeglid püüavad leida valestarte ja märgendada neid osalausepiiri märgenditega, kuid see õnnestub ainult juhul, kui valestart sisaldab verbi (näide 3).

(3)
mul (CLB) on kassetil (CLB-C) oleks ruumipuudus tekkinud

Väga raske on tuvastada sisemiste osalausete lõppu ning need on peamised vigade tekkimise kohad. Näites 4 on tuvastamata osalauselõpp tähistatud tärniga.

(4)
kuna (CLB) ta tundus mulle esmakuulamisel või noh algul kui (CLB) ma kuulasin (*) kuidagi liiga afišlik või noh noh äraleierdatud.

Kirjaliku keele tekstis oleks see osalausepiir tähistatud komaga.

Suulise keele analüüsi grammatika sisaldab 21 osalausepiiride määramise reeglit, mida on oluliselt vähem kui kirjalikus keeles. Seda võib seletada sellega, et kirjaliku keele grammatika oli kohandatud analüüsima ka juriidilisi tekste, mille osalausepiiride määramisel peab arvestama paljude spetsiifiliste juhtudega (loetelud, paragrahvid, sulgudes tekst).

3.4. Parandused süntaktilistes kitsendustes

Nagu eelpool mainitud, koosneb algne kirjaliku keele grammatika 1118 reeglist. Neid reegleid rakendati 2200-sõnalisele eelnevalt süntaktiliselt märgendatud treeningkorpusele (suuline argivestlus) ning tekkinud vigade analüüsimisel parandati või muudeti reegleid. Enamasti tuli muuta reeglite kontekstipiiranguid. ESTKG reegleid võib rakendada kolmes režiimis: a) kontekstitingimused kehtivad kogu lause ulatuses, b) kontekstitingimused kehtivad ainult kindlate osalausepiiride vahel (CLB), c) kontekstitingimused kehtivad kõigi osalausepiiride sees (CLB ja CLB-C). Enamasti oli vigade põhjuseks asjaolu, et reeglid arvestasid liiga kauget konteksti, mis asus väljaspool tegelikku osalauset. Selle parandamiseks tuli muuta vaid konteksti kontrollimise režiimi tunnust. Reegel joonisel 2 eemaldab aluse märgendi, kui osalauses on ainsuse 1. pöördes verb ja sõna ise on osastavas. Esimene on algne reegel, teine modifitseeritud.

```
(5)
(@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB)
 (@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB-C)
```

Ka tuli üle vaadata kõik kirjavahemärke kasutavad reeglid ja võimalusel täiendada neid tingimustega, kui kirjavahemärke lauses ei leidu.

Samuti on sõnavara kasutus suulises keeles erinev. Näiteks tuli arvestada, et relatiivpronoomenit *mis* võidakse kasutada küsilauses küsisõna *kas* asemel või võrdlustes *nagu* ja *kui* asemel.

Reeglite muutmine lõpetati, kui vigade protsent vähenes 7,5-lt 3-le.

4. Hindamine

4.1. Testkorpuse kirjeldus

Analüsaatori töö hindamiseks loodi käsitsi süntaktiliselt märgendatud testkorpus[2], mis koosnes 2543 sõnast. Testkorpus koosnes argivestlustest, milles oli nii pikemaid jutustavaid dialooge kui ka lühikeste remarkidena dialooge. Korpus loodi sel viisil, et parandati analüsaatori poolt põhjustatud vead käsitsi ära. Parandajaks oli üks inimene. Sellisel hindamisel on nõrgad kohad, mis on ka autoritele teada:

1. Korpus on liiga väike ega hõlma suulise kõne kõiki tahke.
2. Automaatselt analüüsitud korpust parandades võivad jääda mitmed vead märkamata, sest esialgsel vaatamisel tundub märgend olevat korrektne ning inimene ei süüvi peensustesse.
3. Korpust peaks vähemalt esialgsel etapil märgendama mitu inimest, mis tagaks kõigi vigade avastamise ja järjekindlama märgenduse.

4.2. Tulemused

Süntaksianalüsaatori väljundit võrreldi käsitsi märgendatud korpusega ning saadi järgmised tulemused (kirjaliku keele andmed on toodud sulgudes):

- sõnade arv korpuses – 2543
- saagis – 97,3% (98,5%)
- täpsus – 89,2% (87,5%)
- ühesus – 91,5% (89,5%)

[2] http://www.ut.ee/~kaili/Korpus/Spoken

4.3. Vigade tüübid

Vead võib jagada järgmistesse klassidesse (arvulised näitajad on toodud treeningkorpuse vigade põhjal):

1. Osalausepiiride valesti määramisest tingitud vead – 16. Näiteks:

(6)
selle taga on saad aru selline lähenemine

Sõnavormilt lähenemine eemaldati aluse märgend, sest samas osalauses on ainsuse 2. pöördes verb.

2. Tundmatu süntaktiline funktsioon – 12. Suulises kõnes on palju poolikuid, väljajättelisi või vigaseid lauseid, mistõttu ei ole isegi inimesel võimalik kõigi sõnade süntaktilisi funktsioone määrata. Tundmatud funktsioonid märgendatakse märgendiga @T. Samas ei ole võimalik koostada reegleid, kus üks või teine sõnavorm on ilmutatult määratlemata süntaktilise funktsiooniga. Analüsaator lisab sõnavormile morfoloogilise info ja konteksti põhjal kõik võimalikud märgendid ning hakkab siis süntaktilisi kitsendusi rakendades neid ükshaaval eemaldama. Kui lause on poolik, siis pole piisavalt kontekstiinfot ning sõnale võib jääda mitu märgendit. Võib olla ka juhtumeid, et reeglite põhjal üritatakse eemaldada kõiki märgendeid, kuid analüsaator ei luba kunagi eemaldada viimast. Nii võivad mittegrammatilises lauses olla sõnadel ka üsna juhuslikud süntaktilised funktsioonid. Näide 7 on pealeräkimise analüüsist. Nurksulgudes on inimese määratud märgendid, sümboliga@ aga need, mis tegelikult leiti. Sõnavorm nad jäi mitmeseks aluse ja sihitise vahel, käsitsi oli ta analüüsitud kui tundmatu süntaktilise funktsiooniga, sest kontekst, mille põhjal selgub tegelik funktsioon, paikneb alles järgmises kõnevoorus.

(7)
A: et [J] @J nad [T] @SUBJ @OBJ
B: mhmh [B] @B
A: sobivad [FMV] @FMV kätte [ADVL] @ADVL

Näites 8 on tundmatu funktsiooniga sõnal juhuslik märgend:

(8)
Pöidla [T] @NN> ja [J] @J kõik [OBJ] @OBJ on [+FCV] @+FCV tehtud [-FMV] @+FMV

3. Omadussõna nimisõna rollis – 9. Omadussõnad võivad elliptilistes lausetes esineda aluse või sihitise rollis (näide 9). Seda tüüpi vigu esineb ka kirjaliku keele analüüsil, kuid harvem.

(9)
Ma [SUBJ] @SUBJ ostaks [+FMV] @+FMV selle [NN>] @OBJ maasikamaitselise [OBJ] @ADVL

4. Varasem vale analüüs – 5. Mõni varem rakendatud reegel on analüüsinud mõnda sõnavormi valesti ning see vigane analüüs põhjustab ka naabersõnade vigast analüüsi.

5. Kordused – 3 (aluse kordamisel rikutakse unikaalsuse printsiipi).

(10)
noh [B] @B se [SUBJ] @NN> see [SUBJ] @SUBJ on [FMV] @FMV tähtis [PRD] @PRD

6. Muu – 14 (sellised vead võivad esineda ka kirjalikus tekstis).

4.4. Mitmesused

Kui võrrelda kirjaliku keele ja suulise kõne allesjäänud mitmesuste klasse, siis ilmneb, et nende arvuline järjestus on erinev. Kirjaliku keele automaatselt analüüsitud tekstides oli domineeriv mitmesus määruse ja järeltäiendi vahel, kolm korda harvem esines sihitise ja eestäiendi, määruse ja eestäiendi ning aluse ja sihitise vahelist mitmesust. Suulise keele tekstides põhjustavad arvukaimat mitmesust infiniitsed verbivormid, millele on jäänud nii öeldise kui ka määruse märgendid. Järgnevad määrus ja alus, määrus ja järeltäiend, alus ja sihitis, alus ja öeldistäide.

5. Mitteladususes

Suulise kõne süntaktilisel analüüsil tuleb hakkama saada mitmete suulisele kõnele omaste nähtustega, mida üldiselt nimetatakse mitteladususteks (ingl *disfluency*). Mitteladususteks peetakse näiteks kordusi, parandusi, poolikuks jäänud lausungeid. Eelpool nägime, et mitte kõik suulise kõne süntaktilisel analüüsil ilmnevad probleemid pole reeglitega lahendatavad. Üheks võimaluseks mitteladususteg a hakkama saada on eeltöötlemisetapil need märgendada, mille käigus ebagrammatilised lausungid tehakse süntaksianalüsaatori jaoks

grammatilisteks. Sageli nimetatakse seda protsessi ka normaliseerimiseks (ingl *normalization*). Mitteladususi on iseseisva analüüsietapina märgendatud näiteks suulise keele korpuses Switchboard (Meteer jt 1995) ja ICE-GB (Meyer 2002: 96).

Switchboardi korpuses on mitteladususte märgendamiseks välja töötatud spetsiaalne märgendamisskeem, mis on võetud ka eesti suulise kõne märgendamise aluseks. Kasutusel olevad märgendid on toodud tabelis 2.

Tabel 2. Mitteladususte analüüsil kasutatavad märgendid, nende seletus ja näited

Märgend	Seletus	Näide
{D ...}	partikkel	**{D** nagu**}**; **{D** noh**}**
{F ...}	täidetud paus	**{F** ee**}**; **{F** õ**}**
{B ...}	hingamine	**{B** hh**}**
{A ...}	raskesti analüüsitav	meil kül `präegu sin `kohapeal **{A** meil**}** `sellist vari`anti ei `paista. /
[RE ... + ...]	kordus	**[RE** nii + nii**]**
[RP ... + ...]	parandus	**[RP** kuidas + kas**]** **[RP** selli- + sellist**]**
/	lausung	H: tere, ma `sooviksin saada `infot õppe`laenu kohta. /
-/	lõpetamata lausung	V:siis saate sealt -/ minu=arust ´dekanaat väljastab niisugused ´tõendid /
--	lausung jätkub	H: ma tahaksin tellida teie kataloogist seda ´juuksehooldusvahendit, -- V: [jaa] / H: -- [Blisaana.] /

Vastavate märgenditega on analüüsitud 16 dialoogi (5631 sõna), kuid eelmärgendatud teksti pole veel jõutud automaatselt süntaktiliselt analüüsida, et hinnata selle skeemi rakendatavust tegelikkuses.

Märgendamisel tekkinud raskusi:

1. Lausungid ei lõpe alati intonatsiooni langusega, nt

> H: ma:=ei=olnd `Tartus / mul jäi mine`mata=ja=nüüd ma [tahaks] uuesti `aega võtta /

2. Paranduse algus ja lõpp pole alati üheselt määratav. Sageli on raske vahet teha, kas eelnev üksus jäeti pooleli ja alustati uut või on tegu parandamisega, nt

H: [RE et + et] -/ [RP kas te k- + mille põhjal te] {D nagu} ´otsustate seda. /

Kuid on ka väga selgeid näiteid sellest, kus üks lausung jääb pooleli ja alustatakse uuega, nt

ma=i=´oska {D nagu} {D nimodi} kohe täpselt ´öelda et {D noh} selline: -/ {B .hh} {Fm} ´tegemist on ühe üliõpilasorganisatsiooni ´aastapäevaga. /

3. Pealeräägitud kõne tulemuseks on palju poolelijäänud lausungeid, mis iseseisvana ei kanna endas just palju informatsiooni, seepärast on analüüsis lubatud minna üle kõnevooru piiri, et tekiks lauseline tervik, nt

V: `nii on väga `raske $ teile anda `infot, $ {D noh} =et see on [väga] --
H: {D [mhmh]} /
V: -- erineva `hinnaklassiga, erineva `pikkusega ja igal `maal on {D noh} omad ekskursi`oonid, oma `vaata[mis]väärsused. /

Vaatamata sellele, et ka käsitsi märgendamisel tekib küsitavusi, on selline märgendamine siiski suureks abiks. Kui vaadata eespool välja toodud automaatsel analüüsil tekkinud veatüüpe, siis mitteladususte märgendamine aitab osalt kindlasti lahendada lausungipiiride leidmise, poolelijäänud lausungite, paranduste ja kordustega seotud probleeme.

6. Järeldused ja tulevikuplaanid

Kirjaliku keele jaoks loodud süntaksianalüsaatori kohandamine suulisele kõnele osutus kergemaks ülesandeks, kui algul loodetud. Kõige olulisem oli lahendada osalausepiiride määramise probleem, süntaktilised kitsendused vajasid vaid vähest muutmist. Üllatuslikult olid suulise keele süntaksianalüsaatori tulemused paremadki kui kirjaliku keele korral. Sellele on kaks põhjendust: 1) suulises kõnes kasutatakse palju partikleid ja adverbe, mille süntaktiline analüüs on triviaalne, 2) lausungid on lühemad.

Meie hinnangul on kitsenduste grammatika sobiv formalism suulise kõne süntaktiliseks analüüsiks: 1) analüüs on pindmine ega näita ilmutatult isegi põhja ja laiendi vahelist seost; see võimaldab ka valesti ühilduvad sõnavormid mõningatel juhtudel õigesti analüüsida; 2) sõnad, mida ei õnnestunud analüüsida, jäetakse mitmese analüüsiga, mis samuti vähendab vigade hulka.

Väga raske on hinnata, kui lihtne oleks kohandada ESTKG morfoloogilist ühestajat (Puolakainen 2001) suulisele kõnele. E. Bick

(1998) jõudis oma katsetes järeldusele, et portugali keele kitsenduste grammatika morfoloogilise ühestamise reeglite muutmine andis parema tulemuse kui süntaktiliste kitsenduste modifitseerimine, kuna morfoloogiline ühestamine vajab väiksemat konteksti. Sama ei pruugi kehtida eesti keele korral, sest keeled on selleks liiga erinevad. Loodetavasti annavad edaspidised eksperimendid sellele küsimusele peagi vastuse.

Töö käigus selgus, et kitsenduste grammatika ei sobi mõnede suulisele kõnele iseloomulike nähtuste märgendamiseks, nagu kordused, valestardid ja eneseparandused. Sedalaadi märgendus peaks olema tehtud juba enne süntaktilist analüüsi, kasutades mõnda teist formalismi. Sellise märgenduse olemasolu hõlbustaks süntaksianalüsaatori tööd.

Me ei ole veel välja töötanud meetodit suulise kõne lausungi esitamiseks süntaksipuuna. Võib-olla oleks mõtet kasutada sedasama lähenemist nagu poolautomaatsel kirjakeele puudepanga Arborest[3] loomisel (Bick jt. 2004).

Kirjandus

Bick, Eckhard 1998. Tagging Speech Data – Constraint Grammar Analysis of Spoken Portuguese. – Proceedings of the 17th Scandinavian Conference of Linguistics. Odense.

Bick, Eckhard, Uibo, Heli, Müürisep, Kaili 2004. Arborest – a VISL-Style Treebank Derived from Estonian Constraint Grammar Corpus. – Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). Tübingen, Germany.

Erelt, Mati, Kasik, Reet, Metslang, Helle, Rajandi, Henno, Ross, Kristiina, Saari, Henn, Tael, Kaja, Vare, Silvi 1993. Eesti keele grammatika. II Süntaks. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu, 56–73.

Hennoste, Tiit, Koit, Mare, Rääbis, Andriela, Strandson, Krista, Valdisoo, Maret, Vutt, Evely 2003. Developing a Typology of Dialogue Acts: Tagging Estonian Dialogue Corpus. – DiaBruck 2003. Proceedings of the 7th Workshop on the Semantics and Pragmatics of

[3] http://corp.hum.sdu.dk/arborest.html

Dialogue. Eds I. Kruijff-Korbayová, C. Kosny. Saarland University, Saarbrücken, 181–182.

Hennoste, Tiit; Lindström, Liina, Rääbis, Andriela, Toomet, Piret, Vellerind, Riina 2000. Tartu University Corpus of Spoken Estonian. – Congressus Nonus Internationalis Fenno-Ugristarum 7.–13. 8. 2000. Pars IV. Dissertationes sectionum: Linguistica I. Eds T. Seilenthal, A. Nurk, T. Palo. Tartu, 345–351.

Karlsson, Fred, Anttila, Arto, Heikkilä, Juha, Voutilainen, Atro 1995. Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter.

Meteer, Marie, Taylor, Ann, MacIntyre, Robert, Iyer, Rukmini 1995. Disfluency Annotation Stylebook for the Switchboard Corpus. Linguistic Data Consortium. www.ldc.upenn.edu/Catalog/CatalogList/LDC99T42/DFLGUIDE.PS

Meyer, Charles F. 2002. English Corpus Linguistics: An Introduction. Cambridge University Press.

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu.

Roosmaa, Tiit, Koit, Mare, Muischnek, Kadri, Müürisep, Kaili, Puolakainen, Tiina, Uibo, Heli 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? – Keel ja Kirjandus 3, 192–209.

Millist leksikoni vajab arvuti tähenduse mõistmiseks?[1]

Heili Orav, Kadri Vider
Tartu Ülikool

1. Sissejuhatus

Ükski ulatuslikum arvutuslingvistiline rakendussüsteem (infootsingu-, keeleõppe-, tõlkesüsteem jne.) ei toimi piisava leksikonita, kuid see leksikon ei saa olla ainult sõnade vorme fikseeriv (analüüsiv, sünteesiv) süsteem, vaid peab sisaldama ka piisavat semantilist ja pragmaatilist informatsiooni. Viimasest johtuvalt on probleemid leksikonide koostamisest tunnistatud tänapäeva arvutuslingvistiliste rakendussüsteemide "pudelikaelaks".

Arvutileksikonid on arenenud teoreetilise lingvistika leksikoni- ja üldisemalt semantikakontseptsioonide mõjul, kuid vaatamata sõnastike kesksele kohale teoreetilises keeleteaduses ei ole leksikaalsel semantikal tänaseni üldist teooriat ega ole ühist semantilist esitust kõigi sõnaklasside ning eri liiki sõnastike jaoks. Seevastu on palju teooriaid, mis keskenduvad mõnele leksikaalse semantika aspektile, analüüsides kas teatud tüüpi sõnu või sõnavara üksuse tähendust mõnest konkreetsest aspektist, kuid mis ei suuda katta kõikide sõnatüüpide ega kogu sõnavara semantilise struktuuri analüüsi (Orav 1998).

Tulevikku vaadates on kogu maailmas sõnastike tegemisel kaks põhisuunda (Muischnek jt 2003):

• Leida uusi lähenemisi, kuidas teha nii korpustest kui teistest allikatest arvutis loetavaid sõnastikke, mida saaks kasutada nii keeletehnoloogilistes rakendustes kui ka tavaliste, trükitud (või CD-ROM-idel) sõnastike väljaandmiseks;

---

[1] Artikkel ja selle aluseks olnud uurimistöö on edenenud ETFi grandi nr 5534 "Tähenduspõhise keeletöötluse ressursid ja töövahendid eesti keele jaoks" (2003–2006) ja sihtfinantseeritava teadusteema nr 0182541s03 "Eesti keele arvutimudelid ja keeleressursid: teoreetilised ja rakenduslikud aspektid" (2003–2007) toel.

• Leida organisatoorsetele küsimustele uusi lahendusi: erinevate leksikaalsete andmete standardiseerimine, uurimine, arendamine, lahendused sõnastike avalikuks kasutamiseks.

Kuna eesti keele (arvuti)sõnastike koostajatel on samad eesmärgid, siis tahame siinses artiklis rääkida arvutileksikonidest, ka eesti keele baasil tehtuist, nende hetkeseisust, uutest lähenemistest ja perspektiividest. Keskendume muuhulgas ka arvutisõnastike keeleteaduslikele käsitlustele, sest igasugune praktiliselt töötav süsteem tugineb eelnevatele teoreetilistele uurimustele. Artikli algusosas anname ülevaate arvutisõnastike arengust, mis traditsioonilistest sõnaraamatutest on jõudnud leksikaalsete andme- ja teadmusbaasideni. Edasi tutvustame leksikaalsete andmebaaside ja teadmusbaaside teaduslikku tausta ning viimases osas toome mõned näited eestikeelsetest freimidest.

2. Arvutileksikonid

Arvutileksikograafia areng on kulgenud arvutisse sisestatud sõnastikutekstidelt leksikaalsete andmebaasideni. Tulevikku jääb leksikaalsete teadmusbaaside koostamine.

Eestis on paljud pabersõnastikud antud välja ka elektrooniliselt. Elektroonilised sõnastikud erinevad traditsioonilistest, inimese jaoks mõeldud (paber)sõnastikest nii oma struktuuri kui sisu poolest. Hulk elektroonilisi sõnastikke on kasutatavad ka interneti kaudu[2].

Arvutisõnastik ei ole siiski vaid arvutisse viidud sõnaraamatu tekst. Sõnastikuartikli erinevad funktsionaalsed osad (märksõna ise, grammatiline info, seletus, näited) peavad olema formaalselt identifitseeritavad, nt varustatud spetsiifiliste märgenditega. Just tänu sellisele liigendusele on sõnastikus esitatud materjal ka "arvuti poolt loetav" (*machine readable dictionary*, MRD) ning mitte ainult inimese poolt kasutatav raamatu asendaja. Arvuti abil võidakse otsida või analüüsida eraldi sõnastikuartikli erinevaid osi, nt seletusi, näiteid, grammatilist infot. Pabersõnastike arvutiversioone kasutatakse leksikaalsete andmebaaside ja teadmusbaaside tegemiseks, aga ka igasuguse muu lingvistilise info otsinguks.

---

[2] Ülevaate neist leiad nt Langemets 2000, Muischnek jt 2003, aga ka http://keeleveeb.edu.ee/ ning http://www.keelevara.ee.

1980. aastatel leiti, et on vaja andmebaasi vormi, mis oleks kasulik automaatselt taksonoomiate, seletuste jms tegemiseks. Arvuti poolt loetavaid sõnastikke hakati kasutama erinevate semantiliste hierarhiate ehitamiseks. Võtmesõnaks sai *leksikaal-semantiline andmebaas.*

3. Leksikaalsed andmebaasid

Leksikaalses andmebaasis (*lexical database,* LDB) kui arvutileksikonis seovad andmebaasi kirjeid, antud juhul leksikaalseid üksusi, omavahel eri tüüpi viidad, antud juhul semantilised suhted (Calzolari 1990). Nii andmebaasis sisalduvad andmed kui ka andmebaasi struktuur on esitatud täiesti eksplitsiitselt ning tänu sellele on võimalik koostada paindlikult liigendatud päringuid.

Maailmas loodud semantiliste arvutileksikonide seas ringi vaadates jääb silma tuntuim – WordNet. Järgnevalt anname ülevaate WordNet'ist ja eestikeelsest wordnet-tüüpi sõnastikust: eesti WordNet'ist.

3.1. WordNet

Princetoni Ülikoolis on loodud leksikaal-semantiline andmebaas **WordNet**[3] (WN), mida loojad iseloomustavad kui "leksikaalsete viidete süsteemi, mille ülesehitus põhineb psühholingvistilistel teooriatel inimpsüühika leksikaalsest organisatsioonist ja mälust."(Miller jt 1990; Fellbaum 1998).

WordNeti elementaarosake on sünonüümirida – **sünohulk** (ingl *synonym set, synset*), mille moodustavad ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid. Teiseks WordNet-tüüpi tesauruse eristavaks tunnuseks on sünohulki ühendavad erinevad semantilised seosed, peamiselt hüpo- ja hüperonüümia, antonüümia, osa-terviku suhted, põhjuslikkus- ja rollisuhted jpm, ühtekokku ligi 60 erinevat suhtetüüpi.

[3] Princentoni WordNet'i koduleheküIg ja kasutajaliides http://wordnet.princeton.edu/ (nov 2005).

3.2. Eesti WordNet

Eesti WordNeti (EstWN) ehk TEKsauruse[4] põhjalik kirjeldus on antud autorite eelmistes artiklites (vt Orav, Vider 1998; Vider jt 2000; Orav, Vider 2005). Hetkel (nov. 2005) on sünohulki 11 000 ringis – põhiliselt substantiivi- (67%) ja verbimõisted (26%), kuid vähesel hulgal ka adjektiive (3%) ja pärisnimesid (4%). Üle poole sünohulkadest on üheliikmelised, veidi üle veerandi on kaheliikmelisi sünohulki ja 3- kuni 9-liikmelisi sünohulki on 19% kõigist sünohulkadest. Kõigist tesauruses esitatud leksikaalsetest üksustest (sõnadest ja sõnaühenditest) 80% on esindatud ühe tähendusega, mitmetähenduslikest (kaks ja enam tähendust) üksustest omakorda 68% moodustavad kahe tähendusega esindatud sõnad, kusjuures verbid on polüseemsemad kui ülejäänud sõnaliigid.

Semantilisi seoseid on ühel sünohulgal üle kahe, domineerivad hüpo- ja hüperonüümiasuhted, kuigi kasutusel on 43 erinevat semantilise suhte tüüpi.

Leksikaalselt põhineb loodav tesaurus olemasolevatel traditsioonilistel sõnaraamatutel (peamiselt "Eesti kirjakeele seletussõnaraamatul") ja tekstikorpusel[5] (mis annab teavet sõnakasutusest), seega võib semantilist informatsiooni, mida andmebaas sisaldab, pidada keelelisel teadmisel, mitte maailmateadmisel põhinevaks informatsiooniks.

4. Leksikaalsed teadmusbaasid

Püüdlused keele semantilise struktuuri esitamiseks on universaalsed: töötada välja kindel esitusviis, semantiline keel, mille abil võiks kirjeldada mistahes sõna vm keeleüksuse tähendust (Õim 1974:146). See püüdlus realiseerub suures vajaduses uute sõnatähenduste kirjeldusmeetodite järele, sest olemasolevad leksikaalsed andmebaasid ei anna piisavalt põhjalikku infot kontseptuaalsel tasandil. Sellest johtuvalt on leksikaalsete andmebaaside kõrval üha enam hakatud rääkima leksikaalsetest teadmusbaasidest (*lexical knowledge base,* LKB).

---

[4] TEKsaurus on päringu kaudu kasutatav ka veebis: http://www.cl.ut.ee/ressursid/teksaurus/ (nov 2005).

[5] Peamiselt kasutatakse Tartu Ülikoolis loodud kirjakeele tekstikorpust TÜKK: http://www.cl.ut.ee/ressursid (nov. 2005)

Üks peamisi erinevusi leksikaalsete teadmusbaaside ja leksikaalsete andmebaaside vahel on esimeste võimes esile tuua üldistusi ja tuletada järeldusi. Leksikaalne andmebaas võimaldab lihtsalt esitada andmeid sõnahaaval ning teeb võimalikuks nende andmete otsimise. Näiteks on inimese jaoks tavaline, et sõnad nagu *klaas, kruus, kann* võivad tähistada mitte ainult teatud nõusid, vaid ka vedeliku kogust, mis neisse mahub. See on kogu vastava semantilise sõnaklassi üldine omadus ja järelikult peaks selline üldistus – selle võimalikkus – ka arvutileksikonis kajastuma. Leksikaalne andmebaas seda ei võimalda.

Nagu öeldud, on arvutileksikonid arenenud teoreetilise lingvistika leksikoni- ja üldisemalt semantikakontseptsioonide mõjul. See seos tiheneb kahtlemata veelgi tulevikus, eriti näiteks leksikaalsete teadmusbaaside loomisel, kus on tarvis teoreetilises semantikas välja töötatud üldistus- ja järeldusmehhanisme. Tutvustame seda seost teoreetilisest lingvistikast välja kasvanud **freimisemantika** näite varal.

Semantilised freimid on situatsioonitüüpide (nt söömine, eemaldamine, uurimine jne) skemaatilised esitused koos hulga osaliste ja teiste kontseptuaalsete rollidega, mida nähakse selliste situatsioonide komponentidena[6]. Sõnad realiseerivad selliseid situatsioone erinevatest vaatepunktidest, fokuseerides selle erinevaid komponente (Fillmore, Atkins 1992).

Freimi mõiste toodi keeleteadusesse teatud kindlas ideoloogilises kontekstis: freimides nähti eelkõige vahendeid, mille abil puhtkeelelisi teadmisi saaks siduda oluliste mittekeeleliste argiteadmistega. Charles Fillmore'i esitatud tüüpnäide on 'ostma'. Kui seletussõnaraamat annab sellele paarisõnalise seletuse, siis maailmateadmistele tuginedes tuleb esitada kogu situatsioon: kui keegi ostab, siis keegi müüb; müüb midagi, millel on ostjale väärtus; selle väärtuse eest ostja maksab müüjale midagi jne (Fillmore, Atkins 1992). Samasugused kirjeldused peavad põhimõtteliselt olemas olema igasuguste tegevuste, situatsioonide, omaduste jm tunnuste kohta. Siin tulevadki appi freimid (Fillmore 1985).

Konkreetsemalt on freimide kasutamist seni seostatud leksikoniga, sõnade tähenduste kirjeldamisega. Seetõttu on lõviosa teoreetilistest diskussioonidest ja ka üksiknäidete käsitlustest seotud frei-

[6] http://framenet.icsi.berkeley.edu/ (nov 2005)

mide käsitlemisega leksikaalse semantika vahendina – freimid kui leksikaalse semantika uus kontseptuaalne vahend, freimid kui vahend, mille abil sõnade tähenduste kirjeldustesse saab sisse tuua relevantseid argiteadmisi, mis seostuvad sõna poolt tähistatava situatsiooniga (Fillmore 1977).

4.1. FrameNet

Võib väita, et kuigi freimisemantilise analüüsi põhiideed on pärit 1970. aastatest (vt nt Fillmore 1975), pole freime sõnatähenduste analüüsimiseks laialdaselt kasutusele võetud – uurimusi selle kirjeldusmeetodi efektiivsuse ja parima esitusviisi kohta alles tehakse. Freime sõnatähenduste kirjeldusmeetodina vajavad eelkõige arvutisõnastikud. Parim näide freimide kasutusest on FrameNet'i projekt[7], kus valitud sõnadele on püütud teha põhjalikku freimisemantilist analüüsi, ülejäänute puhul on leitud algseks analüüsiks vajalikud freimielemendid.

FrameNet (FN) on freimisemantikal põhinev projekt, mille käivitas Charles Fillmore California Ülikoolist Berkeleys 1997. a. Hetkel (nov 2005) sisaldab andmebaas umbes 8900 leksikaalset üksust, millest rohkem kui 6100 on täielikult märgendatud ja jagatud 625 semantilisse freimi ja varustatud ligikaudu 135 000 märgendatud näitelausega[8].

Projekti kallal on ametis projektijuhid, programmeerijad, uurijad ja leksikograafid ning töö on jagatud järgmiselt. Projektijuht esitab katselise semantilise freimi kirjelduse, kus on määratud sõna semantilised rollid ja tema grammatilised realisatsioonid. Kasutades kindlat hulka märgendeid, kontrollivad uurijad süstemaatiliselt iga sõna kasutust erinevates korpustes ja leiavad näiteid andmebaasi jaoks. Leksikograafid valmistavad ette sisendi lõpliku kuju, mis sisaldab lemmat, viidet tema freimile, esinemisvõimaluste loendit ja iga võimaluse kohta illustreerivat näitelauset. Kogu töö lihtsustamiseks on loonud programmeerijad tarkvara.

---

[7] http://framenet.icsi.berkeley.edu/ (nov 2005)

[8] Andmed pärinevad FrameNet'i kodulehelt http://framenet.icsi.berkeley.edu/ (nov 2005).

Projektis pööratakse esmajoones tähelepanu sõnade prototüüpsele kasutusele, sest esialgsed uurimused on näidanud, et on raske koostada freime sõnade metafoorsete kasutuste kohta.

4.2. Näiteid eestikeelsetest freimidest

Peale FrameNet'i projekti ei leidu keeleteaduses teisi laiaulatuslikke katseid kasutada freime empiirilise keelematerjali kirjeldamises ega süstemaatilisi käsitlusi freimi mõiste kasutamise tingimuste ning tagajärgede kohta (Õim, Saluveer 2002). Sama kehtib eesti keeleteaduses, kus freimisemantikal põhinevaid uurimusi on ainult üksikud katsed.

Kuna käesoleva artikli maht ei võimalda esitada näitlikustamiseks kõigi sõnaliikide freimisemantilist esitust, käsitleme ülevaatlikult ainult verbide *rääkima* ja *kõnelema* ja omadussõna *temperamentne* freime. Valik just nende sõnade analüüsimiseks on juhuslik[9].

4.2.1. Verbide freimid. Üksiksõnade tasemel on seni enim tegeldud eesti keeles tegusõnadega kui tegevusfreimi realiseerijatega. Nii on üks käesoleva artikli autoritest Heili Orav (1998) käsitlenud direktiivverbe ja hilisematest töödest väärib äramärkimist Katrin Pükke (2005) bakalaureusetöö liikumis- ja paiknemisverbidest. Tegevusfreimides kajastuvad muutusi tekitavad sündmused, millel on eeltingimused (situatsioon, kus tegevus saab üldse toimuda), on vahetu tegu ise (mis muudab) ja tagajärjeks olev uus situatsioon (mis on muutunud).

Järgnevalt püüame kirjeldada üleminekut wordneti-tüüpi leksikaalsest andmebaasist freimisemantilisele teadmusbaasile verbide *rääkima* ja *kõnelema* näitel. Eesti WordNetis on verbil *rääkima* kokku kuus tähendust ja verbil *kõnelema* kolm tähendust, nende ühisosa on kolmes sünohulgas – {rääkima(1), kõnelema(1)}, {rääkima(2), kõnelema(2)}, {rääkima(4), kõnelema(3)}. Nopime ühestatud sõnatähenduste korpusest vastavate tähendusnumbritega laused ja leiame uuritavate verbitähenduste lausemallid. Saadud süntaktilised kirjeldused on FrameNeti leksikaalsete kirjete (*Lexical Entry, LE*) aluseks.

---

[9] Põhjus seisneb autorite doktoritööde uurimisobjektis: Kadri Videri teema hõlmab sagedasemaid verbe ja Heili Orava teema isiksuseomadusi, mida eesti keeles antakse edasi valdavalt omadussõnadega.

Üks leksikaalne kirje on sisuliselt sama, mis sõnatähendus EstWN-s või selle põhjal ühestatud tekstikorpuses. Edasi on võimalik valida kahe erineva tee vahel: (a) vaadelda uuritavate sõnatähenduste tõlkevastete (WN1.5 põhjal) esindatust FrameNetis ja kontrollida, kas mõni neist sobib oma (kasutus)freimi poolest kirjeldama ka eestikeelset verbitähendust; (b) püüda leida sobiv freim semantiliste tunnuste põhjal otse FrameNetist.

Freim ise koosneb definitsioonist ja freimielementidest (*FE*), mis omakorda jagunevad tuumelementideks (*core*) ja perifeerseteks elementideks (*non-core*). Ideaalis on iga elemendi kirjelduse juures ka sellele freimile vastav näitelause. Freimi kirjesse kuuluvad ka freimi realiseerivad leksikaalsed üksused (*Lexical Unit, LU*), mille seotud hulk võib olla vägagi arvukas. Siin ilmnebki WN ja FN leksikaalse organisatsiooni erinevus kõige selgemalt, sest ühe freimi leksikaalsete kirjete seas võib põhimõtteliselt peale sünonüümide esineda ka hea hulk hüponüüme.

Eelnevalt mainitud *rääkima* ja *kõnelema* sünohulgad paistavad sobivat üsna hästi kolme FN freimiga (Tekstiloome, Vestlemine, Sedastus), mida on võimalik tõestada vastavaid näitelauseid analüüsides.

Freimi nimi: **Tekstiloome;** LE: {rääkima(4), kõnelema(3)}

Definitsioon: Autor loob Teksti, ükskõik kas kirjaliku või suulise, mis sisaldab tähendusega keelelisi üksusi, võib olla suunatud kindlale Vastuvõtjale.

FE: Tuumelemendid: Autor[semantiline_tüüp:aistiv], Tekst

FE: Perifeersed elemendid: Vastuvõtja[semantiline_tüüp:aistiv], Komponendid, Kirjeldav, Vorm, Instrument[semantiline_tüüp:asi], Viis, Vahend, Koht[semantiline_tüüp:koht], Eesmärk, Ajend, Aeg[semantiline_tüüp:aeg]

(1)
... kas suutsin talle$_{\text{Vastuvõtja}}$ seekord$_{\text{Aeg}}$ midagi arukat$_{\text{Tekst}}$ rääkida, ... (tkt0033)

(2)
... ega ole ühtegi inimest, kellele$_{\text{Vastuvõtja}}$ võiks südame tühjaks$_{\text{Eesmärk}}$ rääkida. (tkt0112)

Freimi nimi: **Vestlemine;** LE: {rääkima(1), kõnelema(1)}

Definitsioon: Rühm inimesi vestlevad, kuid keegi neist pole konkreetne Kõneleja või Vastuvõtja, iga osavõtja on kord kuulaja, kord kõneleja rollis. Eesmärgiks on vestlemine üldiselt, mitte otsustamine või infovahetus või vaidlemine.

FE: Tuumelemendid: Vestluskaaslane1, Vestluskaaslane2, Vestluskaaslased

FE: Perifeersed elemendid: Kirjeldav, Kestus[semantiline_tüüp: kestus], Keel, Viis, Vahend, Suhtlusvahend, Koht[semantiline_tüüp: koht], Eesmärk, Aeg[semantiline_tüüp:aeg], Teema

(3)
$\text{Ema}_{\text{Vestluskaaslane1}}$ kõneles $\text{minuga}_{\text{Vestluskaaslane2}}$. (tkt0001)

(4)
$\text{Kasahhitarid}_{\text{Vestluskaaslased}}$ kõnelesid Smerdjakovi poole pilke $\text{heites}_{\text{Viis}}$ eesti $\text{keeles}_{\text{Keel}}$. (tkt0073)

(5)
$\text{Nad}_{\text{Vestluskaaslased}}$ rääkisid $\text{autodest}_{\text{Teema}}$, ... (tkt0055)

Freimi nimi: **Sedastus;** LE: {rääkima(2), kõnelema(2)}

Definitsioon: Freim sisaldab verbe ja nimisõnu, millega Kõneleja suhtlusaktis edastab Sõnumi Vastuvõtjale, kasutades keelt.

FE: Tuumelemendid: Kanal, Sõnum, Kõneleja[semantiline_tüüp:aistiv], Teema

FE: Perifeersed elemendid: Vastuvõtja[semantiline_tüüp:aistiv], Määr, Sisemine_põhjus, Viis, Vahend, Sündmus, Korduvus, Koht[semantiline_tüüp:koht], Aeg

(6) $\text{Sergei}_{\text{Kõneleja}}$ oli kõnelnud, et...

(7)
Täna_{Aeg} räägiti, et... (tkt0030)

4.2.2. Omadussõnade freimid. Edasi võtame vaatluse alla "omadusfreimid". Omadusfreimid erinevad muutumis- või tegevusfreimidest, sest omadusi ja ka objekte (millega on tegu enamiku nimisõnade puhul) tuleb kirjeldada teisiti, neis puudub tavaliselt muutuse ja selle tagajärje idee. Omadusfreimid on pigem suguluses seisundifreimidega, kus määravad elemendid on seisundis olija, seisund ja selle võimalikud mõjud teistele situatsioonis osalejatele. Viimast komponenti ei pruugi kõigis freimides alati eksisteerida, aga inimese iseloomu ja selle omadusi kirjeldavates freimides oleks see ilmselt kohustuslik. Seos teiste situatsioonis osalejatega, st teiste inimestega on olemas kas või ainult selles mõttes, et inimese iseloomustamisel on alati tegu hinnanguga, mitte inimtegevust kirjeldava väljendiga. Kui inimene iseloomustab kedagi (nt *ta on sõbralik, agressiivne, viisakas*), siis

objektiivselt edastab ta informatsiooni, kuid subjektiivselt lisab infole oma suhtumise, millega taotleb omapoolset mõju kuulajale.

Isiksuseomadusi uurides tekitab probleeme muidugi küsimus, millised on hinnangu kategooriad ehk väärtused. On need jagatavad positiivseks või negatiivseks? Mis hinnangu anname, kui ütleme kellegi kohta, et ta on vaikse iseloomuga? Või seltskondlik? Küllalt on, kui ütleme, et inimene on lärmakas (= negatiivne omadus). Seega hinnang tuleneb käitumise põhjal antavast klassifitseeringust, mida tuleb mõista seoses taustsündmustega ja sotsiaalsete ootustega (Vainik, Orav 2005) ja mis on suurele osale keelekogukonnast üheselt mõistetav ning teada. Siiski leidub mõisteid, kus hinnangu väärtus tuleneb individuaalsest arusaamast.

Ka siia valitud näite – temperamentsuse ja tema sünonüümide (*emotsionaalne, tujukas, kapriisne, keevavereline, kirglik, äge, impulsiivne, pöörane, ülevoolav* jne) – puhul tekib uurijal küsimus neisse väljenditesse peidetud hinnangu loomusest: millised väljendid peegeldavad taunitavat ja millised aktsepteeritavat omadust? On oluline, et need komponendid kajastuksid freimikirjelduses, kuna eristavad mõistete tähendusi ning tähendusvarjundeid.

Seega on eelnevat arvestades võimalik fikseerida situatsioon, kus on HINDAJA ja freimi sisu on tema HINNANG ja selle HINNANGU OBJEKT: konkreetsemalt inimene (= Agent (A)) ja tema emotsioon(id).

Kuna omadusfreimi esitamiseks puudub eeskuju, mida järgida, esitame n-ö freimiskeemi, mis on koostatud suhteliselt vabas vormis:

A on TEMPERAMENTNE, kui (prototüüpiliselt):
a) A tunneb emotsiooni või emotsioone
b) A väljendab oma emotsioone
c) hindaja arvates tunneb ja väljendab A oma emotsioone keskmisest kiiremini (seos emotsioonide vaheldumise ajaga) või keskmisest rohkem (seos emotsioonide väljendamise hulgaga) võrreldes kellegi teis(t)ega.

Väärtus: hindajapoolne hinnang Negatiivne või Positiivne, sõltuvalt avaldumisvormist ja kontekstist (kuid mitte neutraalne!); leksikaalne realisatsioon: *temperamentne, emotsionaalne* vms.

Toodud näide osutab, et omadussõnade puhul on üsna keerukas leida kõiki olulist rolli mängivaid aspekte. Seega võib ainult täheldada, et

iseloomuomaduste analüüsimisel ei pruugi freimisemantika olla kõige efektiivsem vahend sõnavara semantiliseks kirjelduseks, kuna arusaam iseloomust pole alati piisavalt täpne ja üheselt mõistetav ega kajasta inimpsüühika ja -käitumise kogu nüansirikkust.

5. Kokkuvõte

WordNet-tüüpi sõnastiku loomine avas keeleuurimises uusi aspekte, kus püüti ühendada sõnavara kontseptuaalseid, leksikaal-semantilisi ja leksikograafilisi dimensioone. Kuid tundub, et nii selle suuna teoreetiline kui ka praktiline areng on veel üsna poole peal, sest pole selge, kuidas leksikaalseid andmeid põhjalikult ja ühtmoodi kirjeldada. Seda kinnitab ka vastavasisulise kirjanduse nending, et otsingud ikka veel kestavad. Küsimus, mis vajab uurimist on järgmine: kui sügavalt peaks sõnavara olema esitatud kontseptuaalsel tasandil (nt mis rolli määrab sõnaliigi eristus) või iseloomustatud organisatsioonilisel tasandil (semantilised suhted)?

Eelnevalt tutvustasime FrameNet'i projekti, mis on osaliselt inspireeritud WordNet'ist ja esitab keeleandmeid märksa põhjalikumalt eriti kontseptuaalsel tasandil. Kas WordNet'i ja FrameNet'i lähenemised keeleandmetele võiksid olla ühendatud ja anda efektiivse kirjeldusviisi sõnatähenduste esitamiseks? See oleks kindlasti üks võimalus, kuid oma artikliga seadsime eesmärgiks vaid esitada mõned näited, mis ilmestavad selget vajadust edasiste uuringute järele.

Kirjandus

Calzolari, Nicoletta 1990. Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base. – Using On-line Resources to Build a Lexicon. Ed by U. Zernik. Hillsdale, NJ: Erlbaum, chapter 8, 191–208.

Fellbaum, Christiane 1998. Introduction. – WordNet: An Electronic Lexical Database. Ed by Ch. Fellbaum. MIT Press, 1–19.

Fillmore, Charles 1975. An alternative to checklist theories of meaning. – Papers from the First Annual Meeting of the Berkeley Linguistics Society, 123–132.

Fillmore, Charles 1977. Scenes-and-frames semantics. – Linguistic Structures Processing. Fundamental Studies in Computer Science, 59. Ed by A. Zampolli. North Holland Publishing.

Fillmore, Charles 1985. Frames and the Semantics of Understanding. – Quaderni di Semantica, vol 6, 222–254.

Fillmore, Charles, Atkins, B. T. Sue 1992. Towards a Frame-based Lexicon: the Semantics of RISK and its Neighbors – Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization. Ed by A. Lehrer, E. F. Kittay. Hillsdale, NJ: Erlbaum, 75–102.

Langemets, Margit 2000. Sõnaraamatu arvutilingvistiline analüüs. Eesti Keele Instituut. Magistritöö, käsikiri.

Miller, George, Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, Miller, K. J. 1990. Introduction to WordNet: An On-line Lexical database. – International Journal of Lexicography 3, 235–312.

Muischnek, Kadri, Orav, Heili, Kaalep, Heiki-Jaan, Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Toim U. Talvik. Tallinn: Eesti Keele Sihtasutus. 86.

Orav, Heili, Vider, Kadri 1998. Sõna tasandilt mõiste ruumi. – Keel ja Kirjandus, 1, 57–64.

Orav, Heili 1998. Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. Magistritöö. Tartu Ülikool, üldkeeleteaduse õppetool, käsikiri.

Orav, Heili, Vider, Kadri 2005. Estonian wordnet and Lexicography. Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography. May 2–4, 2002 at the University of Copenhagen. Ed by H. Gottlieb, J. E. Mogensen, A. Zettersten. Tübingen: Max Niemeyer Verlag, 549–555.

Pükke, Katrin 2005. Liikumise ja paiknemisega seotud verbide semantika arvutirakenduste jaoks. Bakalaureusetöö. Tartu Ülikool, üldkeeleteaduse õppetool, käsikiri.

Vainik, Ene, Orav, Heili 2005. Tee tööd ja näe vaeva, ...aga ikka oled vihane. – Keel ja Kirjandus 4, 257–277.

Vider, Kadri, Kahusk, Neeme, Orav, Heili, Õim, Haldur, Paldre, Leho 2000. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim T. Hennoste. Tartu, 127–152.

Õim, Haldur 1974. Semantika. Tallinn: Valgus.

Õim, Haldur, Saluveer, Madis 2002. Freimid keelekirjelduses. – Akadeemia 12, 2663–2678.

Sõnatähendused ja nende ühestamine tekstides[1]

Kadri Kerner, Kadri Vider, Neeme Kahusk
Tartu Ülikool

1. Sissejuhatus

Tähendusel on psühholoogiline alus, õigupoolest on andmeid selle kohta, et selline alus on olemas. Siiski on leksikograafidki teadlikud sellest, et tähenduse suhtes pole üksmeelele jõutud. Tähenduse üle arutles juba Aristoteles, kes väitis, et sõnad vastavad kontseptidele ja objektidele. Kuid Wittgenstein oli filosoof, kes rõhutas, et sõnade tähendus on vaid nende kasutuses ning (lingvistiline) tähendus on kasutuse funktsioon, millesse keeleväljendid asetatakse.

On lekseeme, millel on rohkem kui üks tähendus, sellised sõnad on mitmetähenduslikud ehk polüseemsed (Karlsson 2002: 243). Polüseemia võib tuleneda regulaarsest tähenduse laiendamisest, seda võib põhjustada kontekst, või võib polüseemia olla tingitud metafoorist või metonüümiast (Ravin, Leacock 2002: 23).

Seega saab esitada kaks põhimõtteliselt erinevat küsimust:

1. Mida see sõna tähendab?

Polüseemsete sõnade korral annab vastus sellele küsimusele sõna põhitähenduse. Sõna põhitähendus (otsene tähendus, denotatiivne tähendus – kõiki neid võiks siin võrdseteks lugeda) on selline, mis aktiviseerub mentaalses leksikonis kontekstita algvormi peale. Võtame näiteks sõnad *koor* ja *hang*. Küsides, mida need sõnad tähendavad, võime saada erinevatelt vastajatelt täiesti erinevad seletused, millest ilmneb, et keelekasutajate teadvuses vastavad neile sõnadele erinevad kontseptid ehk mõisted ja tegu on hoopistükkis homonüümiaga. Nagu näha, aitab see küsimus eristada homonüümiat polüseemiast.

[1] Artikkel ja selle aluseks olnud uurimistöö on edenenud ETFi grandi 5534 "Tähenduspõhise keeletöötluse ressursid ja töövahendid eesti keele jaoks" (2003–2006) ja sihtfinantseeritava teadusteema nr 0182541s03 "Eesti keele arvutimudelid ja keeleressursid: teoreetilised ja rakenduslikud aspektid" (2003–2007) toel.

2. Mis tähenduses seda sõna kasutatakse? See on Wittgensteini lähenemine sõna tähendusele, sest tema määratleb sõna tähendust vaid sõna kasutamise kaudu.

Sõnatähenduse (ingl *word sense*) täpse definitsiooni üle vaieldakse palju. Adam Kilgarriff kirjutab oma artiklis "I don't believe in word senses" (1997), et sõnatähenduste ühestamine eeldab erinevate tähenduste olemasolu ning erinevad tähendused on teada-tuntud komistuskohaks. Artikli autor väidab, et mõistel sõnatähendus puuduvad põhialused.

Siin me ei räägi sõna tähendusest leksikoloogilises mõttes, vaid tähendustest, mida kannavad sõnad kasutuses – seega võivad nad olla sõnaraamatu seisukohalt ehk denotatiivsest tähendusest erinevad ja hõlmata kõiki tähenduse nihestamise või varjundamise võimalusi metafooridest alates ja idiolektiliste ning stiililiste nüanssidega lõpetades.

2. Sõnatähenduste ühestamine

Sõnatähenduse ühestamine (STÜ, ingl *word sense disambiguation*, lühendina WSD) on keeletehnoloogia ülesanne, mille puhul tuleb otsustada, millises tähenduses sõna antud kontekstis esineb. Keele automaatanalüüsis on STÜ lause ja väiksemate süntaktiliste konstruktsioonide semantilisele analüüsile eelnev etapp.

Sõnatähenduste ühestamist peetakse loomuliku keele töötlemise juures üheks kesksemaks probleemiks (Ide, Veronis 1998), see on vaheülesanne, mida vajavad näiteks masintõlge, infootsing, kõnetöötlus jt.

Infootsingu puhul on suureks abiks seegi, kui suudame eristada homonüümseid vorme või lemmasid ja saada vastuseid ühe kindla tähenduse kohta. Kõnesünteesis pole võimalik õigesti hääldada näiteks sõna *palk*, kui me ei tea, kas seda on vaja palataliseerida või mitte. Samuti pole võimalik tõlkida eesti keelest inglise keelde sõna *naine*, kui me ei tea, kas tegemist on lihtsalt naisterahvaga (*woman*) või abielunaisega (*wife*) (Kahusk, Kaljurand 2002).

Kui me jätame kõrvale vormihomonüümia, jäävad meil alles leksikaalne homonüümia ja polüseemia. Mis on homonüümia ja polüseemia vahe, pole antud juhul oluline. Tähtis on see, et on eristatud vähemalt homonüümsed tähendused ja et need erisused on fikseeritud mingis leksikonis.

Sõnatähenduste ühestamise protsess keeletehnoloogias hõlmab endas kahte sammu.

1. Teha kindlaks kõik erinevad tähendused kindlas kontekstis. Selle alamülesande lahendamisel kasutatakse:

a) elektroonilist sõnastikku, mis esitab sõnade kõikvõimalikud tähendused;

b) kategooriate, omaduste gruppi või seoseis olevaid sõnu (nt sünonüüme nagu tesauruses);

c) tõlke- või ülekandesõnastikku, mis hõlmab endas teise keele tõlkeid.

2. Teha kindlaks, milline tähendus on kõige sobivam antud kontekstis. Püütakse leida meetodit, mis kirjeldaks vastavust sõna võimalike kontekstide ja võimalike tähenduste vahel. Kasulikku andmestikku mingi sõna seostamiseks tema tähendusega saab, kui kasutada keeleväliseid teadmiste allikaid: leksikaalsed, entsüklopeedilised ressursid (Ide, Veronis 1998).

Ülesanne on põhimõtteliselt sama mis morfoloogilise või süntaktilise ühestamise korral, kuid sisult märksa keerulisem, sest semantika jaoks puuduvad sarnased traditsioonilised ja samas formaliseeritud käsitlused, nagu need on olemas morfoloogia ja süntaksi jaoks.

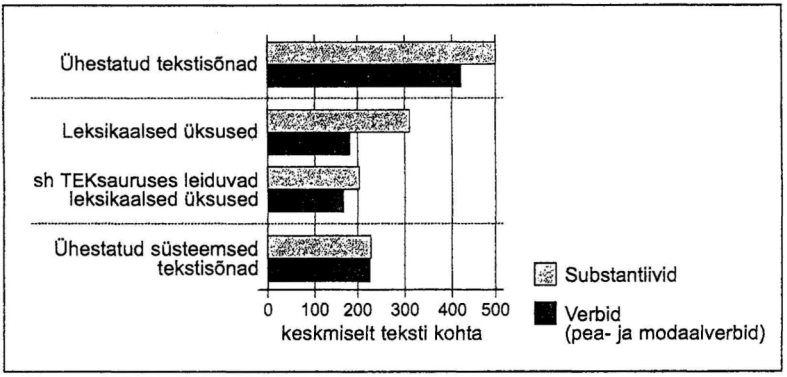
Semantilise ühestamise programmi loomisel on sageli eelduseks käsitsi ühestatud sõnatähenduste korpus, mille najal programmi testida ja arendada. Lisaks võib STÜ korpus huvi pakkuda nii sõnasemantikutele kui ka tähenduspõhiste sõnastike loojatele (selle kohta vt Orav, Vider 2002).

3. Eesti sõnatähenduste ühestatud korpus

STÜ korpus sisaldab annoteeritud (ühestatud) sisusõnu. Annoteeritavate sisusõnade leidmine omakorda sõltub nii tekstide morfoloogilise eeltöötluse kui kasutatava sõnatähenduste leksikoni sõnaliigituse printsiipidest (Muischnek, Vider 2005).

Leksikoniks, millest sisusõnade tähendusi otsitakse, on TEKsaurus[2], mille tähendusnumbrite põhjal ühestati tekstides ainult nimisõnu (substantiive), põhiverbe ja modaalverbe. Üks mis kindel – esialgu pole lootust leida korpusest ühestatud funktsioonisõnu ega lause/fraasi semantilist annotatsiooni.

[2] http://www.cl.ut.ee/ressursid/teksaurus/



Joonis 1. Ühestatud sisusõnad ja nende tähendused STÜ korpuses

Iga teksti on ühestanud teineteisest sõltumatult vähemalt kaks inimest; tulemused on hiljem ühtlustatud ühestajate läbirääkimiste teel.

Tekstisõna tähendus märgendati talle vastava numbriga TEKsaurusest. Kui sõna leksikonist puudus, märgiti tähendusnumbriks '0'. Kui sõna küll esines leksikonis, aga puudus konkreetsele kasutusjuhule vastav tähendus, märgiti tähendusenumbriks '+1'. Erinevaid tõlgendusi tekstitähendustele oli ühestajatel algul rohkesti, kogemuste kasvades, ja peale sagedamini esinevate erimeelsuste analüüsi, jäi neid vähemaks ning töö edenes jõudsamalt, kuid keskmiselt ühestati erinevalt umbes 20% tekstisõnadest.

Praegu on enamus eesti keele sõnatähenduste tasandil ühestatud 100 000-sõnelise korpuse tekstidest ilukirjandustekstid, mis on pärit eesti kirjakeele korpuse 1980. aastate ilukirjanduse alamkorpusest[3].

Ühe sõna analüüsi tulemus on sarnane sellele, mida on kirjeldatud morfoloogiliselt ühestatud korpuse[4] juures, lisaks on info sõne tähenduse kohta (semantiline info):

```
sõne tüvi + lõpp // morfanalüüs // semantiline info
```

Semantiline info sisaldab lemmat, käsitsi lisatud tähenduse numbrit, sünohulga (kirje) numbrit ja tähenduste arvu TEKsauruses.

[3] http://www.cl.ut.ee/korpused/baaskorpus/
[4] http://www.cl.ut.ee/korpused/morfkorpus/

4. Kui palju tähendusi sõnal on?

Siiani on üheks kesksemaks küsimuseks, kui peenelt peavad sõnatähendused eristatud olema. Sõnatähendused ei tohiks loomulikult olla ei liiga üle- ega ka liiga ala-eristatud. Erinevad rakendused vajavad erinevat eristatuse taset – mis on piisav näiteks lihtsamale masintõlkesüsteemile, pole samas eriti informatiivne leksikograafile. Ka arvutilingvistilised rakendused vajavad erinevat eristatuse taset: infootsing saaks peaaegu 100%-se täpsuse ka väga üldise eristatusega, kuid masintõlkesüsteemile on vaja siiski pisut detailsemat sõnatähenduste eristatust. Nancy Ide ja Yorick Wilks püstitavad oma artiklis "Making Sense about Sense" (2004) küsimuse: kui vähe informatsiooni sõna tähenduse kohta on vaja loomuliku keele töötluses? Nad leiavad, et praegused ressursid on selgelt liiga üle-eristatud ja vajavad ümbertegemist ning aktuaalseks on seega muutunud sarnaste sõnatähenduste klasterdamine ehk grupeerimine. Üks aktuaalne suund sõnatähenduste klasterdamisel on paralleelkorpuste (ja tõlkevastete) kasutamine, teine mingi olemasoleva leksikaal-semantilise ressursi, näiteks wordneti kasutamine.

Princetoni WordNetis[5] on umbes 20%-l sõnadel rohkem kui üks tähendus ning polüseemsel sõnal on keskmiselt kolm erinevat tähendust. Sarnased on polüseemianäitajad ka TEKsauruses, kus polüseemsetel sõnadel on keskmiselt 2,58 tähendust. Seega on kasulik uurida, kuidas ja millistele meetoditele tuginedes saab polüseemiataset vähendada ehk sarnaseid sõnatähendusi klasterdada.

W. Gale, K. Church ja D. Yarowsky (1992) leidsid sõnatähenduste ühestamise süsteemi väljatöötamisel ka seda, et sõna tähendus sõltub ühest kindlast tekstist. Kui sõna esineb kaks või enam korda ühes tekstis, siis tema tähendus on tavaliselt üks ja seesama kogu teksti piires. Enamasti kehtib see printsiip ka eesti keele nimisõnade kohta ning (harva esinevate) verbide kohta.

5. Tähendusklastrid

Kadri Kerneri 2004. aastal TÜ arvutilingvistika erialal kaitstud bakalaureusetöös (Kerner 2004) on eesmärgiks uurida käsitsiühestajate arvamuste erinevust eesti wordneti ehk TEKsauruse tähendusnumbrite põhjal. Arvamuste erinevuste uurimine annab võimaluse näha, millis-

[5] http://wordnet.princeton.edu/

te sõnade tähenduste osas on erimeelsused suured ja millised tähendusklastrid (sarnaste tähenduste grupid) tekivad. Samuti on püütud leida uusi reegleid ja juhiseid käsitsiühestajate jaoks.

Osa töös analüüsitud sõnade tähendusi moodustasid olulisi tähendusklastreid, osa mitte. Raskusi tekstisõnadele tähenduste omistamisel oli vähem, kui tesauruse sõnatähendused:

- ei sisalda autohüponüümia suhteid. Autohüponüümia on suhe, milles üks sõna tähendus on teisele sama sõna tähendusele hüponüümiks (Peters jt 1998);
- ei sisalda ko-hüponüüme (ingl *sisters*), mille puhul on sõnatähendustel üks ja seesama hüperonüüm (Peters jt 1998);
- ei sisalda osalist ko-hüponüümiat, mille puhul sõnatähenduste mingi hulk hüperonüüme kattuvad;
- kuuluvad väiksema polüseemsusega sõna tähenduste hulka. Kõrge polüseemiatasemega sõnade puhul on keerulisem üht ja ainsat sobivat tähendust määrata. Sageli tekib olukordi, kus kontekst lubab mitmeid võimalikke tähendusi või polegi otstarbekas teatud sõnatähendusi eristada (Vider, Orav 2003: 315–317).

Kui sõnatähendustel tekivad tähendusklastrid kindlate tähenduste vahel, on põhjust kahtlustada, et

- TEKsauruses on tähendused eristatavad, kuid tekstis mitte;
- TEKsauruse kirjel on esituses teatavad puudused: näidete puudumine, ähmane seletus, näidete ja/või seletuste ja/või sünohulkade kokkulangemine.

Tähendusklastrite uurimine annab võimaluse viidata mingi konkreetse sõnatähenduse ebaselgusele, mõni sõnatähendus kombineerub kõikide erinevuste seas või paljude teiste tähendustega. Juhul kui ühe sõna kõik tähendused saavad üksteisega koos esineda (peaaegu kõik tähendused kombineeruvad peaaegu kõigiga), võib oletada, et terve sõna tähendusjaotus on puudulik, ebaselge ning vajab täiendamist. See tendents ilmneb abstraktse ja/või laia tähendusampluaaga sõnatähenduste korral, näiteks sõnad *asi*, *aeg*, *jääma*.

Mõned sõnatähenduste uurijad (Kilgarriff 1998; Agirre, Martinez 2000) on viidanud probleemile, et ebaselged ja raskesti märgendatavad on tipmistesse sünohulkadesse kuuluvad sõnad. Samasugused tulemused on välja toodud ka Kadri Kerneri bakalaureusetöös eesti STÜ korpuse põhjal. Lisaks leiti, et tipmisesse sünohulka kuuluvat sõnatähendust on raskem teistest eristada.

Tähendusklastrid võivad olla ka oluliseks abiks automaatse sõnatähenduste ühestaja töös. Automaatsel ühestamissüsteemil on kergem valida sobivat ja õiget tähendusnumbrit, kui tähendused pole liialt üle-eristatud, s.t süsteemi töö kiirus ja täpsus suurenevad. Kui ühe sõna tähendused moodustavad mingeid kindlaid ja olulisi tähendusklastreid, on mõttekas need teatud keeletehnoloogiliste rakenduste jaoks liita üheks tervikuks ning anda neile üks tähendusnumber.

6. Kokkuvõte

Leksikaalses semantikas on kaks põhimõtteliselt erinevat küsimust: (1) Mida mingi sõna tähendab? ja (2) Mis tähenduses mingit sõna kasutatakse?

Keeletehnoloogias põletavaks muutunud sõnatähenduste ühestamise probleem puutub peamiselt kokku küsimusega (2). Probleemi lahendamiseks on vajalik (1) teha kindlaks mingi sõna tähendused erinevates kontekstides ning seejärel (2) omistada konkreetsele tekstisõnale kontekstiga sobiv tähendus. Esimeses etapis kasutatakse tähenduste määratlemiseks sageli mingit leksikoni, teise etapi tulemuseks on tihti sõnatähenduste ühestatud korpus, mida saab kasutada ka automaatse STÜ süsteemi treenimiseks.

Ühest vastust küsimusele, kui peenelt või täpselt peab ikkagi sõnatähendused eristama, pole veel antud ja erinevad keeletehnoloogilised rakendused vajavad erinevat eristatuse astet. Üldiselt võib oletada, et liiga üle-eristatud sõnatähendused kipuvad tekitama automaatses sõnatähenduse ühestamise programmis liiga palju müra ning süsteemi töökiirust aeglustama. Ka ühestatud korpuse loomiseks vajalik käsitsiühestamise ülesanne muutub liiga peente sõnatähenduste korral liiga keeruliseks – tekib palju eriarvamusi inimeste-ühestajate vahel. Üks võimalikke lahendusi oleks lubada tähenduste klastrite moodustamist.

Kirjandus

Agirre, Eneko, Martinez, D. 2000. Exploring automatic word sense disambiguation with decision lists and the web. – Proceedings of the 18th International Conference on Computational Linguistics. Saarbrucken, Germany, 11–19.

Gale, William, Church, Kenneth, Yarowsky, David 1992. One Sense Per Discourse. – DARPA Workshop on Speech and Natural Language. New York, 233–237.

Ide, Nancy, Véronis, Jean 1998. Introduction to the special issue on word sense disambiguation: the state of the art. – Computational Linguistics 24, 2–40.

Ide, Nancy, Wilks, Yorick 2004. Making Sense About Sense. – Word Sense Disambiguation: Algorithms, Applications and Trends. Ed by E. Agirre, P. Edmonds. Kluwer. Chapter 3.

Kahusk, Neeme, Kaljurand, Kaarel 2002. *Semyhe* tulemusi: kas tasub *naise* pärast WordNet ümber teha? – Tähendusepüüdja. Catcher of the Meaning. Toim R. Pajusalu, T. Hennoste, TÜ üldkeeleteaduse õppetooli toimetised 3, Tartu, 185–195.

Karlsson, Fred 2002. Üldkeeleteadus. Tõlkinud ja kohandanud R. Pajusalu, J. Valge, I. Tragel. Tallinn: Eesti Keele Sihtasutus.

Kilgarriff, Adam 1997. I don't believe in word senses. – Computers and the Humanities 32 (2), 91–113.

Kilgariff, Adam 1998. Gold standard datasets for evaluating word sense disambiguation programs. – Computer Speech and Language 12(4), Special Issue on Evaluation.

Kerner, Kadri 2004. Sõnatähendused tekstides ja tesauruses ühestajate erimeelsuste põhjal. Käsikiri. Tartu Ülikooli üldkeeleteaduse õppetooli bakalaureusetöö.

Muischnek, Kadri, Vider, Kadri 2005. Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis. – Eesti Rakenduslingvistika Ühingu aastaraamat 1 (2004). Tallinn: Eesti Keele Sihtasutus, 99–114.

Orav, Heili, Vider, Kadri 2002. Kas tesaurus ja tekstid lähevad kasutuses kokku? – Tähendusepüüdja. Catcher of the Meaning. TÜ üldkeeleteaduse õppetooli toimetised 3. Toim R. Pajusalu, T. Hennoste. Tartu, 297–303.

Peters, Wim, Peters, Ivonne, Vossen, Piek, 1998. Automatic sense clustering in EuroWordnet. – Proceedings of the First Interational Conference on Language Resources and Evaluation, Granada.

Ravin, Yael, Leacock, Claudia (eds) 2002. Polysemy: theoretical and computational approaches. Oxford University Press.

Vider, Kadri, Orav, Heili, 2003. Idee ja rakenduse vahe tesauruse näitel. – Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 14. Toim M. Langemets, H. Sahkai, M.-M. Sepper. Tallinn: Eesti Keele Sihtasutus, 313–322.

Sõnade *mees* ja *naine* kollokatsioonide võrdlemise võimalusi eesti, saksa ja inglise keele korpustes

Liisi Piits
Tartu Ülikool, Eesti Keele Instituut

1. Sissejuhatus

Siinses artiklis kirjeldan ma oma esimesi katseid võrrelda sõnade *naine* ja *mees* eesti kollokatsioone saksa ja inglise keele vastavate sõnade kollokatsioonidega (vt ka Piits 2004). Andmete töötlemisel ei kasutanud ma keerulist statistilist analüüsi, vaid piirdusin kõige lihtsamate meetoditega, lugedes kokku sõnade *mees* ja *naine* vasakule hargnevad kollokatsioonid ja reastades need sageduse järgi. Sageduse mõiste ongi tihti kollokatsiooni mõistega seotud. Kollokatsiooni nähakse kui sagedamini sõna naabruses asetsevat sõna (Singleton 2000; Stubbs 2001). Sõna ja tema kollokatsioonide vahelisi seoseid on defineeritud kui üht keele siduvatest jõududest, mis organiseerib sõnavara, arvestades, millised sõnad esinevad tüüpiliselt koos, ja näidates sõnaassotsiatsioonide võrke (Johnson 1999: 57). Artiklis lähtutakse eeldusest, et kollokatsioonides peegelduvad ka sõna tähendusnüansid. Sõnade võime omada erinevaid kollokatsioone võib olla väga varieeruv. Sara Mills viitab Carrolli ja Kowitzi uuringule, mille kohaselt hulk adjektiive nagu *tähtis, kuulus, rikas* esinevad inglise keeles palju sagedamini viitega mehele, samas kui naisele viitavad adjektiivid *hõivatud, ilus, kaunis* (Mills 2002: 210).

2. Uurimismaterjalist

Sõnad *mees* ja *naine* on tänuväärne uurimismaterjal eelkõige seetõttu, et need sõnad kuuluvad kümne eesti keeles kõige sagedamini esineva substantiivi hulka (Kaalep, Muischnek 2002: 154). Sageduse aspekt on eriti oluline arvestades minu kasutatud eesti keele korpuse suhtelist väiksust. Harvemini esinevate sõnade esinemuse kohta oleks sellise materjali põhjal raske järeldusi teha.

Sõnade *mees* ja *naine* eesti kollokatsioonid olid algselt kogutud kogu Tartu Ülikooli Eesti kirjakeele korpuse (TÜKK) eelmise sajandi ajakirjanduskorpuse tekstidest (1,8 miljonit sõnet), aga kuna

inglise ja saksa keele materjalid kajastasid peamiselt vaid viimase aastakümne keelt, siis tuli ka eesti keele materjal ainult 1990. aastate allkorpusest valida. Kui eesti keele tekstid oleks kajastanud kogu eelmise sajandi sõnakasutust, aga inglise ja saksa keele korpuste tekstid vaid sajandi lõpu sõnaesinemust, siis oleks see seganud andmete võrdlemist. Sõnade *mees* ja *naine* 30 sagedamini esineva eesti kollokatsiooni seas oleks olnud sõnu, mis 1990. aastateks olid juba kas vananenud või historismid. Näiteks sõnu *vaimulik*, *valla* ja *õpetatud* kasutati sõna *mees* täienditena just eriti 1890. ja 1900. aastate allkorpustes ja sõna *nõukogude* esines *naine* täiendina peamiselt aastakümnetel 1950–1980. Seega tuli niigi väikest korpust veel kriitilise piirini piirata, arvestades vaid viimase kümnendi tekste (maht 384,5 tuhat sõnet). Võrdluseks vajalikud andmed koguti korpusest 2002. ja 2003. aastal. Praegusel hetkel oleks võrdluseks võimalik kasutada ligi sada korda suuremat ajakirjanduskorpust (u 40 miljonit sõnet).

Sõnade *mees* ja *naine* saksa ja inglise keele kollokatsioonid saadi automaatselt Leipzigi ülikooli koduleheküljel (http://wortschatz.uni-leipzig.de) oleva sõna esinemisümbruse leidmise programmi abil, mis võimaldas teha otsinguid suurtest inglise ja saksa keele korpustest. Eesti, saksa ja inglise keele korpuste suurimaks erinevuseks jäigi maht: viimased olid sadu kordi suuremad. Lisaks olin võrdluse võimaluse suhtes skeptiline, kuna saksa ja inglise keele kohta saadud andmed sain automaatselt, aga eesti keele sõnade kollokatsioonide leidmine oli suures osas käsitsitöö.

3. Sõnade *mees* ja *naine* kollokatsioonide leidmisest

Algselt leidsin TÜKKi leheküljel oleva konkordantsi otsijaga korpusest kõik konkordantsiread, kus esinesid tähejadad *mees*, *mehe* ja *naine, nais*. Saadud andmetest eemaldasin kõik juhud, kus sõna esines liitsõna (*esimees, perenaine, naislendur, meessugu* jne) või tuletise (*meeskond, naiselik* jne) osisena. Lisaks tuli eemaldada näiteks kohanime *Häädemeeste* esinemisjuhud. Kui alles olid jäänud ainult konkordantsiread, kus sõnu *mees* või *naine* kasutati lihtsõnana, asusin leidma kollokatsioone.

Üks raskusi kollokatiivsete suhete mõõtmisel avaldub probleemis, kui kaugel võib asetseda sõna otsingusõnast, et seda saaks veel nimetada koosesinemiseks. Stubbsi sõnul on olemas konsensus, et olulised kollokatsioonid leitakse vahemaalt 4:4, st neli sõna otsin-

gusõnast vasakule ja neli sõna paremale (Stubbs 2001: 29). Kindlasti oleneb aga vahemaa suurus palju keele tüübist: Stubbs on oma järeldusele jõudnud rohkem analüütilist väljendusviisi kasutavat inglise keelt uurides, eesti keeles võiks optimaalne kollokatsioonide kaugus otsingusõnast väiksem olla. Lähtesõnast oleneb, kas tähenduslikud kollokatsioonid jäävad sõnast pigem vasakule või paremale. Näiteks kui lähtesõnaks on substantiiv, siis jäävad sõnaga süntaktiliselt seotud adjektiivid ja numeraalid vasakule, aga kui lähtesõnaks on numeraal või adjektiiv, siis tuleks kindlasti vaadelda neist paremale hargnevaid substantiivseid kollokatsioone.

Mina piirasin kollokatsioonide leidmise kauguse kolme sõnani vasakult, s.t vaadeldava sõna kollokatsioonidena lähevad arvesse vaid sõnad, mis ei esine kaugemal kui kolm sõna lähtesõnast. See tundub olevat optimaalne kaugus, kust kollokatsioone leida, sest kaugemal esinevatel sõnadel on väga harva seost vaadeldava sõnaga. Kollokatsioonid leidsin poolautomaatselt programmi *WordSmith Tools* abil. Aga saadud tulemust tuli käsitsi lemmatiseerida ja selle käigus oli otstarbekas välja jätta kõik sõnad, mis ei olnud lähtesõnaga grammatiliselt või tähenduslikult seotud. Viimane toiming oli vajalik eelkõige seetõttu, et lähtuvalt analüüsi aluseks oleva keelekorpuse suhtelisest väiksusest, võisid kollokatsioonide hulgas tooni hakata andma ka juhuslikud lähtesõnaga mitteseostuvad sõnad. Isegi sama sõnavorm võis kord laiendada lähtesõna ja teisal kuuluda hoopis teise osalausesse või fraasi.

Tihti võis mõne automaatselt leitud kollokatsiooni esinemissagedus kahaneda pärast käsitsi kontrolli poole võrra. Kollokatsioonide leidmise kauguseks oli kolm sõna vasakule, aga mida kaugemal asetses sõna otsitavast sõnast, seda suurem oli tõenäosus, et automaatselt leitud kollokatsioon ei olnud otsitava sõnaga süntaktiliselt ega semantiliselt seotud. Ilmnes seaduspärasus, et mida kaugemalt kollokatsioone otsida, seda enam kasvab erinevate sõnade hulk, aga seda vähem on nende hulgas suurema esinemissagedusega sõnu, ehk mida kaugemalt kollokatsioone otsida, seda enam võib leida sõnu, mis esinevad vaid korra. Kirjeldatud tendents toetab igati arvamust, et sõna sageduse arvestamine peaks välja praakima juhusliku sõnaümbruse.

Kollokatsioonide esinemissagedust väljendav arv ei sisalda sõnakasutust teoste pealkirjades, sest kahjuks tingis 1990. aastate

ajakirjanduskeele korpuse väiksus olukorra, kus sõna *naine* 30 sagedasema kollokatsiooni hulka oleks sattunud sõnu vaid ühe raamatut või filmi arvustava artikli tõttu. Kollokatsioonidena ei ole loendatud sõnu ajakirjade "Eesti Naine", "Nõukogude Naine", raamatute "Niskamäe naised", "Kojamehe naine" ega filmi ja kujutava kunsti teoste pealkirjadest "Kihnu naine", "Setu naised". Samuti pole arvestatud kollokatsioonide hulka nimesid: näiteks eraldati omadussõnaliste täiendite hulgast pärisnimi *Aus* jne.

4. Sõnade *mees, man* ja *Mann* kollokatsioonide võrdlus

Mõningad erinevused tulenesid veel asjaolust, et eesti keele korpusest oli leitud sõnade *mees* ja *naine* nii ainsuse- kui ka mitmusevormide kollokatsioonid, aga saksa ja inglise keele korpusest oli otsitud vaid ainsustüvega koosesinevaid sõnu. Kuna mitmed arv- ja asesõnad saavad esineda vaid koos mitmusevormiga, siis need inglise ja saksa kollokatsiooniloenditest puuduvad. Näiteks sõna *mees* ja *naine* eesti kollokatsioonide seas esinenud arvsõnad (mis eesti keeles esinevad küll koos ainsusvormiga) *kaks, kolm, neli, üheksa, tuhat* ning asesõnad (mis nõuavad ka eesti keeles mitmust) *need, kõik, palju, mõlemad, rohkem* nõuavad inglise keeles mitmusetüve *men* või *women*. Eelnevalt mainitud erinevustele vaatamata tegin katset sõnade *mees* ja *naine* kollokatsioone kolmes keeles võrrelda.

Tabel 1. Sõnade *man, mees* ja *Mann* 30 sagedasemat kollokatsiooni esinemissageduse järjekorras

	Inglise MAN	Eesti MEES	Saksa MANN
1.	NOOR (young)	SEE/ NEED	NOOR (jungen, -er)
2.	MUST (black)	ÜKS	TEMA/ NENDE (ihr, -em, -en)
3.	VANA (old)	EESTI	VANA (alte, -n, -r)
4.	ÜKS (one)	MÕNI	MINU (mein, -em, -en)
5.	VALGE (white)	OMA	TUGEV (starke, -en, -er)
6.	TEINE (another)	MEIE	VÄIKE (kleine, -en, -er)
7.	SURNUD (dead)	KAKS	TEIE (Ihr)
8.	VANEM (elderly)	NOOR	LIIGUTATUD (bewegte)
9.	PAREMAKÄELINE (right-hand)	KOLM	SEE/ NEED (diesen, -er)
10.	KODUTU (homeless)	SAMA	VÕIMSAIM (mächtigste, -en, -er)
11.	TUNDMATU (unidentified)	MU (MINU)	VANEM (älteren, -er)

	Inglise MAN	**Eesti MEES**	**Saksa MANN**
12.	PETIS (con-man)	UUS	TEINE (zweite, -er)
13.	RIKKAIM (richest)	TUHAT	ÕIGE (richtige)
14.	KESKEALINE (middle-aged)	PALJUD	TUNDMATU (unbekannter, -en)
15.	KOLMAS (third)	MÕLEMAD	VABA (freier)
16.	HABEMEGA (bearded)	IGA	RIKKAIM (reichste, -en)
17.	IGA (every)	VALGE	KOLMAS (dritte, -r)
18.	VÄRVATUD (enlisted)	TUGEV	HAIGE (kranker, -en)
19.	MÄNGU ALUSTAV (leadoff)	TEMA (TA)	RIKAS (reicher)
20.	VEIDER (odd)	ÕIGE	ÜKSKI (kein)
21.	ÕRNAHÄÄLNE (soft-spoken)	ESIMENE	NÕUTUD (gefragter)
22.	MASKEERITUD (masked)	VABA	VIGASTATUD (gebrochener)
23.	PEREKONNA (family)	TÖÖTU	MASKEERITUD (maskierter)
24.	AUS (honest)	TEINE	IGA (jener)
25.	NOOREM (younger)	SOBIV	RELVASTATUD (bewaffneter)
26.	PALESTIINA (Palestinian)	RIKAS	PÄRIT(stammende)
27.	PIME (blind)	MUST	KÕHETU (hagere)
28.	SUUR (great)	ETTE-VÕTLIK	MEIE (unser)
29.	SÜÜTU (innocent)	AUS	VÕIMAS (mächtiger)
30.	ABIELUS (married)	ANDEKAS	TEHTUD (gemahter)

Tabelis on kõrvutatud sõna *mees* kollokatsioone inglise, eesti ja saksa keeles. Tabelist on välja jäetud vanust märkivad kollokatsioonid, inglise keeles oli sellised 30 sagedamini esineva kollokatsiooni hulka pääsenud sõnadest veerand, näiteks *20-aastane, 22-aastane, 25-aastane, 19-aastane* jne.

Võrdluseks, saksa korpuses olid sagedamini mainitud vanused kõrgemad: kõige tihedamini oli mainitud 30-aastast meest, sellele järgnes 40-aastane. Sõna *naine* kollokatsioonide hulgas võis samuti märgata, et inglise keeles oli rohkem mainitud nooremat vanust (*19-aastane, 18-aastane*) ja saksa keeles oli samal ajal sagedamini mainitud 35-aastast ja 25-aastast naist. Edasised uuringud võiksid näidata, kas keelekorpuses ilmnev erinevus näitab tegelikku eri vanusegruppide väärtustamist eri kultuurides või mitte.

Tabelist on näha sõnakobarate ühisosa ja erinevused. Sõnade *man*, *mees* ja *Mann* kõigist kollokatsioonidest on umbes pooled ka-

hes või kolmes keeles ühised. Kolm kollokatsiooni *noor, iga* ja *teine* on kõigil kolmel sõnal ühised. Võrreldes sõnade *mees* ja *naine* kollokatsioone kõigis kolmes keeles, paistab silma, et sõna *noor* on ülekaalukalt kõige sagedasem kollokatsioon. Inglise ja saksa keele korpustes on see omadussõna sagedustabelis esikohal nii *mehe* kui *naise* (vt tabel 1 ja 2) täiendina. Eesti keele kõige sagedamini esinevate kollokatsioonide hulka kuuluvatele sõnadele *see* ja *üks* nii saksa kui ingliskeelsete vastete leidmist segas tõik, et neis keeltes täidavad sama funktsiooni üldjuhul artiklid, mis kollokatsioonide seast olid kõrvaldatud.

Sõna *mees* eesti kollokatsioonidel on suurim ühisosa saksa kollokatsioonidega: peale mainitud kolme kõigile ühise kollokatsiooni on neil veel 8 ühist sõna: *see, meie, minu, tugev, tema, õige, vaba, rikas.* Inglise kollokatsioonidega on ühisosa väiksem, vaid sõnad *üks*, valge, *must* ja *aus*. Seega võib väita, et eesti sõnale *mees* on saksa *Mann* tähenduslikult lähemal kui inglise vaste. Seda kinnitab ka sõnaraamatutähenduste võrdlemine: kui inglise *man* on kasutusel vaid erandjuhtudel ja teatud ühendites abielumehe tähistamiseks (vt *Colliens concise dictionary*), siis saksa ja eesti vasted on abikaasa tähistajatena tavalised (vt EKSS III, 1992: 386). Kollokatsioonid *minu* ja *tema* ongi mõeldavad sõna *mees* laienditena vaid neil juhtudel, kus see esineb abikaasa tähenduses.

Sõnade *man* ja *Mann* ühised kollokatsioonid on *vana, vanem, tundmatu, rikkaim, kolmas, maskeeritud.* Märkimisväärne on sõnade sama esinemisjärjekord sagedustabelis.

Siiski on kõigil kollokatsioonikobaratel ka suur eriosa. Mingil määral tuleneb see erinevast kultuurikontekstist. Ingliskeelsete kollokatsioonide seas on väga suure esinemissagedusega sõnad *must* ja *valge*, kuna ameerika kultuurikontekstis on vastandus *must mees – valge mees* väga olulisel kohal. Eesti ja saksa kultuurikontekstis on sellisel vastandusel väiksem roll. Kui eesti kollokatsioonide seas esinevad mainitud sõnad tabeli teises pooles, siis saksa korpust analüüsides pole *must* ja *valge* jõudnud üldse sõna *mees* 30 sagedamini esineva kollokatsiooni hulka. Viimane ei tähenda muidugi, et saksa keeles selliseid sõnaühendeid ei kohtaks: ka saksa korpuses esines sõnaühend *valge mees* ligi 60 korda. Veel üks väike erinevus: eesti ja saksa keeles on rohkem kasutatud täiendit *valge*, aga inglise keeles on nii *mehe* kui *naise* täiendina levinum sõna *must.*

5. Sõnade *woman*, *Frau* ja *naine* kollokatsioonide võrdlus

Sõna *woman* ja *Frau* kõigist kollokatsioonidest on pooled, ning sõna *naine* kollokatsioonidest on kolmandik kahes või kolmes keeles sarnased (tabel 2). Võrreldes sõna *mees* kollokatsioone sõna *naine* kollokatsioonidega kõigis kolmes keeles, hakkab silma, et sõna *naine* kollokatsioonide seas on poole rohkem kõigile kolmele kollokatsioonikobarale ühised sõnu.

Tabel 2. Sõnade *woman, naine* ja *Frau* 30 sagedasemat kollokatsiooni esinemissageduse järjekorras

	Inglise WOMAN	Eesti NAINE	Saksa FRAU
1.	NOOR (young)	EESTI	NOOR (junge)
2.	ESIMENE (first)	NOOR	MINU (meine/r)
3.	VANEM (elderly)	ÜKS	VANA (alte)
4.	RASE (pregnant)	VANA	TEMA (seine)
5.	ÜKS (one)	OMA	ÜKS (eine)
6.	MUST (black)	TEINE	ESIMENE (erste)
7.	TEINE (another)	MEHE	RASE (schwangere)
8.	KESKEALINE (middle-aged)	TARTU	VANEM (ältere, -en)
9.	VALGE (white)	TUHAT	LAHUTATUD (geschienene)
10.	ILUS (beautiful)	KAKS	AINUKE (einzige)
11.	TUNDMATU (unidentified)	MU	ILUS (schöne)
12.	VANA (old)	NEED	HABRAS (zierliche)
13.	AMEERIKA (American)	SOOME	TEINE (zweite)
14.	VANEM (older)	ÜHEKSA	BLOND (blonde)
15.	PALESTIINA (Palestinian)	KÕIK	ABIELUS (verheiratete, -en)
16.	ALASTI (nude)	LASTEGA	SÜNNIPÄEV(geburtstag)
17.	KODUTU (homeless)	MAAILMA	IGA (jede)
18.	MARYLANDI (Maryland)	MÕLEMAD	ARMASTATUD (geliebte, -en)
19.	ALASTI (naked)	NELI	ALASTI (nackte, -en)
20.	KORISTUS (cleaning)	PALJAS	NOOREM (jüngere)
21.	ABIELUS (married)	PRESIDENDI	ARMULINE (gnädige)
22.	NOOREM (younger)	RIIETATUD	RASE (hochschwangere)
23.	TÜRGI (Turkish)	ROHKEM	TÖÖTAV (berufstätige)
24.	AINUKE (only)	TEMA	KOLMAS (dritte)
25.	PÕLLUHARIJA (peasant)	VÄGISTATUD	RESOLUUTNE (resolute)
26.	SURNUD (dead)	VENE	ÜKSIK (alleinstehende)

	Inglise WOMAN	**Eesti NAINE**	**Saksa FRAU**
27.	LIIBANONI (Lebanese)	VANEM	AUSTATUD (geehrte)
28.	TEADVUSETU (comatose)	IGA	ELUS/ELAV (lebende)
29.	KARJÄÄRI (career)	KOHALIK	HAIGE (kranke)
30.	SALAPÄRANE (mystery)	TÖÖTAV	VEETLEV (attraktive)

Need kõigis kolmes keeles sagedamini korduvad kollokatsioonid on *noor, vana, vanem, üks, paljas* ja *teine*. Inglise ja saksa keeles laiendavad naist lisaks veel *rase, ilus, abielus, esimene, noorem* ja *ainuke*. Neist vähemalt kolme esimese sõna esinemine oleks väga ootuspärane ka eesti ajakirjandustekstides, kuna need märgivad niivõrd olulisi aspekte, mida naiste puhul on tähtsaks peetud: atraktiivsust, viljakust ja perekonnaseisu.

Tähelepanuvääriv on aspekt, et sõna *naine* ja *woman* kollokatsioonide seas pole ühtegi sõna, mis oleks ühised vaid neile kahele ja puuduksid sõna *Frau* kollokatsioonide hulgast. Samas ainult saksa ja eesti kollokatsioonikobaratele on ühised *minu, tema, iga* ja *töötav*. Viimasele vastab inglisekeelsest korpusest kõige paremini *career woman* 'karjäärinaine'. Kollokatsioonid *minu* ja *tema* viitavad sõna *naine* tähendusnüansile, mis saksa ja eesti keeles esineb, aga inglise keeles puudub. Nii saksa *Frau* kui eesti *naine* tähistavad ka abikaasat, sõnal *woman* see tähendus aga puudub. Eestikeelne sõna *naine* omab inglise keeles vasteid *woman* ja *wife*, saksakeelne vaste *Frau* on lisaks kasutusel ka kõnetlussõnana. Sõna *Frau* kasutus tähenduses *proua* avaldub saksa kollokatsioonides *armuline, austatud* ja *armastatud*. Sõnast paremale jäävate kollokatsioonide seas on olulisel kohal mitmed isikunimed. Seega võib öelda, et *Frau* on kõige laiema tähendusega, tähistades nii naissoost inimest, abikaasat kui prouat, ja *woman* on tähenduselt kõige kitsam, tähistades vaid naissoost inimest.

Kohanimede *Tartu* ja *Soome* esinemist sõna *naine* kollokatsioonidena võib pidada juhuslikuks. Siiski on tähelepanuväärne, et ka inglise keeles on sõna *naine* 30 sagedasema kollokatsiooni hulgas palju kohatäiendeid: *Marylandi, Palestiina, Türgi* ja *Liibanoni*. Kui konkreetsed kohanimed *Tartu* ja *Soome* võivad tõesti korpuse väikesest mahust tulenevalt olla sõna *naine* kollokatsioonidena juhuslikud, siis ei saa eitada, et ka suuremast korpusest otsinguid tehes sa-

tuks nii mõnigi kohanimi sõnade *mees* ja *naine* 30 sagedasema kollokatsiooni hulka.

6. Kokkuvõtteks

Sõnade *mees* ja *naine* eesti, inglise ja saksa keele kollokatsioonide analüüsist võib teha ettevaatliku järelduse, et sõna konnokatsioonid peegelduvad kollokatsioonides: kui kahes keeles on sõnal sarnane tähendusnüanss, mis kolmandas keeles puudub, siis väljendub see kollokatsioonides, mis sõnal on esimestes keeltes ühised, aga kolmandas keeles puuduvad. Lisaks tuli võrdluskatsel välja, et sõnadel *mees* ja *naine* on rohkem ühiseid kollokatsioone saksa keele sõnadega *Mann* ja *Frau* kui inglise *man* ja *woman* kollokatsioonidega.

Kirjeldatud võrdluskatse eesmärk ei olnudki põhjalike järelduste tegemine. Pigem huvitas mind, kas niivõrd erineva mahuga korpustest saadud andmed on võrreldavad, ja kuidas kollokatsioonide käsitsi kontroll mõjub võrdlusele. Selgus, et väikse korpuse kasutamisel täitis käsitsi kontroll sama rolli, mis automaatne sageduse arvestamine suuremate korpuste puhul.

Praegu on eesti keele sõnade kollokatsioonide uurimiseks paremad võimalused. Korpused on suurenenud ja kõne alla tuleks ka eesti keele morfoloogilise analüsaatori kasutamine, mis aitaks sõnavormide lemmatiseerimisel. Keeletarkvara abi ja keerulisemat statistilist töötlust on juba kasutatud kollokatiivsete püsiühendite leidmisel (Kaalep, Muischnek 2003). Lisaks on nüüd Leipzigi ülikooli otsingumootoriga võimalik saksa, inglise, prantsuse, hollandi, islandi ja sorbi keele kõrval automaatselt leida kollokatsioone ka eesti keelest.

Kirjandus

Collins Concise Dictionary 1999. Ed. J. M. Sinclair. Glasgow: Harper Collins Publishers.

EKSS = Eesti kirjakeele seletussõnaraamat 1992. III kd. Tallinn: Eesti Keele Instituut.

Johnson, Keith, Johnson, Helen 1999. Encyclopedic Dictionary of Applied Linguistics. A Handbook for Language Teaching. London: Blackwell.

Kaalep, Heiki-Jaan, Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli kirjastus.

Kaalep, Heiki-Jaan, Muischnek, Kadri 2003. Püsiühendite leidmine suurtest tekstikorpustest. – Toimiv keel I. Töid rakenduslingvistika alalt. (Eesti keele instituudi toimetised 12.) Toim M. Langemets, H. Sahkai, M.-M. Sepper. Tallinn: Eesti Keele Sihtasutus, 101–118.

Leipzigi ülikooli kollokatsioonide otsingumootor [WWW] http://wortschatz.uni-leipzig.de (2003)

Mills, Sara 2002. Post-feminist text analysis. – Critical Discourse Analysis. Critical Concepts in Llinguistics. Vol I. Ed by M. Toolan. London, New York: Routledge, 202–203.

Piits, Liisi 2004. Sõnade *mees*, *naine* ja *inimene* esinemisümbrusest ja stiilivärvingust. Magistritöö Tallinna Ülikooli eesti keele õppetoolis.

Singleton, David 2000. Language and the Lexicon. An Introduction. London: Arnold.

Stubbs, Michael 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

TÜKK = Tartu Ülikooli Eesti kirjakeele korpus www.cl.ut.ee/ee/corpusb/ (2002)

Eestikeelsete tekstide sisukokkuvõtjast EstSum

Kaili Müürisep
Tartu Ülikool

1. Sissejuhatus

Tekstidest automaatne sisukokkuvõtete tegemine on protsess, mille käigus luuakse tekstist olemasoleva põhjal uus, lühendatud versioon, mis sisaldab ainult kasutajale vajalikku informatsiooni.

Lühikesed ja informatiivsed ülevaated dokumentide, ajalehe- ja teaduslike artiklite jms sisust osutuvad tänapäeva tohutu informatsioonitulva juures asendamatuteks. Neid on näiteks tarvis nii mobiiltelefonide või pihuarvutite puhul, mille ekraanidelt on ebamugav liiga pikka teksti lugeda; Interneti otsingumootorites, et väljastatavate päringute hulgas oleks lihtsam orienteeruda; kiire ülevaate saamiseks juhul, kui lugejal on vaja läbi töötada suurem hulk tekste; kui ka hoopis teksti sünteeskõnena ettelugemisel arvuti poolt, kuna lühem tekst on kuulajale arusaadavam ja mugavam. Omaette uurimisvaldond on suulise kõne lühendamine, mille käigus eemaldatakse pausid jutus, kordused ja eneseparandused.

Arvutiprogrammi abil kokkuvõtete tegemisel kerkivad esile järgnevad probleemid: mida üldse lugeda tähtsaks informatsiooniks? Kui palju informatsiooni on vaja eraldada? Millised on need omadused või kriteeriumid, mille põhjal kõige tähtsamaid lauseid välja valida? Kuidas vältida seosetute lausete kaasamist? Kuidas vältida tarbetuid kordusi?

Töö eesti keele sisukokkuvõtjaga on käinud juba mitu aastat, kuid enamasti on see piirdunud diplomi- ja bakalaureusetöö tasemel eksperimentidega (Lippur 2000; Mutso 2005; vt ka Müürisep, Mutso 2005). EstSum on esialgu orienteeritud ainult veebiuudiste ja elektrooniliste ajaleheartiklite sisukokkuvõtmisele.

2. Mis on sisukokkuvõte

Radev jt annavad sisukokkuvõtte mitterange definitsiooni: "Sisukokkuvõte on tekst, mis on saadud ühe või rohkema teksti töötlemisel; mis annab edasi originaalteksti(de) olulist informatsiooni ja mis pole

pikem kui pool originaaltekstist, enamasti oluliselt lühem." (Radev jt 2002: 399). Teksti mõistet on siin kasutatud üsna vabalt: selle all mõeldakse nii tavalist teksti kui ka kõnet, multimeediadokumente, hüperteksti jne.

Sisukokkuvõtte põhieesmärk on esitada teksti peamised ideed väiksemas mahus.

Sisukokkuvõtteid on väga mitmeid tüüpe. Neid eristatakse selle järgi, kas saadud sisukokkuvõtte laused on originaaltekstist välja valitud (väljavõte, ingl k *extract*) või on nad genereeritud (ülevaade, ingl k *abstract*). Väljavõtte laused on täielikult kopeeritud originaaltekstist ning sisu antakse edasi autori täpses sõnastuses. Ülevaate puhul on sisukokkuvõttes lauseid, mida originaaltekstis ei esine. Tavaliselt on ülevaadetes kasutatud parafraseerimist (asesõnade asendamist, osalausete eemaldamist jms). Üldiselt võimaldavad ülevaated sisu tihedamini kokku pakkida kui väljavõtted, samas on neid automaatselt keerulisem genereerida.

Teine sisukokkuvõtete liigitus jagab sisukokkuvõtted indikatiivseteks ja informatiivseteks. Indikatiivne sisukokkuvõte peab andma arusaama, millest on dokumendis juttu, ilma detailidesse laskumata. Informatiivsed sisukokkuvõtted peavad edastama lühidalt kogu olulise informatsiooni.

Sisukokkuvõtteid saab liigitada ka teema põhjal: eristatakse üldisi ja teemale orienteeritud (*topic-oriented*, ka *user-focused*) sisukokkuvõtteid, kus viimased annavad informatsiooni lugeja huvidele vastavatel teemadel. See, millisel kujul lugeja huvi on ilmutatud, sõltub süsteemist: arvesse võidakse võtta kasutajaprofiili ja seniseid harjumusi, kasutaja koostatud päringut või ka loomulikus keeles esitatud küsimust.

3. Sisukokkuvõtmise meetodid

Klassikaline sisukokkuvõtmise protsess on kolmeetapiline (Mani 2001: 13).

1. Analüüs – sisendi analüüsimine ja selle sisekujule viimine.
2. Transformatsioon – toimub teisendus sisendi sisekujult kokkuvõtte sisekujule.
3. Süntees – sisukokkuvõtte sisekuju muudetakse loomuliku keele tekstiks.

Tegelikes süsteemides erinevate etappide piirid nii selged pole ning sageli nimetatakse neid etappe teiste nimedega.

Enamik tänapäeva sisukokkuvõttesüsteeme kasutab väljavõtte tüüpi sisukokkuvõtete tegemise metoodikat, st originaaltekstist valitakse välja laused, mis kannavad süsteemi "meelest" olulist informatsiooni, ja esitatakse need kasutajale.

Selliste süsteemide loomise ajalugu ulatub 1950. aastatesse. Levinuim tehnika seisneb selles, et iga lause jaoks arvutatakse selle lause skoor ehk kaal, mis põhineb lause asukohal tekstis, sõnade ja fraaside sagedustel, võtmefraaside esinemisel jne. Suurima skooriga laused kaasatakse sisukokkuvõttesse. Uuemad meetodid kasutavad tunnuste leidmiseks masinõppimise meetodeid või oluliste tekstipassaažide määramiseks süntaktilist analüüsi, samuti uuritakse pigem sõnadevahelisi seoseid kui sõnade hulki.

Lausetele kaalu arvutamise meetod põhineb Edmundsoni klassikaks saanud paradigmal (Edmundson 1969): lausete väljavalimisel kasutatakse nende hindamiseks valemit (1), kus W(s) on lause s kaal, C(s) on selle lause märgusõnade (*cue words*) skoor, K(s) võtmesõnade (*key words*) skoor, L(s) asukoha (*location*) skoor ja T(s) pealkirja sõnade (*title words*) skoor ning α, β, γ ja δ on konstandid:

(1)
$$W(s) = \alpha\, C(s) + \beta\, K(s) + \gamma\, L(s) + \delta\, T(s)$$

Märgusõnade skoor. Märgusõnadeks või -fraasideks loetakse sõnu, mis viitavad, et autor ise on selles lauses sisu kokku võtnud, nt "kokkuvõtteks", "järelikult", "selles artiklis", aga ka kesk- ja ülivõrdes omadussõnad "parim", "edukam", kõikvõimalikud hinnangut väljendavad sõnad ("õnnestus", "edukas") jpt. Kui sellised sõnad või fraasid esinevad lauses, siis saab see lause lisapunkte. Kui aga lauses leidub sõnu, mis võivad välistada lause sobimise sisukokkuvõttesse, nt "juhuslik", "vaevalt", siis lause kaalu vähendatakse. Mitmed hilisemad uuringud on näidanud, et märgusõnade ja -fraaside kasutamine on eriti õigustatud teaduslike tekstide sisukokkuvõtmisel. Ainult seda meetodit kasutades on võimalik leida 55% olulistest lausetest tekstis (Teufel, Moens 1997).

Võtmesõnade skoor. Olulised laused sisaldavad sõnu, mis esinevad tekstis mõnevõrra sagedamini. Muidugi tuleb seejuures arvestada sõnade üldist sagedust. Samas eksperimendid näitavad, et võt-

mesõnade skoori arvestamine ei pruugi sisukokkuvõtja tulemust parandada (Marcu 2003).

Asukoha skoor. Paljud sisukokkuvõtjad eeldavad, et laused, mis paiknevad teksti algul, on olulisemad kui tagapool paiknevad. Eksperimendid näitavad, et selle eeldusega töötavad sisukokkuvõtjad on parimad ühe tunnuse põhjal lauseid valivatest kokkuvõtjatest (Marcu 2003). Kuigi see meetod sõltub paljuski teksti žanrist ja sisukokkuvõtte pikkusest, on see üks tõhusamaid meetodeid 33-protsendilise pakkimise korral.

Pealkirjas esinevate sõnade skoor. Sõnad, mis esinevad teksti pealkirjas, on ilmselt temaatilised ning neid sõnu sisaldavad laused peaksid olema olulised.

Edmundson määras nende parameetrite väärtused käsitsi, samuti konstantide α, β, γ, ja δ väärtused. Tema eksperimendid näitasid, et märgusõnade, asukoha- ja pealkirja sõnade skooride kasutamine andis parima tulemuse, eraldi katsetatuna oli edukaim asukohapõhine meetod ja nõrgim võtmesõnade meetod. Tänapäeval on välja töötatud mitmed meetodid parameetrite väärtuste määramiseks masinõppimise meetodil.

Uuemad meetodid võtavad arvesse lausete süntaktilist ja semantilist infot.

Kohesioonipõhised meetodid eeldavad, et olulised on need laused, millel on kõige rohkem seoseid teiste lausetega. Üheks kohesioonipõhiseks meetodiks on näiteks leksikaalsete ahelate (*lexical chain*) meetod (Barzilay, Elhadad 1997). Leksikaalsesse ahelasse kuuluvad omavahel semantiliselt tugevalt seotud sõnad (nt kapsas, porgand, juurvili jmt). Ahelad genereeritakse automaatselt tesauruste või Wordneti abil. Tähtsateks lauseteks loetakse need, mille mõni sõna kuulub tugevasse ahelasse. Ahela tugevus arvutatakse tema pikkuse ja liikmete sageduse põhjal.

Teine näide kohesioonipõhisest meetodist on seotuse (*connectedness*) meetod (Mani 2001: 95–102): sõnadest moodustatakse graafi tipud. Tippude vahele luuakse kaared vastavalt sellele, kas antud sõnade vahel on grammatilisi, koreferentsi- või leksikaalse sarnasuse seoseid. Laused, mis sisaldavad enim seotud sõnu, loetakse olulisteks.

Teksti diskursuse struktuuri põhised meetodid. Teksti hierarhilist diskursuse struktuuri saab kasutada oluliste lausete leidmi-

seks. Kui lõik *p* arendab edasi lauset *l*, siis on mõistlik eeldada, et sisukokkuvõte peaks sisaldama ainult lauset *l*. See meetod on osutunud eriti sobivaks väga lühikeste sisukokkuvõtete tegemisel (Marcu 2000).

4. Sisukokkuvõtja EstSum kirjeldus

Eesti keele sisukokkuvõtja EstSum kasutab lausete väljavalimismeetodit ehk siis genereerib väljavõtte. Hetkel on EstSum orienteeritud veebis avaldatud uudiste ja ajaleheartiklite indikatiivsetele sisukokkuvõtetele.

EstSumi kavandamisel oli eeskujuks rootsi keele sisukokkuvõtja SWESUM[1] (Dalianis jt 2003).

EstSum koosneb kolmest moodulist: HTML-konverter, lausestaja ja väljavõtete tegija.

HTML-konverter eemaldab sisukokkuvõtte jaoks ebaolulised HTML-märgendid, normaliseerib ristuvad märgendid, eemaldab tabelid ja konverteerib sisendi SGML-formaati. SGML-formaadis märgendatakse pealkirjad ja alapealkirjad, autorid, pildiallkirjad. Samuti märgendatakse oluline laadiinformatsioon, eristades rasvast, kald- ja tavalist kirja.

Lausestaja kasutab reeglipõhist meetodit sisendi töötlemisel, lause alguse ja lõpu märgendamiseks kasutatakse 30 Perli regulaaravaldist.

EstSum arvutab sisukokkuvõtte pikkust kahel viisil. Esimene moodus on tavaline lausepõhine meetod, mille korral sisukokkuvõte 30% tähendab, et tekstis on 30% esialgsetest lausetest. Samas on pikemad laused informatsioonirikkamad ning tegelikult ei pruugi teksti pikkus nii palju lüheneda. Teine võimalus on arvutada sisukokkuvõtte pikkus sõnades. Sel viisil saadud sisukokkuvõtted on tõepoolest 30% esialgse teksti pikkusest.

EstSum kasutab oluliste lausete väljavalimiseks informatsiooni lausete asukoha, formaadi ja sõnavara kohta. Lausetele skoori arvutamiseks kasutatakse Edmundsoni valemile (1) sarnast valemit (2):

(2)
$$W(s)=\alpha P(s)+\beta F(s)+\gamma K(s)$$

[1] http://swesum.nada.kth.se/index-eng.html

W(s) on lause s kaal, P(s) on positsioonipõhine skoorifunktsioon, F(s) formaadipõhine skoorifunktsioon ja K(s) on sõnasagedustepõhine skoorifunktsioon; α, β ja γ on konstandid.

EstSumis puudub märgusõnade arvestamise võimalus, sest see nõuaks eestikeelsete ajaleheartiklite sõnastuse põhjalikumat analüüsi, samas kasutatakse ära formaadiinformatsiooni, erinevalt algsest valemist.

Tunnuste kaalud ja konstandid α, β ja γ on määratud käsitsi, kasutades selleks väikest treeningkorpust (20 teksti). Ehkki kasutatud korpus on suhteliselt väike, võimaldas see siiski määrata parameetrite esialgsed väärtused. Täpsem korpuse ja algoritmi kirjeldus on toodud (Müürisep, Mutso 2005).

Lausete asukohta uurides selgus, et olulised laused paiknevad pealkirja järel. Esimene lause oli sisukokkuvõttes 100% juhtudest, teine ja kolmas 65% juhtudest. Suurendati ka nende lausete skoore, mis järgnesid alapealkirjale, samuti lõigu esimese, teise ja kolmanda lause ning artikli viimase lause skoori.

EstSum loeb tähtsateks neid lauseid, milles kasutatakse rasvast või kaldkirja. Formaadipõhine skoorifunktsioon arvestab ka lause kirjavahemärkidega: hüüu- ja küsimärgid vähendavad lause kaalu, samuti jutumärgid. Täielikult välistatakse pildiallkirjade lisamine sisukokkuvõttesse.

Praegune EstSumi versioon ei kaasa lingvistilist moodulit, seepärast kasutatakse võtmesõnade statistika tegemiseks sõnavorme, mitte sõnade põhivorme.

Võtmesõnade tuvastamiseks kasutatakse kahte meetodit: 1) leitakse sõnad, mis on artiklis väga sagedased, kuid mitte nii sagedased üldises sagedustabelis; 2) pealkirjas ja alapealkirjades leiduvad sõnad loetakse olulisteks.

Samas, treeningkorpuse lähemal uurimisel ilmnes, et ainult 48% lausetest, mis sisaldasid pealkirjas leiduvaid sõnu, esinesid ka sisukokkuvõttes. Tekstis sagedasti esinevate sõnadega laused olid sisukokkuvõttes ainult 25% juhtudest.

Et leida konstantide α, β ja γ väärtusi, testisime, millise tulemuse annaks iga skoorifunktsioon eraldi. Ilmnes, et olulisim on positsiooni arvestav skoorifunktsioon. Formaati arvestav skoor erineb ainult üksikutel lausetel ning see iseseisvana lausete valikuks ei sobi.

Võtmesõnu arvestav skoorifunktsioon oli ebatäpsem ning seetõttu määrati konstantide α, β ja γ väärtusteks vastavalt 0,4, 0,4 ja 0,2.

Tallinn sai lapsesõbraliku linna tiitli kolme aasta pikkuse katseajaga
23.11.2005 00:01
Triin Olvet, reporter/toimetaja

Lapsesõbralikuks linnaks tunnistatud pealinn peab kolme aastaga näitama, et tahab ja suudab tegutseda ka noorimate linnakodanike huvides.

UNICEFi jagatava lapsesõbraliku linna tiitli pälvimiseks peab linn olema turvaline ning arvestama igakülgselt laste õigusi ja vajadusi.

Samuti peab UNICEF oluliseks lastele mitmekülgse huvitegevuse pakkumist ning nende kaasamist otsuste tegemisse. Lisaks Tallinnale said sel aastal austava nimetuse ka Jõhvi, Põlva ja Viljandi.

UNICEFi esindaja Eestis Toomas Palu ütles, et tiitli andmine ei tähenda veel, et linn ongi lapsesõbralik ja valmis. Mõte on pigem selles, et lapsed ja noored märkaksid, mis on nende kodulinnas hästi ja mida võiks linnavõimud paremini teha. "Laste teema peaks olema mitte ainult sotsiaal- ja haridusameti, vaid kogu linnavalitsuse ühine asi," rääkis Palu.

Katseaeg kolm aastat

Palu sõnul jälgib UNICEFi ja linna ühiskomisjon kolme aasta jooksul, et tahtmine lastele tähelepanu pöörata kuskile ei kaoks. Vastasel korral on õigus linnalt tiitel ära võtta. Palu rõhutas, et tiitel on nagu avanss, ja kolme aasta jooksul peab olukord paranema. "See ei ole finiš, vaid start," märkis Palu. "Kui tuleb valida näiteks Vabaduse väljaku asfalteerimise ja koolile uue katuse panemise vahel, siis peaks Tallinna eelistus nüüd selge olema."

Kuigi lapsesõbralikkust määravate tingimuste hulka kuulub ka puuetega laste toetamine, ei ole Tallinna puuetega inimeste koda praeguse olukorraga rahul. Koja esindaja Külli Urb oli Tallinnale antud tiitlist üllatunud, sest nendega polnud keegi sel teemal ühendust võtnud.

Urb nentis, et kuigi linna üldilme on liikumispuudega laste jaoks pisut paranenud, pääseb terves pealinnas ratastooliga sisse vaid Nõmme gümnaasiumisse ja Inglise Kolledžisse. Ülejäänud koolid on ilma isiklike abistajateta ratastoolilastele kättesaamatud.

Urbi sõnul on tore, et madalapõhjalisi busse ja mitmeid teenuseid on juurde tulnud,

Joonis 1. Katke EstSumi sisendtekstist
http://www.postimees.ee/231105/esileht/siseuudised/183960_print.php

Tallinn sai lapsesõbraliku linna tiitli kolme aasta pikkuse katseajaga

Lapsesõbralikuks linnaks tunnistatud pealinn peab kolme aastaga näitama, et tahab ja suudab tegutseda ka noorimate linnakodanike huvides.

UNICEFi jagatava lapsesõbraliku linna tiitli pälvimiseks peab linn olema turvaline ning arvestama igakülgselt laste õigusi ja vajadusi.

UNICEFi esindaja Eestis Toomas Palu ütles, et tiitli andmine ei tähenda veel, et linn ongi lapsesõbralik ja valmis.

Palu sõnul jälgib UNICEFi ja linna ühiskomisjon kolme aasta jooksul, et tahtmine lastele tähelepanu pöörata kuskile ei kaoks.

Urb nentis, et kuigi linna üldilme on liikumispuudega laste jaoks pisut paranenud, pääseb terves pealinnas ratastooliga sisse vaid Nõmme gümnaasiumisse ja Inglise Kolledžisse.

Urbi sõnul on tore, et madalapõhjalisi busse ja mitmeid teenuseid on juurde tulnud, aga üldise arenguga võrreldes on puudega lapsed unustusse jäetud.

Põhja prefektuuri liiklusjärelevalve osakonna konstaabel Liina Soodla oli Tallinna lapsesõbralikuks tunnistamise üle rõõmus.

Joonis 2. EstSumi poolt genereeritud sisukokkuvõte

Selliste väärtuste korral kattus 51% EstSumi poolt leitud sisukokkuvõtete lausetest inimese poolt valitutega.

Joonistel 1 ja 2 on toodud näide lähtetekstist ja 30-protsendilisest sisukokkuvõttest. Selle artikli puhul on näha, et EstSum eelistab lõikude esimesi lauseid. Formaadipõhine skoorifunktsioon aitab vältida jutumärkides tsitaatide kaasamist sisukokkuvõttesse. Ainult positsiooni ja võtmesõnade baasil otsustades oleks sisukokkuvõtte viimaseks lauseks "Kõik need mänguväljakud ja noortekeskused on väga vahvad."

EstSumi loodud sisukokkuvõte ei ole sidus. Selles artiklis ei kaasatud sisukokkuvõttesse asesõnadega lauseid, kuigi enamasti valitakse välja ka mõni lause, milles on ilma igasuguste selgitusteta asesõnad *ta* või *see*. Näiteväljavõttes kasutatakse isiku tähistamiseks perekonnanime Urb, lähtetekstis olid küll antud tema täisnimi ja tiitel, kuid see lause ei osutunud valituks.

Artikli viimane osa käsitles Tallinna liikluskultuuri, millega konstaabel Soodla ei olnud rahul, kuid sisukokkuvõttesse sattus tendentslikult ainult positiivne lause.

5. Tulemuste hindamine

Kuidas hinnata sisukokkuvõtte headust? Mis teeb ühe sisukokkuvõtte heaks ja teise halvaks? Tavaliselt toimub automaatselt genereeritud sisukokkuvõtte hindamine sel viisil, et seda võrreldakse inimes(t)e poolt koostatud sisukokkuvõttega (väljavõttega) ning leitakse kattuvate lausete osakaal. Hea, kui neid inimesi, kes sisukokkuvõtteid käsitsi koostavad, oleks mitu. Samas on sisukokkuvõtja töö hindamisel tähtis teada fakti, et kahe inimese poolt koostatud väljavõtetes kattub ainult 70% lausetest (Hassel 2003).

EstSumi hindamiseks loodud korpus koosnes 11 tekstist, milles oli keskmiselt 23 lauset. EstSumi poolt valitud laused kattusid 60% ulatuses inimese poolt valitud lausetega. Parimal juhul oli samu lauseid 85% ja halvimal juhul ei kattunud ükski (väga lühike artikkel).

6. Järeldused ja plaanid edaspidiseks

Eestikeelsete tekstide sisukokkuvõtja EstSum on veel eksperimentaalses arengujärgus ning selle edendamiseks on vaja teha palju tööd.

Olulisim oleks lingvistilise mooduli (morfoloogiaanalüsaator, morfoloogiline ühestaja ja süntaksianalüsaator) ühendamine EstSumiga. Eelkõige võimaldaks see paremat võtmesõnade statistikat, mis praegu on EstSumi nõrgim koht. Semantikal põhineva heuristika kasutuselevõtt vajaks Wordneti (või sellel baseeruva sõnastiku) ühendamist sisukokkuvõtjaga. Süntaktilise informatsiooni olemasolu lubaks lauseid automaatselt lühendada, eemaldades näiteks osalauseid, samuti oleks see eelduseks anafooride lahendamisele. Süntaktiliselt analüüsitud sisend ja lahendatud anafoorid võimaldaksid katsetada ka keerukamaid kohesioonipõhiseid sisukokkuvõttemeetodeid.

Samas on ilmselge, et 10–20 teksti käsitsi koostatud sisukokkuvõtte põhjal tehtud üldistused ei ole piisavalt täpsed. Vaja on suuremat test- ja treeningkorpust, mis võimaldaks parameetrite väärtused leida statistiliste masinõppimise meetoditega.

Kirjandus

Barzilay, Regina, Elhadad, Michael 1997. Using Lexical Chains for Text Summarization. – Proceedings of the Intelligent Scalable Text Summarization Workshop, ACL, Madrid.

Dalianis, Hercules, Hassel, Martin, Wedekind, Jürgen, Haltrup, Dorte, de Smedt, Koenraad, Lech, Till Christopher 2003. Automatic text summarization for the Scandinavian languages. – Nordisk Sprogteknologi 2002: Årbog for Nordisk Språkteknologisk Forskningsprogram 2000–2004. Ed by H. Holmboe. Museum Tusculanums Forlag, 153–163.

Edmundson, H. P. 1969. New methods in automatic abstracting. – Journal of the Association for Computing Machinery 16 (2). 264–285. Reprinted in: Advances in Automatic Text Summarization. Ed by I. Mani, M.T. Maybury. Cambridge, Massachusetts: MIT Press, 21–42.

Hassel, Martin 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. – Proceedings of NODALIDA '03 – 14th Nordic Conference on Computational Linguistic., Reykjavik, Iceland.

Lippur, Andres 2000. Automaatne sisukokkuvõtete tegemine eestikeelsetele tekstidele. Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut.

Mani, Inderjeet 2001. Automatic summarization. Amsterdam: John Benjamins.

Marcu, Daniel 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. – Computational Linguistics, 26 (3), 395–448.

Marcu, Daniel 2003. Automatic Abstracting. – Encyclopedia of Library and Information Science, 245–256.

Mutso, Pilleriin 2005. Automaatne sisukokkuvõtete tegemine. Diplomitöö. Tartu Ülikool, arvutiteaduse instituut.

Müürisep, Kaili, Mutso, Pilleriin 2005. ESTSUM – Estonian newspaper texts summarizer. – Proceedings of The Second Baltic Conference on Human Language Technologies. Tallinn, 311–316.

Radev, Dragomir R., Hovy, Eduard, McKeown, Kathleen 2002. Introduction to the special issue on summarization. – Computational Linguistics 28(4), 399–408.

Teufel, Simone, Moens, Marc 1997. Sentence Extraction as a Classification Task. – Proceedings of Intelligent and Scalable Text Summarization Workshop. Madrid, 58–65.

Lisa. Automaatselt genereeritud 10-protsendiline sisukokkuvõte artiklist

Eestikeelsete tekstide sisukokkuvõtjast EstSum

Kaili Müürisep

Automaatne sisukokkuvõtete tegemine tekstist on protsess, mille käigus luuakse tekstist olemasoleva põhjal uus, lühendatud versioon, mis sisaldab ainult kasutajale vajalikku informatsiooni.

Radev jt annavad sisukokkuvõtte mitterange definitsiooni: "Sisukokkuvõte on tekst, mis on saadud ühe või rohkema teksti töötlemisel; mis annab edasi originaalteksti(de) olulist informatsiooni ja mis pole pikem kui pool originaaltekstist, enamasti oluliselt lühem." (Radev jt 2002: 399).

Klassikaline sisukokkuvõtmise protsess on kolmeetapiline (Mani 2001: 13).

Märgusõnadeks või -fraasideks loetakse sõnu, mis viitavad, et autor ise on selles lauses sisu kokku võtnud, nt "kokkuvõtteks", "järelikult", "selles artiklis", aga ka kesk- ja ülivõrdes omadussõnad "parim", "edukam", kõikvõimalikud hinnangut väljendavad sõnad ("õnnestus", "edukas") jpt.

Olulised laused sisaldavad sõnu, mis esinevad tekstis mõnevõrra sagedamini.

Paljud sisukokkuvõtjad eeldavad, et laused, mis paiknevad teksti algul, on olulisemad kui tagapool paiknevad.

Sõnad, mis esinevad teksti pealkirjas, on ilmselt temaatilised ning neid sõnu sisaldavad laused peaksid olema olulised.

Eesti keele sisukokkuvõtja EstSum kasutab lausete väljavalimismeetodit ehk siis genereerib väljavõtte.

EstSum kasutab oluliste lausete väljavalimiseks informatsiooni lausete asukoha, formaadi ja sõnavara kohta.

W(s) on lause s kaal, P(s) on positsioonipõhine skoorifunktsioon, F(s) formaadipõhine skoorifunktsioon and K(s) on sõnasageduste põhine skoorifunktsioon; α, β ja γ on konstandid.

Tunnuste kaalud ja konstandid α, β ja γ on määratud käsitsi, kasutades selleks väikest treeningkorpust (20 teksti).

Joonistel 1 ja 2 on toodud näide lähtetekstist ja 30-protsendilisest sisukokkuvõttest.

Suuline keel, dialoog ja arvuti. Sissejuhatuseks

Tiit Hennoste
Helsingi Ülikool / Tartu Ülikool

Käesolev artikkel on sissejuhatuseks uurimustele suulise keele, dialoogi ja arvuti suhete kohta, millega tegelevad järgnevad artiklid.[1]

Esimeses osas annan lühikese ülevaate suulise kõne korpustest, teises osas suulise keele ja dialoogi uurimisest. Ühelt poolt on see üldine pilt, teiselt keskendun Tartu ülikooli suulise kõne uurimisrühma töödele. Sealjuures ei tegele ma teemade ja probleemide detailide kirjeldamisega, vaid viitan artiklitele ja monograafiatele, kus nendest juttu on.[2]

1. Korpused, programmid ja arvuti

Suuline kõne on hetkeline nähtus, mille uurimine saab toetuda ainult sellele, et ta on kogutud, fikseeritud ja korpuseks organiseeritud.

Varasemad suulise kõne korpused olid lihtsalt kastid kuulamise põhjal käsitsi üleskirjutatud sedelitega. Alates 1960. aastatest, eriti aga viimase paarikümne aasta jooksul on ka suulise kõne korpused arvutis olevad keelekogud (vt Leech jt 1995).

1.1. Suulised arvutikorpused

Korpus vajab loomiseks aluspõhimõtteid, millest lähtudes materjali koguda ja süstematiseerida. Selliseid aluseid on palju ja erinevad korpused on ehitatud vägagi erinevalt (vt nt Crowdy 1993). Järgnevas jagame korpused kolme mõõtme järgi.

---

[1] Tänan Mare Koitu abi ja täienduste eest.

[2] Seega ei anna käesolev artikkel ülevaadet sellest, mida on tehtud eesti suulise keele ja dialoogi uurimisel või modelleerimisel väljaspool seda uurimisrühma (Leelo Keevallik, Renate Pajusalu jt; vt ülevaatlikult Hennoste 2003a; Keevallik 2003). Samuti jäävad kõrvale kõnetuvastuse ja -sünteesi uurimused ja modelleerimised, millega tegelevad TTÜ küberneetika Instituudi ja EKI uurijad (Einar Meister, Meelis Mihkla, Arvo Eek, Tanel Alumäe; vt Alumäe 2005; Mihkla, Meister 2002; Meister 2001; Mihkla jt 2000).

Üks liigendus on kõnekorpused (*speech corpora*) ja keelekorpused (*linguistic corpora*). Kõnekorpused on rangete tehniliste kvaliteedinõuete alusel koostatud materjalikogud, millest uurijad kogu-vad materjali kõne foneetiliseks analüüsiks ja sünteesiks. Keelekorpused on kogutud tavaliselt lõdvemate kvaliteedinõuete alusel ja neid kasutatakse sõnavara, grammatika ja suhtluse uurimisel. Samas on kõnekorpuste teemade ja situatsioonide valik üsna piiratud, keelekorpuste nõuded aga lubavad teha lindistusi praktiliselt igal pool.[3]

Teine liigendus on vastavalt korpuse lindistusviisile. Varasemad korpused olid audiolindistused. Viimasel ajal kasutatakse võimalusel alati videolindistusi. Meie korpuse lindistuste põhiosa on audiolindistused.

Kolmas võimalus on jagada korpused kaheks vastavalt valitud tekstidele: universaalsed korpused ja spetsiifilised koprused.

Universaalsed korpused püüavad haarata ühe (all)keele maksimaalselt paljusid variante ja võimaldavad lahendada suhteliselt paljusid uurimisülesandeid. Sellised on nt Tartu Ülikooli Eesti suulise keele korpus või ülikooli ja EKI koostöös loodud eesti traditsioonilise murrete korpus (vt Hennoste 2000, 2003b; Lindström jt 2001; Lindström, Pajusalu 2003).

Spetsiifilised korpused on sellised, millesse kogutakse või valitakse tekste kindlate parameetrite alusel kindla uurimisülesande tarvis. Nii on nt tehtud Tartu Ülikooli Eesti dialoogikorpus EDiK, millesse on kogutud kindlat tüüpi institutsionaalseid dialooge selleks, et kasutada neid suulise infodialoogi analüüsiks ja modelleerimiseks (vt Hennoste jt 2002, 2004a; Koit 2003, Gerassimenko jt 2004a). Sellised on enamasti ka erinevate uurijate individuaalsed korpused (nt Riina Kasterpalu uurimuses kasutatav müügivestluste korpus jm).

1.2. Transkriptsioon ja transkribeerimine

Suuline kõne vajab ülesmärkimiseks transkriptsiooni, mis erineb tavalisest kirjakeele ortograafiast. Põhjusi on mitu. Näiteks esineb kõnes sõnu ja häälitsusi (*ee, mhmh* jms), kirjakeelest erinevalt hääldatud sõnu (*sis, vä*) jms, mille ülesmärkimiseks puudub ortograafia. Suulises kõnes pole tihti lauset kirjakeele mõttes, vaid pigem intonatsioonilised

[3] Kõnekorpuste kohta vt Eek, Meister 1999; Meister jt 2003.

üksused. Suhtlejad kasutavad mitteverbaalseid vahendeid, millel on suhtluses oma funktsioonid (pausid, intonatsioon jms).

Pole olemas universaalset ja kõiki keeleseiku fikseerivat transkriptsiooni. Esiteks on selle tegemine tohutult töömahukas ja selliselt väga kallis. Teiseks teeb kogu info korraga esitamine teksti inimesele loetamatuks. Asja teeb keerukamaks see, et eri teoreetilised kontseptsioonid ja uurimisülesanded vajavad vägagi erinevaid transkriptsioone. Seetõttu on maailmas kasutusel erinevaid transkriptsioone. Tuntumad on vast eeskätt murdetekstide litereerimisel kasutatav foneetiline transkriptsioon ning vestlusanalüüsi transkriptsioon, mida kasutatakse ka TÜ Eesti suulise keele korpuses (vt Pajusalu jt 2002: 298–301; Hennoste 2000).

Eri transkriptsioonid esitavad detailselt just neid nähtusi, millele vastava korpuse kasutajad eeldatavasti keskenduvad ja märgivad muud seigad pealiskaudselt.

Nii esitab foneetiline transkriptsioon vastavalt klassikalise murdeuurimise eesmärkidele väga detailselt sõnade hääldust, kuid väga pealiskaudselt ja puudulikult süntaksit. Vestlusanalüüsi transkriptsioon aga keskendub dialoogi loomisel oluliste üksuste väljatoomisele (pausid, pealerääkimised jms), kuid kirjutab sõnad tavalist ortograafiat kasutades, n-ö kuulmise järgi.

Vajadusel esitatakse sama transkriptsiooni sees lisainfot või transkribeeritakse samad tekstid mitmesse eri transkriptiooni. Nt USA suulise kõne korpuse transkriptsioonis antakse vajadusel sulgudes sõna eripärane hääldus foneetilises transkriptsioonis (vt Du Bois jt 1993). Eesti murdekorpuses on aga samad tekstid antud nii foneetilises kui lihtsustatud ehk süntaktilises transkriptsioonis (vt Lindström jt 2001).

Arvutikorpuste tulek muutis nõudeid transkriptsioonidele. Kui käsitsi tehtav ja kasutatav trankriptsioon loodi suhteliselt vabalt, siis arvuti jaoks on tarvis hoopis rangemaid reegleid erinevate üksuste piiride määramiseks. Ka ei saa arvuti puhul kasutada kõiki märke vabalt, sest see võib raskendada või takistada hiljem transkriptsiooni abil materjali sorteerimist. Lisaks kujundati välja eraldi transkriptsioonid, mis olid mõeldud nimelt ja üksnes arvuti jaoks, nt TEI (vt Leech jt 1995; Hennoste jt 2000).

Transkribeerimine on aastakümneid olnud inimese poolt kuulates ja käsitsi tehtav tegevus. Sealjuures on see osalt olnud asjaoludest

tingitud vajadus, osalt aga põhimõte. Nt vestlusanaüüsi klassikaliste seisukohtade järgi tuleb transkribeerida nimelt kuuldeliselt, et imiteerida inimest, kes kõnet vastu võtab. Ka tänapäeval ei ole olemas täisautomaatseid trankriptsiooniprogramme, mis teeksid töö ilma inimese abita valmis. Siiski on arvuti muutunud viimasel kümnekonnal aastal transkribeerija oluliseks abiliseks.

Kõigepealt on loodud tehnilisi abivahendeid. Selline on nt Voice Walker, pisike programm, mis lubab digitaalset materjali kergesti n-ö edasi-tagasi kerida, automaatselt mitu korda korrata jne (vt http://linguistics.ucsb.edu/resources/resources.htm). See on tegelikult klassikaliste transkribeerimismagnetofonide asendaja arvutis. Sellest keerukam ja enam võimalusi pakkuv on CLAN (vt http://childes.psy.cmu.edu).

Teiseks on väike osa transkribeerimisest endast antud tänapäeval ka arvutile. Nt pausid mõõdetakse tihti foneetikaprogrammide abil, millest tavalisim on Praat (http://www.fon.hum.uva.nl/praat/). Muide, kõrvaga hinnatud pauside pikkus erineb üsna palju arvuti abil arvutatust. See omakorda teeb eri meetoditel transkribeeritud tekstide kvantitatiivse võrdlemise raskeks.

Ja lõpuks on väga oluline see, et transkribeerimine ei ole mehhaaniline tegevus, vaid juba materjali interpretatsioon. Eri inimesed kuulevad nt eesti keele välteid või lausungi intonatsiooni pisut erinevalt, tõlgendavad erinevalt ja märgivad erinevalt üles. See johtub paljudest seikadest, mis ei ole kõrvaldatavad: lintide kvaliteet, uurijate erinevad teoreetilised arusaamad, mis ebateadlikult mõjutavad nende interpretatsiooni, lihtsalt erinev kuulmine jne.

1.3. Taustainfo

Tänapäeval on üldaktsepteeritav, et keele kasutamine oleneb palju erinevatest konsituatiivsetest seikadest. Argivestluste keel erineb nii kvantitatiivselt kui kvalitatiivselt ametidialoogide omast. Viimastes omakorda saame välja tuua erinevaid alarühmi, mille keelelised valikud on üsna erinevad. Seega on vaja tekstidele lisaks fikseerida taustateave situatsioonide kohta, kasutades neid parameetreid, mille alusel keelekasutus võib eeldatavasti varieerida. Vaid korpus, mis on mõeldud ainult keeles olevate universaalsete nähtuste kvalitatiivseks uurimiseks, ei vaja erilist taustateavet.

Taustateabe valimiseks ja esitamiseks pole mingit ühtset või laiemalt aktsepteeritud süsteemi (vt Hennoste 2000). Enamik korpusi on siiski lisanud infot tekstide kasutajate sotsiaalsete parameetrite ja suhtlussituatsioonide kohta.

Arvutite puhul võib nt kõnelda TEI korpusepäisest, milles tuuakse välja nii tekstide tehnilised kui sisulised erijooned teatud süsteemis. Seda on suulise korpuse tarvis arendatud ja kasutatud vähe (vt Johansson 1995). Tartu Ülikooli Eesti suulise keele korpuse kohta on loodud n-ö maksimaalne taustakirjeldus, milles on võetud arvesse maksimaalselt need seigad, mis on leitud mõjutavat keelejooni. Kuid reaalses tekstikogumises kasutatakse sellest lühikest varianti (vt http://www.cl.ut.ee/suuline/).

1.4. Korpuste kasutamine

Kõigepealt, suuliste korpuste puhul moodustab omaette probleemi nende avalikuks tegemine (nt arvutivõrku riputamine). See toob kaasa keerukaid eetilisi ja juriidilisi probleeme, millel siiani ühised lahendused eri maade jaoks puuduvad. Põhjuseks on asjaolu, et tihti on tegemist tekstidega, mis on eravestlused või dialoogid väga delikaatsetel teemadel (nt arsti-patsidendi vestlused). Teisalt tekib rida tehnilisi probleeme. Need seostuvad tekstide erinevate lindistusformaatidega, erinevustega transkriptsioonides jne. Lisaks tekib probleeme sellega, et suuline korpus on suure töö tulemus ja eriti üksikuurijad ei soovi loovutada oma tööd tasuta teistele kasutamiseks. Seetõttu võib võrgust leida enamasti vaid transkribeeritud tekstide üksiknäiteid, mille kasutamiseks tuleb üldjuhul ka korpuse loojatelt või omanikelt luba küsida.

Ka Tartu ülikooli korpuse kasutamiseks on vaja sõlmida korpuse administraatoriga eraldi leping (vt http://www.cl.ut.ee/suuline/). Eesti dialoogikorpus (transkribeeritud tekstid) on veebis ligipääsetav, kuid parooliga kaitstud.

Korpuste materjali kasutatakse väga erinevalt, kuid need viisid võib koondada kahte rühma.

Ühel juhul on korpus lihtsalt materjalipank. Sel juhul peab see võimaldama materjali eri viisil otsida, nt leida erinevate taustaomadustega tekste (dialoogid, mitte monoloogid, argivestlused, mitte ametidialoogid jne), otsida kindlaid keelelisi nähtusi (nt asesõnu, teatud tüüpi lauseid jne). Pisikesed (ja eriti uurijate personaalsed)

korpused võimaldavad sellist materjali ilma erilise vaevata leida. Suured korpused, mille tegemisel on osalenud kümneid inimesi, vajavad otsimisel tavaliselt arvutiprogrammide abi.

Lihtsaimad seda tüüpi programmid on kasvõi Wordist tuntud käsk Otsi. Samuti saab selleks kasutada nt UNIXi käsurida. Aga loomulikult on loodud palju erinevaid otsimisprogramme, mis on kindlate uurimisülesannete jaoks kohandatud detailsemad otsijad. Nii on Tartu ülikooli dialoogikorpuse tarvis loodud programm, mis võimaldab otsida kindlaid dialoogiakte või sõnu ja sõnajärjendeid kas kogu korpusest või teatud dialoogidest (Treumuth 2005, vt ka http://math.ut.ee/~treumuth/).

Teisel juhul analüüsitakse materjali arvutiprogrammide abiga. Siia kuuluvad nt programmid, mis lubavad teha erinevaid statistikaid ja mingeid keelelisi üksusi rühmadesse sorteerida.

Väikeste korpuste puhul piisab tihti inimese enda arvutusvõimest. Suurte korpuste käsitsi analüüsimine tähendab kõigepealt programmide loomist, mis analüüsivad transkribeeritud teksti.

Lihtsaimad seda tüüpi programmid on samuti Wordist tuntud käsud, nagu Sordi ja Sõnaarvestus. Samuti saab neid ülesandeid teha UNIXi käsurea või CLANi abil. Aga siia kuuluvad ka programmid, mis analüüsivad nt foneetikat või morfoloogiat.

Lihtsamad ülesanded on sellised, mille puhul sobivad universaalsed programmid, mis aitavad leida, süstematiseerida või analüüsida materjali, kasutades korpuse transkriptsioonis olevaid tunnuseid. Nt liigendavad korpuse lausungiteks, kasutades kirjavahemärke, loevad üle pausid, kasutades ära pauside transkriptsiooni või analüüsivad foneetilisi nähtusi nagu Praat.

Keerukamate ülesannete jaoks on aga vaja korpus eelnevalt ette valmistada ja/või luua sobivad analüüsiprogrammid.

Nt kui me tahame koostada eesti keele erinevate käänete ja sõnaliikide kasutuse sagedustabelit, siis on meil vaja programmi, mis nimelt eesti keele käänded ja sõnaliigid ära tunneb ja välja otsib, ehk eesti keele morfanalüsaatorit. See loodi esmalt kirjakeele tarvis. Kuna suuline keel erineb oma sõnaliikidelt ja käänete vormistuselt mõnes osas kirjakeelest, siis tuli seda programmi suulise kõne tarvis täiendada (vt Kaalep, Vaino 2000; Hennoste jt 2000).

Samasugune programm on ka pindsüntaksi analüsaator, mis praegu on suulise kõne analüüsimiseks alles algusjärgus (vt Müürisep jt käesolevas kogumikus).

Selliselt töödeldud ja otsitud materjali on võimalik kasutada keele ja dialoogi uurimisel, aga ka erinevateks praktilisteks lahendusteks, nt sõnastike tegemiseks, mis tänapäeval on eeskätt korpustest sõnade ja nende kasutust illustreerivate näidete valimine. Aga samuti on selline materjal aluseks uute arvutiprogrammide tegemisele.

2. Suulise keele ja dialoogi uurimused

2.1. Suulise keele uurimused

Suulise keele uurimise võib jagada laias laastus kolme rühma.

Esimene rühm on selline, mille puhul suuline kõne on eeskätt materjaliks, või üheks materjaliks muude allkeelte kõrval.

Siia võib arvata nt klassikalise murdeuurimise. Traditsioonilised murded on suulised allkeeled, kuid nende uurimine ei võtnud arvesse nt keelelisi erijooni, mis on tingitud suulisusest. Eriti hästi torkab see silma süntaksis, kus kõne ja kiri üksteisest tugevalt erinevad (vt Pajusalu jt 2002: 107–115).

Samasse jäävad ka kvantitatiivsed sotsiolingvistilised uurimused, mille puhul üheks uuritavaks objektiks on suulised argisituatsioonid ja neis kasutatav keel võrrelduna muudes situatsioonides kasutatavaga. Eesmärgiks pole siin aga uurida suulist kõnet, vaid leida neid inimrühmade sotsiaalseid tunnuseid, mille alusel keelekasutus varieerub.

Teine osa suulise kõne uurimusi tegeleb nimelt suulise kõne erijoonte analüüsiga.[4] Üks osa selle suuna uurimusi vaatleb suulise keele erijooni võrreldes kirjaga. Teine suund uurib erinevusi suulise kõne sees, nt argikeele ja ametikeele erinevusi. Meetodite poolest on tegemist väga kirju seltskonnaga.

Ühe rühma moodustavad deskriptiivsed suulise kõne kirjeldused, mis toovad välja selle, kuidas suuline kõne nt moodustab lauseid, kasutab erinevaid sõnu, eriti partikleid, moodustab käändeid

[4] Võibolla võib siia paigutada ka omaette seisva foneetika, mis uurib häält, selle tekitamist, erinevaid häälikuid ja nende moodustamist jms häälega seotud nähtusi.

jne. Kuna erinevused kirja ja kõne vahel võivad olla kvalitatiivsed ja kvantitatiivsed, siis kasutab uurimine ka mõlemaid meetodeid.

Teise rühma moodustavad nt psühholingvistilised uurimused, mis tänapäeval kasutavad üha enam materjalina reaalset suulist kõnet ja mille eesmärgiks on seletada suulise kõne eri fenomene psühholoogiliselt.

Kolmas rühm on võrdlev uurimine, mis tegeleb eri keelte võrdlusega, vaatleb nt seda, mismoodi kasutatakse samu partikleid eri keeltes või mille abil vastatakse eri keeltes samasugustele küsimustele või milliseid üneeme kasutavad eri keeled ja kuidas see on seotud vastava keele foneemikomplektiga. Samuti otsivad need uurimused suulise kõne universaale.

Omaette rühma annab võõrkeele (L2) õppimise ja kasutuse uurimine, mille eesmärgiks on eeskätt saada teavet selle kohta, kuidas kõneleja kasutab teist keelt, millised on emakeelse ja võõrkeelse suhtleja probleemid, kuidas neid lahendatakse jne.

2.2. Dialoogi uurimine

Omaette ala moodustab vestluse, dialoogi uurimine, mis pole mitte keele, vaid suhtluse põhimõtete uurimine. Seda suunda huvitab eeskätt see, kuidas kasutatakse erinevaid keelevahendeid suhtluses. See toetub osalt suulise keele uurimisel kasutatavatele tulemustele. Siin on erinevaid meetodeid, millest keskne on tänapäeval vestlusanalüüs (vt selle kohta nt Tainio toim 1997).

Sellest on osaliselt välja kasvanud interaktiivne keeleuurimine, mille aluseks on vaateviis, et grammatikat, ja eriti suulise keele grammatikat, tuleb uurida suhtlusest lähtudes, kuna see, kuidas me keelt kasutame, on tugevalt seotud sellega, mida me keele abil teha tahame (vt ülevaadet Keevallik 2002).

Selle sees annavad omaette rühma uurimused, mille eesmärgiks on suulise kõne ja suhtluse modelleerimine.

Nende lõppeesmärgiks on töötava ja suulises kõnes suhtleva programmi loomine. Kõige enam on tegemist küsimus–vastus süsteemidega, milles klient suhtleb arvutiga (nt rongiaegade ütleja, restoranide tutvustaja jne). Teine suur rühm on referaatide, lühikokkuvõtete tegijad. Aga siia kuuluvad ka pisemad kaubalised rakendused, nt lapse mängutelefon, mis ütleb klahvidele vajutamisel telefoninumbreid jne jne.

Seni on seda tüüpi programmid olnud suhteliselt lihtsad ja ühe ülesande keskses (nt vastates rongide väljumisaegu). Selliste programmide tegemine vajab kõigepealt vastava (suulise) keele analüüsi ja selle põhjal tehtud sünteesi, mille tulemuseks on arvutiprogrammid, mis suudavad mõista vastavas keeles teksti ja öelda ise võimalikult hästi vastava keele sõnu ja lauseid. Siin on keskseks olnud foneetilised ja intonatsioonilised probleemid, st see, et arvuti suudaks öelda sõnu ja ka lauseid võimalikult vastavat keelt emakeelena kõneleja sarnaselt. Sealjuures ei arvesta need programmid eriti tegeliku suulise suhtluse põhimõtetega vastavas keeles või kultuuris.

Viimasel ajal on suund keerukamatele programmidele. Need suudavad arvesse võtta suulise suhtluse üldisi erijooni (nt eristada välja inimese küsimuses olevad takerdumised ja saada neile vaatamata küsimusest aru). Ja suudavad ka suhelda võimalikult inimlähedaselt, st arvestada vastava kultuuri inimeste tüüpilisi suhtlusstrateegiaid. Lisaks on programmid üha enam sellised, mis kasutamise käigus ise oskusi juurde õpivad.

Need lahendused vajavad konkreetsete keelte ja neis peetavate dialoogide mikrouurimist. Sealjuures kasutatakse ära eelnevaid uurimusi, kuid uuritakse tihti ka keelt ja diskursust omapoolsete, klassikalisest keeleteadusest erinevate meetodite abil.

2.3. Suulise eesti keele ja dialoogi uurimine

Suulise kõne ja dialoogi uurimine Tartu ülikooli suulise kõne töörühmas on liikunud mitmes suunas.

Suuliste tekstide süstemaatiline kogumine ja uurimine algas 1997. aastal Tiit Hennoste juhitud projektiga "Eesti linnakeelte kogumine ja uurimine" (ETF grant 3105, 1996–2000). Seda tööd on jätkanud mitteformaalne suulise kõne töörühm, kuhu praegu kuuluvad Tiit Hennoste, Andriela Rääbis, Riina Kasterpalu, Krista Strandson ja Olga Gerassimenko. Praegu toetab korpuse täiendamist programm "Eesti keel ja rahvuslik mälu".

Selle rühma ja Tiit Hennoste suulise kõne kursuse kuulajate töö üheks tulemuseks on Tartu Ülikooli Eesti suulise keele korpus, mis sisaldab erinevat tüüpi tekste ja tekstikatkeid, mahuga ligi 1 000 000 tekstisõna (vt http://www.cl.ut.ee/suuline/).

See korpus on mõeldud erinevaid allkeeli esindava universaalse korpusena, mida saaksid oma uurimistöös kasutada eeskätt vestluse uurijad, sotsiolingvistid ja kvantitatiivsete keelesuhete uurijad.

Korpuse jaoks on valitud konversatsioonianalüüsi transkriptsioon, mida on kohandatud eesti keelele. Lisaks on iga teksti kohta esitatud taustainfo suhtlussituatsiooni kohta. Kesksed situatsioonide liigendamise alused on situatsiooni argisus/ametlikkus, teksti tegemise spontaansus/ettevalmistatus, suhtluse vahetus/vahendatus ja dialoogilisus/monoloogilisus (vt korpuse teoreetilistest alustest Hennoste 2000; Hennoste 2003b).

Selle korpuse põhjal on eri aegadel valminud rida suulise kõne uurimusi, mida võib kokku võtta mõne märksõna alla.

On kirjeldatud suulise kõne keskseid erijooni ja tehtud suulise kõne esmane kvantitatiivne analüüs. On loodud sõnavormide sagedussõnastik, kesksete bigrammide ja trigrammide sagedussõnastik ja kesksete kokkuhäädatud sõnede sagedussõnastik (vt http://www.cl.ut.ee/suuline/). On uuritud, mille poolest erinevad kirja ja kõne, argivestluste ja ametliku dialoogi sõnastike algusosad, millised on kesksed partiklid ja normikeelest erinevad sõnavormid (vt Hennoste 2000–2001; Hennoste jt 2000).

On tehtud rida kvalitatiivseid mikrouurimisi, mis toetuvad konversatsioonianalüüsile. Siin on kesksed olnud kolm suunda:

- partiklite funktsioonid ja kasutamine suulises vestluses (*jah, mhmh, vä, siis/sis, noh* jt) ning nende kaudu ka suhtluse uurimine (nt erinevate dialoogipartiklite kasutus jms; vt Hennoste 2000–2001: 1773–1806, 2465–2473, 184–193; Jansons 2002; Kasterpalu 2005);
- eneseparandused ja partneri algatatud parandused (Hennoste 2000–2001: 2689–2710; Strandson 2001);
- telefonivestluste sissejuhatused ja lõpetamised (Rääbis 2000, 2002, 2003).

2.4. Dialoogikorpus ja dialoogiaktide analüüs

Viimastel aastatel on selle rühma keskne ühine teema olnud dialoogikorpuse loomine ja dialoogisüsteemi ettevalmistamine. Selles töös osalevad lisaks suulise kõne töörühma liikmetele ka arvutilingvistika ja informaatika üliõpilased Mare Koidu juhendamisel. Tööd on toetanud ETF (grandid 4555, 2001–2003, ja 5685, 2004–2007), samuti

riiklikud programmid "Eesti keel ja rahvuskultuur" (2002, 2003) ning "Eesti keel ja rahvuslik mälu" (2004, 2005).

Selle tegevuse kaugemaks eesmärgiks on inimesega loomulikus eesti keeles suhtleva dialoogsüsteemi loomine. Sellise süsteemi jaoks tuleb arvutil modelleerida keelt kasutava inimese peas toimuvaid protsesse: partneri jutu mõistmist, vastuse planeerimist ja ülesehitamist jm.

Meid huvitab see, et loomuliku dialoogi modelleerimiseks tuleb leida need normid ja reeglid, mille kohaselt tegutsevad inimesed ja millele toetudes peaks toimima inimsuhtluse-lähedane dialoogsüsteem.

Selleks on kõigepealt koostatud infodialoogide erikorpus Eesti Dialoogikorpus (EDiK). Sellesse on valitud institutsionaalsed infodialoogid, milles üheks suhtlejaks on mingi institutsiooni esindaja (infotelefoni vastaja, reisibüroo konsultant jms) ja teiseks pooleks klient. Sealjuures on vestluse eesmärgiks info vahetamine, mitte lihtsalt viisakusvestlus (*small talk*).

Korpus sisaldas 2005. a. detsembris 873 litereeritud teksti, neist 715 telefonikõnet ja 116 silmast silma vestlust, kogupikkusega umbes 150 000 tekstisõna (vt http://math.ut.ee/~koit/Dialoog/EDiC).[5]

Teiseks oleme uurinud seda, mida inimesed sellistes dialoogides keele abil teevad ja kuidas nad oma tegevusi keeleliselt vormistavad. Selle uurimise aluseks on dialoogiaktid (tegevused, mida inimene dialoogis keele abil teeb, nagu küsimine, vastamine, tervitamine jne). Oleme loonud dialoogiaktide süsteemi, mis toetub põhialustes vestlusanalüüsile. Selle põhisüsteem on järgmine (vt Hennoste, Rääbis 2004 ja käesoleva kogumiku Lisa 2):

- Naaberpaariaktid: Rituaalid, Teemavahetusaktid, Partneri algatatud parandused, Vastuse tingimuste täpsustamise aktid, Kontakti kontroll, Direktiivid, Küsimused ja Seisukohavõtud.
- Üksikaktid: Rituaalid, Üksi tehtavad parandused, Infoaktid (sh mitte-interpreteeritavad aktid), Infolisad, Vabatahtlikud reaktsioonid.

[5] Lisaks suulistele dialoogidele sisaldab EDiK ka arvutisimulatsioonides nn võlur Ozi meetodil kogutud kirjalikke dialooge (21 teksti, umbes 2500 sõna, vt Valdisoo, Vutt 2002). Samuti kogutakse dialooge jooksvalt kahe veebis kasutatava programmi – Reisiagent ja Teatriagent – abil, mõlema programmi autor on Margus Treumuth (vt http://www.dialoogid.ee/).

Selle töö tegemise käigus oleme analüüsinud erinevaid dialoogiakte ja nende vormistamisi, mille tulemused on esitatud põhiliselt rahvusvahelistel konverentsidel ja trükitud konverentside kogumikes. Selle töö põhisuunad on järgmised:

• dialoogiaktide defineerimise ja piiride probleemid ning aktide määratlemine üldisemalt – selle töö tulemused võtab kokku raamat *Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs* (Hennoste, Rääbis 2004; vt ka Gerassimenko 2004a; Hennoste jt 2003a, 2004a, 2004b);

• küsimuste ja vastuste analüüs, mis on dialoogisüsteemi modelleerimisel keskne; ning küsimustega tihedalt seotud direktiivide analüüs (Hennoste jt 2003b, 2004b, 2005);

• partneri algatatud paranduste probleemid, mis on arvuti ja inimese suhtluse modelleerimisel väga olulised (vt Gerassimenko jt 2004b; Hennoste 2005; Hennoste jt 2005; Strandson käesolevas kogumikus);

• dialoogide algus- ja lõpurituaalide vormistamine (vt Rääbis 2000, 2002, 2003; ka Rääbis ja Kasterpalu käesolevas kogumikus);

• lisaks on tegeldud ka reaalsete suuliste dialoogide ja arvutisimulatsioonil saadud Võlur Ozi dialoogide võrdlemisega (vt Valdisoo, Vutt 2002; Valdisoo, Vutt, Koit 2003; Eskor 2005; ka Gerassimenko ja Valdisoo käesolevas kogumikus).

Korpuse koostamisel ja dialoogiaktide märgendamisel on omakorda abiks mõned arvutiprogrammid. Aktide analüüs ja märgendamine toimub seni inimese poolt, kuid selle juures on abiks programm, mis aitab kergemini analüüsi tulemusi vormistada. Selle on koostanud Evely Vutt (vt Valdisoo, Vutt 2002). Teiseks on abiks veebis kasutatav programm – dialoogikorpuse tööpink, mis lubab leida erinevalt märgendatud akte, arvutada lausungite pikkust, leida pauside, rõhuliste sõnade, kokkuhäälduste arvu, sõnade ja aktide sagedustabeleid, dialoogiaktide 2-, 3- või 4-liikmelisi ahelaid, avastada transkriptsiooni- ja märgendusvigu, kujutada dialoogi graafiliselt ajateljel jne. Selle autoriks on Margus Treumuth (vt Treumuth 2005; http://math.ut.ee/~treumuth/).

Uueks eesmärgiks on anda osa aktide määramine üle automaadile, st luua programm, mis tunneb erinevate keeleliste tunnuste abil ära teatud aktid ja märgendab need ise (vt Fishel 2005; ka Mark Fišel ja Taavet Kikas käesolevas kogumikus).

Kolmandaks astmeks dialoogsüsteemi loomisel on dialoogi modelleerimise katsed.

Mare Koit ja Haldur Õim on uurinud dialoogides kasutatavaid suhtluseesmärkide saavutamise viise ehk suhtlusstrateegiaid, partneri arutluse mõjutamise (argumenteerimise) viise ja kavandanud dialoogsüsteemi arhitektuuri (Koit, Õim 1993, 2003 ja 2004; Õim, Koit 2005). Liina Eskor on uurinud ja võrrelnud suhtlusstrateegiaid inimestevahelistes ning inimese ja arvuti vahelistes dialoogides ja modelleerinud neid pinustruktuuri abil (Eskor 2005, vt ka Eskor käesolevas kogumikus).

Margus Treumuth on koostanud lihtsa küsimus–vastus-dialoogi formaalse grammatika ja realiseerinud selle veebipõhises dialoogsüsteemis Reisiagent (http://www.dialoogid.ee/), mis vastuseks kasutaja eestikeelsele päringule suudab teatada lennukite väljumisaegu Tallinna lennujaamast. Programm lõimib mõned olemasolevad eesti keele automaattöötluse vahendid – morfoloogiline analüsaator ESTMORF ja morfoloogiline generaator ning tekst-kõnesüntesaator. Koostöös TTÜ küberneetika instituudi foneetika ja kõnetehnoloogia laboriga arendab sama autor samadel põhimõtetel dialoogsüsteemi Teatriagent, mis kasutab Eesti teatrite mängukavade andmebaasi ja millesse on lisatud ka kõnetuvastus (http://www.dialoogid.ee/, vt ka Treumuth käesolevas kogumikus).

Kirjandus

Alumäe, Tanel 2005. Large Vocabulary Continuous Speech Recognition for Estonian Using Morphemes and Classes. – Proceedings of the First Baltic Conference: Human Language Technologies – The Baltic Perspective, 2004, 166–169.

Crowdy, S. 1993. Spoken corpus design. – Literary and Linguistic Computing 8 (4), 259–264.

Du Bois, J. W., Scuetze-Coburn, S., Cumming, S., Paolino, D. 1993. Outline of discourse transcription. – Talking Data: Transcription and Coding in Discourse Research. Ed by J. A. Edwards, M. D. Lampert. Hillsdale, NJ: Lawrence Erlbaum, 45–89.

Eek, Arvo, Meister, Einar 1999. Estonian speech in the BABEL multilanguage database: Phonetic-phonological problems revealed in the

text corpus. – Proceedings of LP'98. Vol II. Ed by O. Fujimura et al, Prague: The Karolinum Press, 529–546.

Eskor, Liina 2005. Dialogue acts and communicative strategies in Estonian dialogues. – Proceedings of Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, Tallinn, 4–5 April, 225–230.

Fishel, Mark 2005. Using artificial neural networks in dialogue act recognition in Estonian dialogues. Proc. of Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, Tallinn, 4–5 April, 231–236.

Gerassimenko, Olga, Hennoste, Tiit, Koit, Mare, Rääbis, Andriela, Strandson, Krista, Valdisoo, Maret, Vutt, Evely 2004a. Annotated Dialogue Corpus as a Language Resource: An Experience of Building the Estonian Dialogue Corpus. – The First Baltic Conference "Human Language Technologies. The Baltic Perspective". Commission of the Official Language at the Chancellery of the President of Latvia, Riga, 2004, 150–155.

Gerassimenko, Olga, Hennoste, Tiit, Koit, Mare, Rääbis, Andriela 2004b. Other-initiated self-repairs in Estonian information dialogues: Solving communication problems in cooperation. – Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, April 30 – May 1, 2004. Ed by M. Strube, C. Sidner. Cambridge, 39–42.

Hennoste, Tiit 2000. Eesti suulise kõne uurimine: transkriptsioon, taust ja korpus. – Keel ja Kirjandus 2, 91–106.

Hennoste, Tiit 2000–2001. Sissejuhatus suulisesse eesti keelde I–IX. – Akadeemia 2000 5, 1117–1150; 6, 1343–1374; 7, 1553–1582; 8, 1773–1806; 9, 2011–2038; 10, 2223–2254; 11, 2465–2486; 12, 2689–2710; 2001 1, 179–206.

Hennoste, Tiit 2003a. Keelekasutuse uurimine. – Eesti keele uurimise analüüs. Emakeele Seltsi Aastaraamat 48 / 2002. Koost M. Erelt. Tallinn: Emakeele Selts, 217–262.

Hennoste, Tiit 2003b. Suulise eesti keele uurimine: korpus. – Keel ja Kirjandus 7, 481–500.

Hennoste, Tiit 2005. Repair-initiating particles and um-s in Estonian spontaneous speech. – Proceedings of DISS'05, Disfluency in Spontaneous Speech Workshop. 10–12 September 2005, Aix-en-Provence, Universite en Provence, France, 83–88.

Hennoste, Tiit Andriela Rääbis 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu.

Hennoste, Tiit, Lindström, Liina, Rääbis, Andriela, Toomet, Piret, Vellerind, Riina 2000: Eesti suulise kõne korpus ja mõne allkeele võrd-

lemise katse. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim T. Hennoste. Tartu, 245–284.

Hennoste, Tiit, Koit, Mare, Kullasaar, Maret, Rääbis, Andriela, Vutt, Evely 2002. Eesti dialoogikorpuse loomise probleemid. – Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks. Toim R. Pajusalu, T. Hennoste. TÜ üldkeeleteaduse õppetooli toimetised 3, Tartu, 143–160.

Hennoste, Tiit, Koit, Mare, Rääbis, Andriela, Strandson, Krista, Valdisoo, Maret, Vutt, Evely 2003a. Developing a typology of dialogue acts: Some boundary problems. – Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo July 5-6, 2003, 226–235.

Hennoste, Tiit, Koit, Mare, Rääbis, Andriela, Strandson, Krista, Valdisoo, Maret, Vutt, Evely 2003b. Directives in Estonian information dialogues. – Text, Speech and Dialogue. Proceedings of the 6th International Conference, TSD 2003, České Budějovice, Czech Republic, September 8–12, 2003. Ed by V. Matoušek, P. Mautner. Berlin: Springer, 406–411.

Hennoste, Tiit, Koit, Mare, Rääbis, Andriela, Valdisoo, Maret 2004a. Developing a dialogue act coding scheme: An experience of annotating the Estonian Dialogue Corpus. – LREC 2004. IV International Conference On Language Resources and Evaluation. Workshop: Compiling and Processing Spoken Language Corpora. 24th May 2004. Lisboa, Portugal. Ed by N. Oostdijk, G. Kristoffersen, G. Sampson. Lisboa, 40–47.

Hennoste, Tiit, Koit, Mare, Strandson, Krista, Rääbis, Andriela, Valdisoo, Maret, Vutt, Evely 2004b. Küsimuste ja direktiivide märgendamine eestikeelsetes infodialoogides. – Toimiv keel II. Töid rakenduslingvistika alalt. Tallinna Pedagoogikaülikooli eesti filoloogia osakonna toimetised 3. Koost H. Metslang, toim M.-M. Sepper, J. Lepasaar. Tallinn, 138–154.

Hennoste, Tiit, Gerassimenko, Olga, Kasterpalu, Riina, Koit, Mare, Rääbis, Andriela, Strandson, Krista, Valdisoo, Maret 2005. Questions in Estonian nformation Dialogues: Form and Functions. Text, Speech and Dialogue. Proceedings of the 8th International Conference TSD 2005. Ed by V. Matousek, P. Mautner. Berlin: Springer, 420–427.

Jansons, Airi 2002. Partikli *siis* funtsioonid eesti suulises kõnes. – Keel ja Kirjandus 5, 344–358.

Johansson, S. 1995. The approach of the Text Encoding Initiative to the encoding of spoken discourse. – Spoken English on Computer.

Transcription, Mark-up and Application. Ed by G. Leech, G. Myers, J. Thomas. London: Longman, 82–98.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim T. Hennoste. Tartu, 87–100.

Kasterpalu, Riina 2005. Partiklid jah, jaa ning jajaa naaberpaari järelliikmena müügiläbirääkimistes. – Keel ja Kirjandus 11, 873–890, 12, 996–1000.

Keevallik, Leelo 2002. Grammatika suhtluses. – Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Toim R. Pajusalu, I. Tragel, T. Hennoste & H. Õim. Tartu, 89–104.

Keevallik, Leelo, 2003. Colloquial Estonian. – Estonian Language. Linguistica Uralica Supplementary Series. Volume 1. Ed by Mati Erelt, Tallinn: Estonian Academy Publishers, 342–378.

Koit, Mare 2003. Märgendatud dialoogikorpus kui keeleressurss. – Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi Toimetised 12, 119–136.

Koit, Mare, Õim, Haldur 1993. A formal model of communicative strategy. – Proceedings of the Scandinavian Conference on Artificial Intelligence '93. Stockholm. 226 –231.

Koit, Mare, Õim, Haldur 2003. Eestikeelse dialoogi modelleerimine. – Keel ja Kirjandus, 10, 721-735.

Koit, Mare, Õim, Haldur 2004. Argumentation in the Agreement Negotiation Process: A Model that Involves Natural Reasoning. Proceedings of the Workshop W12 on Computational Models of Natural Argument. 16th European Conference on Artificial Intelligence, Valencia, Spain, August 2004, 53–56.

Leech, G., Myers, G., Thomas, J. (ed) 1995. Spoken English on Computer. Transcription, Mark-up and Application. London: Longman.

Lindström, Liina, Pajusalu, Karl 2003. Corpus of Estonian Dialects and the Estonian Vowel System. – Linguistica Uralica 4, 241–257.

Lindström, Liina, Lonn, Varje, Mets, Mari, Pajusalu, Karl, Teras, Pire, Veismann, Ann, Velsker, Eva, Viikberg, Jüri 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. Tartu Ülikooli eesti keele õppetooli toimetised 17. Toim R. Kasik. Tartu, 186–211.

Meister, Einar 2001. Towards speech recognition in Estonian. – 21. fonetiikan päivät, Turku 4.-5.1.2001. Publications of the Department of Finnish and General Linguistics of the University of Turku, Turun

yliopiston ja yleisen kielitieteen laitoksen julkaisuja 67. Ed by S. Ojala, J. Tuomainen. Turku, 59–70.

Meister, Einar, Lasn, Jürgen, Meister, Lya 2003. Development of the Estonian SpeechDat-like Database. Proceedings of Eurospeech 2003, Geneva 2003, vol 2, 1601–1604.

Mihkla, Meelis, Meister, Einar 2002. Eesti keele tekst-kõnesüntees. – Keel ja Kirjandus 2, 88–97; 3, 173–182.

Mihkla, Meelis, Meister, Einar, Eek, Arvo 2000. Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. – Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Arvutuslingvistikalt inimesele. Toim T. Hennoste, Tartu, 309–319.

Pajusalu, Karl, Hennoste, Tiit, Niit, Ellen, Päll, Peeter, Viikberg, Jüri 2002. Eesti murded ja kohanimed. Õpik. Toim T. Hennoste. Tallinn: Eesti Keele Sihtasutus.

Rääbis, Andriela 2000. Telefonivestluse sissejuhatus. – Keel ja Kirjandus 6, 409–418.

Rääbis, Andriela 2002. Ametlike telefonikõnede lõpetamine. – Emakeele Seltsi Aastaraamat 47/2001. Tallinn, ETA Emakeele Selts, 107–125.

Rääbis, Andriela 2003. Olukorda kontrollivad küsimused telefonivestluse alguses. – VIRSU. Viro ja suomi: kohdekielet kontrastissa. Toim P. Muikku-Werner, H. Remes. Lähivertailuja 13, Joensuu, 236–246.

Strandson, Krista 2001. Kuidas vestluskaaslane parandusprotsessi algatab. – Keel ja Kirjandus 6, 394–409.

Tainio, Liisa (toim). 1997. Keskustelunanalyysin perusteet. Tampere: Vastapaino.

Treumuth, Margus 2005. A software tool for the Estonian Dialogue Corpus. Proc. of Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, Tallinn, 4–5 April, 341–346.

Valdisoo, Maret, Vutt, Evely 2002. "Võlur Ozi" tehnika ja eesti keeles suhtlev arvuti. – A&A 5, 63–67.

Valdisoo, Maret, Vutt, Evely, Koit, Mare 2003. On a method for designing a dialogue system and the experience of its application. – Journal of Computer and Systems Sciences International 42/3, 456–464.

Õim, Haldur, Koit, Mare 2005. Developing a Dialogue System that Interacts with a User in Estonian. A Finnish Computer Linguist: Kimmo Koskenniemi. Festschrift on the 60th birthday. Ed by Arppe, Carlson, Heinamäki, Linden, Miestamo, Piitulainen, Tupakka, Westerlund, Yli-Jyrä et al. CSLI Publications, 278–288.

Infodialoogi algusrituaalid[1]

Andriela Rääbis
Tartu Ülikool

1. Sissejuhatus

Käesolevas artiklis vaadeldakse, kuidas alustatakse infodialooge. Analüüsi aluseks on 577 institutsionaalset telefonikõnet Eesti dialoogikorpusest (vt http://math.ut.ee/~koit/Dialoog/EDiC).

Suhtluse keskmes on infoküsimine. Helistajaks on kõigis telefonikõnedes klient (infoküsija) ning vastajaks ametiisik (infoandja). Klient on enamasti eraisik, ainult mõnes üksikus telefonikõnes suhtlevad omavahel kaks institutsiooni esindajat (nt helistatakse raadiost politseisse, päästeametisse; sotsiaalabi osakonnast polikliiniku registratuuri). Situatsioonidest on kesksed helistamine infotelefonile (390 telefonikõnet, neist 319 vastab sama töötaja), polikliiniku registratuuri (59, neist 49 vastab sama töötaja), reisibüroosse (47), bussijaama (12), takso tellimine (17, üks dispetšer). Muid situatsioone on 52, igast tüübist on lindistatud 1–6 telefonikõnet (nt helistamine kauplusse, ehitusfirmasse, autoteenindusse, sotsiaalabi osakonda, ülikooli õppetooli, panka, kinnisvarabüroosse, raamatukogusse, hotelli, pensioniametisse, postkontorisse, kinno, spordikeskusesse, lasteaeda jne). Osalejad on enamasti võõrad, vaid 13 dialoogis tuttavad (polikliiniku arst või õde helistab registratuuri). Seega on tegu ametikõneluste tuumrühmaga (vt Hennoste 2003). Peaaegu kõik kõned on lauatelefonile, neli kõnet mobiilile.

Dialoogide litereerimisel on kasutatud konversatsioonianalüüsi transkriptsiooni (vt Lisa 1). Osalejate anonüümsuse huvides on kõik nimed transliteratsioonides muudetud. H tähistab helistajat, V vastajat.

[1] Artikli valmimist on toetanud ETF (grant 5685) ning HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

2. Telefonivestluse sissejuhatuse täielik mudel

Telefonivestlust alustades ei minda tavaliselt kohe teema juurde. Teemale eelneb sissejuhatus: osalejad identifitseerivad teineteise, tervitavad, vahetavad viisakusväljendeid. Väga lühikese aja jooksul vahetatakse palju informatsiooni. Telefonivestluse sissejuhatus koosneb teatud hulgast ülesannetest, mida täidetakse kindlas järjekorras.

Konversatsioonianalüüsi meetodil hakati telefonivestlusi uurima 1960n. aastatel (Sacks 1992 [1964–1972]; Schegloff 1967; 1968; 1979) ning enim tähelepanu on pööratud just vestluste sissejuhatustele. E. Schegloff (1986: 113–133) kirjeldab telefonivestluse sissejuhatuse täielikku mudelit. Ta leidis, et enamik analüüsitud 450 telefonikõnest algab nelja sekventsi järjendiga. Need sekventsid on järgmised:

1. Kutsung – vastus: telefonihelin ja vastaja esimene voor.
2. Identifitseerimine (ja/või äratundmine): osalejad tutvustavad ennast ja väljendavad partneri äratundmist.
3. Tervitused.
4. *Kuidas läheb*-sekventsid.

Pärast neid nelja sekventsi jõuavad vestlejad sellisesse positsiooni, et nad saavad alustada esimest teemat. Schegloff (1986: 116) nimetab seda ankrupositsiooniks. Tavaliselt alustab esimest teemat helistaja, kes teatab oma helistamise põhjuse.

Ingliskeelseid telefonikõnesid uurides on leitud, et niisuguse mudeli järgi viiakse sissejuhatused läbi, kui osalejad on omavahel tuttavad. Lähedaste või võõraste inimeste puhul on sissejuhatus enamasti redutseeritud. Lähedased jätavad vahele identifitseerimissekventsi. Võõrastevahelise suhtluse puhul võidakse ära jätta kõik sissejuhatuse osad peale kutsung–vastuse: enamasti vastaja tutvustab end ning helistaja läheb kohe oma esimeses voorus asja juurde (Whalen, Zimmerman 1987; Hopper, Doany, Johnson, Drummond 1990/91; Hopper, Drummond 1992; Hopper 1992: 77–83).

Eesti telefonivestluse sissejuhatuse täielik mudel koosneb samuti neljast sekventsist, kuid osade järjekord on teistsugune: tervitatakse enne tutvustamist (Rääbis 2000: 418):

1. Kutsung – vastus.
2. Tervitused.
3. Identifitseerimine.
4. Olukorra kontrollimine.

Neljaosaline sissejuhatus ei ole Eestis tüüpiline. Täielik mudel on pigem neutraalne tegevuste loend, mille hulgast valitakse igas konkreetses telefonikõnes mõned sekventsid olenevalt vestluse tüübist, osalejate suhetest ja konkreetsest situatsioonist.

3. Infodialoogide algusrituaalide märgendamine

Rituaalid on tegevused, mis sooritatakse vormeliteks nimetatavate kindlate keeleliste üksuste abil (vormelite kohta vt Hennoste 2000: 2246–2248). Vormeleid iseloomustavad järgmised tunnused:

- kinnistunud grammatiline struktuur,
- kinnistunud sõnajärg,
- kinnistunud leksikaalsemantiline koosseis,
- algse sõnasõnalise tähenduse kitsenemine või teisenemine,
- ehituse ebaregulaarsus (EKG II 1993: 229).

Dialoogide algusrituaalide märgendamiseks kasutame järgmisi akte[2]:
RIE: KUTSUNG
RIJ: KUTSUNGI VASTUVÕTMINE
RIE: TERVITUS
RIJ: VASTUTERVITUS
RY: TUTVUSTUS
RY: ÄRATUNDMINE

Kutsung – kutsungi vastuvõtmine on telefonivestluse esimene naaberpaar. Silmast silma vestlust alustades on tüüpiline sümmeetriline kohtumine, telefonivestluse alustamiseks on kutsung ainus võimalus. Vastaja esimene voor on kutsungile vastuseks. Lausung võib olla ainult kutsungi vastuvõtmise funktsioonis (*jaa*; *hallo*) või täita korraga mitut rolli, nt kutsungi vastuvõtmine ja tutvustus (*ujula*) (vt ka Hennoste, Rääbis 2004: 47–48). Institutsionaalses suhtluses on teine variant ülekaalukalt sagedasem: normikohaselt vastaja tutvustab ennast. Kui vastaja esimene voor koosneb mitmest komponendist, siis saab topeltmärgendite hulga vähendamiseks märgendi RIJ: KUTSUNGI VASTUVÕTMINE ainult vooru esimene komponent.

Infodialoogi teine naaberpaar on tervitused. Eesti institutsionaalsetes telefonikõnedes tervitatakse partnerit peaaegu alati. 505 dialoogis tervitasid mõlemad osalejad, 70 dialoogis jäi tervitus vastuseta. Tervitused puudusid vaid kahes telefonikõnes. Analüüsitud dialoogides on

[2] Dialoogiaktide loend vt Lisa 2.

tervitussõnad järgmised: *tere* (933 korda), *tervist* (83), *tere päevast* (77), *tere õhtust* (12), *tere hommikust/ommikust/ omikust/omigust* (9), *tere hommikut/ommikut* (2), *tervitus* (1), *tšau* (1).

Tutvustus ei ole naaberpaariakt, kuna ta ei nõua partnerilt enesetutvustust. Eestis on tavaks, et võõrastevahelises institutsionaalses dialoogis helistaja ennast ei tutvusta. Erandiks on situatsioonid, kus inimene peab ennast info saamiseks identifitseerima, nt arsti vastuvõtule registreerimine. Kuid ka siis ei tutvustata end sageli kohe vestluse alguses, vaid hilisemas vestluses: enne leitakse patsiendile sobiv vastuvõtuaeg, seejärel küsib registratuuritöötaja patsiendi nime.

Vastaja tutvustab end järgmiselt:

- ettevõtte nimi (üldisem – nt *kauplus*, *näitus*, *ujula* – või täpne nimi, nt *Maria takso*, *Kuku kliinik*) (97 korda);
- allasutus / osakond (*registratuur*, *silmakabinet*) (51 korda);
- identifitseeritakse ka enda positsioon (*politsei korrapidaja*) (1 kord);
- telefoninumber (1 kord);
- eesnimi (3 korda);
- eesnimi + perekonnanimi (1 kord);
- ettevõtte nimi + eesnimi (405 korda);
- ettevõtte nimi + osakond + eesnimi (1 kord);
- ettevõtte nimi + osakond + eesnimi + perekonnanimi (1 kord).

Vastaja ei tutvustanud end 16 dialoogis.

Kui tutvustus koosneb mitmest osast, saab iga komponent eraldi märgendi.

Äratundmine on akt, mida kasutatakse eeskätt eravestlustes. Institutsionaalsetes võõrastevahelistes dialoogides on ta haruldane.

4. Sissejuhatuse põhistruktuurid

Normiks on niisugune sissejuhatus, kus vastaja ennast tutvustab ning mõlemad osalejad tervitavad. Sõltuvalt vastaja tegevustest oma esimeses voorus on sissejuhatusel kaks varianti. 1. mudeli puhul algab dialoog vastaja enesetutvustusega ning helistajat tervitatakse kohe samas voorus. Järgmises voorus tervitab helistaja vastu ning esitab oma soovi või küsimuse.

1. mudel

kutsung
V: tutvustus + tervitus
H: tervitus + soov / küsimus

(1)
((kutsung)) | RIE: KUTSUNG |
V: `info`telefon= | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
`Kersti= | RY: TUTVUSTUS |
tere? | RIE: TERVITUS |
(.)
H: e tere, | RIJ: VASTUTERVITUS |
Pärnu õ `loodussõprade maja palun. | DIE: SOOV |

Sissejuhatuse 2. mudelis viiakse ühes voorus läbi vaid üks tegevus.

2. mudel
kutsung
V: tutvustus
H: tervitus
V: tervitus
H: soov / küsimus

(2)
((kutsung)) | RIE: KUTSUNG |
V: `Nordik=Reisid= | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
`Piia=kuuleb. | RY: TUTVUSTUS |
H: tere? | RIE: TERVITUS |
V: tere | RIJ: VASTUTERVITUS |
H: tahaks odavalt `Inglismaale sõita. | DIE: SOOV |

1. mudel on statistiliselt selgelt ülekaalus: 416 sissejuhatust (72%, vt tabel 1). Statistikat mõjutab infotelefonidialoogide suur hulk, kus vastaja esimene voor on kindel vormel, mida ei varieerita. 2. mudelit kasutatakse suhteliselt harva, vaid 24 telefonikõnes (4%).

Tabel 1. Sissejuhatuse põhistruktuuride esinemissagedus

	1. mudel	2. mudel	Eripärased
Kokku	**416 (72%)**	**24 (4%)**	**137 (24%)**
Infotelefon	332	–	58
Registratuur	35	2	22
Reisibüroo	17	10	20
Bussijaam	6	3	3
Taksofirma	13	–	4
Muu	13	9	30

5. Eripärased sissejuhatused

5.1. Kõige sagedasem kõrvalekalle 1. mudelist on helistaja vastutervituse puudumine, kohe esitatakse oma soov (36 telefonikõnes, neist 30 infotelefon). Nii sagedast sissejuhatuse varianti normiks pidamast takistab asjaolu, et tervitused peaksid vestluses esinema naaberpaarina. Seega on helistaja niisugune tegevus ühelt poolt suhtlusnorme rikkuv, teiselt poolt aga kooperatiivne: oma soov püütakse edastada võimalikult kiiresti.

(3)
((kutsung)) | RIE: KUTSUNG |
V: `infotelefon | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
`Kersti= | RY: TUTVUSTUS |
tere? | RIE: TERVITUS |
H: ma paluksin `Tartust `Soola tänavalt `Talu`puu. | DIE: SOOV |

5.2. Seitsmes telefonikõnes olid 1. mudeli puhul vastaja esimese vooru komponendid teises järjekorras: tervitus + tutvustus. Helistaja esimene voor on kõigis neis dialoogides tervitus + soov/küsimus nagu 1. mudelis.

5.3. 13 telefonikõnes puudus vastaja vastutervitus. Dialoog algab vastaja tutvustusega, helistaja tervitab ning esitab kohe oma soovi või küsimuse. Enamasti järgneb soov tervitusele kiiresti, helistaja ei anna vastutervituseks võimalustki, kolmes sissejuhatuses aga järgneb tervitusele paus ning sellele vaatamata vastutervitust ei tule (näide 4).

(4)
((valitakse numbrit, telefon heliseb)) | RIE: KUTSUNG |
V: Topolaims? | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
H: tere? | RIE: TERVITUS |
(0.5)
sooviks saada informatsi`ooni teie poolt pakutavate sõitude kohta
Eu`roopasse. | DIE: SOOV |

5.4. Vastaja ei tutvustanud ennast 16 dialoogis. Esimene voor on sel juhul *hallo/haloo/aljo* (6 korda), *jaa* (4), *jah* (2), *jaa.* (.) *ma kuulen teid* (1), 3 dialoogis alustas vastaja vestlust tervitusega.

Identifitseerimisprobleem tekkis kolmes dialoogis: helistaja küsis, kas see on soovitud ettevõte.

(5)
((kutsung)) | RIE: KUTSUNG |
V: jaa? | RIJ: KUTSUNGI VASTUVÕTMINE |
H: tere `hommikust. | RIE: TERVITUS |

V: tere? | RIJ: VASTUTERVITUS |
H: e on se `lammutus. | KYE: SULETUD KAS |
V: jah | KYJ: JAH |
H: ee `kas teil Re`noo juppe ka `sees on. | KYE: SULETUD KAS |

5.5. Samasuguse küsimuse esitas helistaja aga ka kahes dialoogis, kui vastaja oli ennast tutvustanud.

5.6. Kahes dialoogis soovis helistaja rääkida konkreetse inimesega.

(6)
((kutsung)) | RIE: KUTSUNG |
V: ((kõrvalolijale:)) {-} | YA: PRAAK |
((helistajale:)) X=vald | RIJ: KUTSUNGI VASTUVÕTMINE || RY: TUTVUSTUS |
tere=päevast | RIE: TERVITUS |
H: £ .hh te:re päevast. | RIJ: VASTUTERVITUS |
.hh vabandage, (.) `mina tahaksin rääkida:: X valla sotsiaal`töötajaga. £
| DIE: SOOV |
V: jaa ma `kuulen. | DIJ: MUU |

5.7. Helistaja tutvustas ennast järgnevatel juhtudel.

• Osalejad on omavahel tuttavad. Polikliiniku registratuuri helistanud kolleeg tutvustas end ühes telefonikõnes eesnimega, ühes perekonnanimega (*doktor Alpikann siin*). Ülikooli õppetooli helistanud töökaaslane ütles oma eesnime.

• Suhtlevad kaks institutsiooni esindajat (3 telefonikõnet). Helistaja ei tutvusta ennast nimepidi, vaid ütleb oma institutsiooni nime (*mina (.) tülitan teid X > linnavalitsuse < `sotsiaalabi osakonnast*).

• Helistaja peab end soovitud info saamiseks identifitseerima: patsient on unustanud, mis kellaks ta on arsti vastuvõtule registreeritud; ema pole saanud lastetoetust (näide 7); klient soovib teada, miks ta arvuti remont venib. Tutvustus oli vaid nelja niisuguse telefonikõne alguses. Ka arsti vastuvõtule registreerides, hotellituba broneerides ja taksot tellides peab helistaja oma nime ütlema, kuid seda tehakse alles hilisemas vestluses, kui ametnik nime küsib.

(7)
((kutsung)) | RIE: KUTSUNG |
V: pensioniamet | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
H: tere. | RIE: TERVITUS |
V: tere? | RIJ: VASTUTERVITUS |
H: .hh minu `nimi=ee on Jaana `Kasesalu, | RY: TUTVUSTUS |
(.)
.hh e ma=i=ole `saanud nüd `viimase `kuu laste`toetust.= | DIE: SOOV |

• Klient tutvustas end kahes infotelefonidialoogis. Üks neist oli välismaalane (*mina helistan `Tartust ma olen x kiriku `preester isa `Johan*), teisel juhul helistati valel numbril.

Seega on institutsionaalses dialoogis väga haruldane, et helistaja end tutvustab.

5.8. Kolmes registratuuridialoogis väljendas helistaja vastaja (kolleegi) äratundmist. Partner tunti ära hääle järgi, oma nime ta ei olnud öelnud.

(8)
((kutsung)) | RIE: KUTSUNG |
V: registratuur | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
tere | RIE: TERVITUS |
(.)
H: Liina, | RY: ÄRATUNDMINE |
V: [noh] | VR: NEUTRAALNE JÄTKAJA |
H: [ole] hea, ütle palun, millal ortopeedid meile siia tulevad, | KYE: AVATUD |

Helistaja nimetas vastajat nimepidi ka kahes infotelefonidialoogis (vastaja oli end eesnimega tutvustanud). Äratundmisega siin tegu ei ole, kuna osalejad on võõrad.

5.9. 11 telefonikõnet olid jätkuks eelmisele kõnele: ühendus oli katkenud, kliendil oli palutud natukese aja pärast uuesti helistada või oli saadud info helistaja arvates vale või ebapiisav. Siis alustab helistaja fraasidega *ma `vist ei `rääkind teiega=meil `katkes ühendus*; *ee vabandust, mul katkes enne kõne `ära*; *ma helistasin minut tagasi*; *ma helistan selle `Tartusse sõidu pärast `uuesti* vms. Neist kuues telefonikõnes puudus helistaja tervitus.

5.10. Normikohaselt peaks telefonivestluses esimesena rääkima vastaja. Kahes infotelefonidialoogis alustas vestlust aga helistaja ning ühes dialoogis hakkasid osalejad korraga rääkima. Vastaja alustab sellegipoolest rutiinse vormeliga ning vestlus jätkub põhimalli järgi:

(9)
((kutsung)) | RIE: KUTSUNG |
H: allo? | KKE: ALGATUS |
V: info`telefon= | RIJ: KUTSUNGI VASTUVÕTMINE | | KKJ: KINNITAMINE |
| RY: TUTVUSTUS |
> Kersti < = | RY: TUTVUSTUS |
tere:. | RIE: TERVITUS |
H: te:re. | RIJ: VASTUTERVITUS |
palun kodumasinate telefoninumber Pärnus. | DIE: SOOV |

Viies infotelefonidialoogis hakkas helistaja kõnelema vastaja tutvustuse ajal ning viimane katkestas oma vooru.

(10)
((kutsung)) | RIE: KUTSUNG |
V: `infotelefon | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
[Kersti] | RY: TUTVUSTUS |
H: [tere.] | RIE: TERVITUS |
`Kesko Aagised `paluks. | DIE: SOOV |

5.11. Kaheksas dialoogis oli kolm tervitust. Tervituste naaberpaari alustab vastaja, helistaja tervitab vastu ning seejärel tervitab vastaja veel kord. See oli iseloomulik eelkõige reisibüroodialoogidele. Vastaja esimene tervitus on rutiinse vormeli osa, teine tervitus aga vastab just selle helistaja tervitusele. Reisibüroodialoogid on vaadeldud dialoogidest ainsad, kus midagi müüa proovitakse ja vestlus ei ole puhtinformatiivne. Reisiinfo andja püüab kliendiga lähedasemat suhet luua kui infotelefoni töötaja.

(11)
((kutsung)) | RIE: KUTSUNG |
V: Euroreisibüroo | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
tere=päevast? | RIE: TERVITUS |
H: tere päevast. | RIJ: VASTUTERVITUS |
V: tere | RIJ: VASTUTERVITUS |
H: mul on selline `küsimus, kas teie büroo `Kreekasse ka reise korraldab.
| KYE: JUTUSTAV KAS |

5.12. Institutsionaalses dialoogis üliharuldane on helistaja tervitusele järgnev jätkaja (vt ka näide 8, kus vastaja väljendab partikliga *noh* kolleegi äratundmist ja õhutab teda edasi rääkima).

(12)
((kutsung)) | RIE: KUTSUNG |
V: ((kõrvale)) mhmh | YA: PRAAK |
((torusse)) registratuur | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
tere? | RIE: TERVITUS |
(.)
H: tere päevast. | RIJ: VASTUTERVITUS |
V: jah? | VR: NEUTRAALNE JÄTKAJA |
H: ee `oskate te `öelda `doktor `Kaldma `telefoni[numbrit.] | KYE: JUTUSTAV KAS |

5.13. Tervitused puudusid vaid kahes telefonikõnes.

5.14. Kuulmishäire on üsna sagedaseks sissejuhatuse põhimallist kõrvalekaldumise põhjuseks. Näites 13 järgneb helistaja tervitusele pikk paus ning vastaja kontrollib suhtluskanali funktsioneerimist.

(13)
((kutsung)) | RIE: KUTSUNG |
V: info`telefon= | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
Kersti= | RY: TUTVUSTUS |
tere | RIE: TERVITUS |
(0.8)
H: tere | RIJ: VASTUTERVITUS |
(2.2)
V: halloo? | KKE: ALGATUS |
H: halloo, | KKJ: KINNITAMINE |
öelge palun=e `Tartu `pensioniameti `number. | DIE: SOOV |

Tabelis 2 on esitatud eripäraste sissejuhatuste tüübid. Mõnes sissejuhatuses esineb korraga mitu nähtust ja need on tabelis eraldi välja toodud. Kuulmishäire tõttu tekkinud sissejuhatuste eripärasid (nt tervituste kordamine) ei ole eraldi variantidena vaadeldud.

Tabel 2. Kõrvalekalded sissejuhatuse põhimudelitest

	Info-telefon	Registratuur	Reisibüroo	Bussijaam	Taksofirma	Muu
1. mudel, H vastutervitus puudub	30	4	–	–	2	–
1. mudel, V 1. voor tervitus + tutvustus	1	–	2	–	–	4
2. mudel, V vastutervitus puudub	–	1	5	2	–	5
V tutvustus puudub	–	1	2	1	1	11
H esitab identifitseerimisküsimuse	2	–	–	–	–	3
H soovib rääkida konkreetse inimesega	–	–	–	–	–	2
H tutvustab ennast	2	5	–	–	–	5
H nimetab vastajat nimepidi	2	3	–	–	–	–
Viide eelmisele kõnele	4	4	2	–	–	1
H alustab varem	8	–	–	–	–	–
1. mudel, 3 tervitust	–	1	6	–	–	1
V jätkaja pärast H tervitust / äratundmist	–	2	–	–	–	–

	Info-telefon	Registratuur	Reisibüroo	Bussijaam	Taksofirma	Muu
Tervitused puuduvad	–	–	–	–	–	2
Kuulmishäire	9	2	4	–	1	1

6. Kokkuvõte

Infodialoogi sissejuhatus on enamasti väga lühike, klient esitab soovi või küsimuse juba oma esimeses voorus. Voorud koosnevad tüüpiliselt mitme sekventsi osadest (nt tutvustus + tervitus; tervitus + küsimus).

Infotelefonidialoogi sissejuhatus viiakse enamasti (85%) läbi 1. mudeli järgi: vastaja esimene voor koosneb tutvustusest ning tervitusest, klient tervitab vastu ning esitab samas voorus oma soovi või küsimuse. Sagedasemad kõrvalekalded sellest mudelist on kliendi tervituse puudumine ning kliendi poolt vestluse alustamine enne vastaja esimese vooru lõppu. Seega on kõrvalekalded põhimudelist tingitud peamiselt sellest, et klient püüab oma soovi esitada võimalikult kiiresti.

Registratuuridialoogide sissejuhatus kulgeb samuti peamiselt 1. mudeli järgi (59%), kuid variatiivsus on märksa suurem: eripäraseid sissejuhatusi on 37%. Selles rühmas põhjustab kõrvalekaldeid põhimudelitest eelkõige asjaolu, et 13 dialoogis on osalejad tuttavad: helistaja tutvustab ennast ja/või väljendab vastaja äratundmist; kuna helistatakse päeva jooksul mitu korda, siis ei tervitata.

Reisibüroodialoogide sissejuhatused viiakse läbi nii 1. kui 2. mudeli järgi (vastavalt 36% ja 21%). Nende sissejuhatuste eripäraks on vastaja vastutervituse puudumine 2. mudeli puhul ning kolmanda tervituse lisamine 1. mudelile.

Bussiinfodialooge on järelduste tegemiseks liiga vähe. 50% sissejuhatustes kasutatakse 1. mudelit ja 25% 2. mudelit.

76% taksotellimisdialoogidest algab samuti 1. mudeli järgi. Ühes dialoogis puudub vastaja tutvustus ning kahes dialoogis kliendi tervitus.

Muude dialoogide rühm on eripäraste variantide poolest kõige mitmekesisem. 1. mudeli järgi alustatakse vaid 25% ja 2. mudeli järgi 17% vestlusi. Erinevalt teistest rühmadest puudub suurel osal vastajatest tõenäoliselt suhtlemiskoolitus ning telefonile vastamine pole

nende põhitöö. Silmatorkavalt palju vastajaid ei tutvusta ennast, kuid suhtlusprobleeme see enamasti ei põhjusta.

Kuna vaadeldud allkorpused on väga erinevad nii dialoogide arvu kui ka erinevate kõnelejate hulga poolest, võib saadud statistikat alles esialgseks tulemuseks pidada. Dialooge tuleb kindlasti juurde koguda, et situatsioonitüüpe oleks korpuses enamvähem võrdselt. Infotelefonidialooge on kogutud küll juba üsna palju, kuid enamik neist on salvestatud ühe töötaja poolt. Ka kõik taksotellimisdialoogid on salvestanud vaid üks dispetšer. Alakogum "Muud" vajab edaspidi liigendamist, kui sinna on kogunenud piisavalt palju erinevaid situatsioone. Tasakaalustatud märgendatud dialoogikorpus on vajalik nii eestikeelse dialoogsüsteemi loomiseks kui ka suulise kõne uurimiseks laiemalt.

Kirjandus

EKG II 1993. = Erelt, Mati (toim) Eesti keele grammatika II. Süntaks. Tallinn: ETA Keele ja Kirjanduse Instituut.

Hennoste, Tiit 2000. Sissejuhatus suulisesse eesti keelde VIII. Lausung suulises kõnes III. Eneseparandused. – Akadeemia 12, 2223–2254.

Hennoste, Tiit 2003. Suulise eesti keele uurimine: korpus. – Keel ja Kirjandus 7, 481–500.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Hopper, Robert 1992. Telephone Conversation. Bloomington: Indiana University Press.

Hopper, Robert, Doany, Nada, Johnson, Michael, Drummond, Kent 1990/91. Universals and particulars in telephone openings. – Research on Language and Social Interaction 24, 369–387.

Hopper, Robert, Drummond, Kent 1992. Accomplishing interpersonal relationship: The telephone openings of strangers and intimates. – Western Journal of Communication 56(3), 185–199.

Rääbis, Andriela 2000. Telefonivestluse sissejuhatus. – Keel ja Kirjandus 6, 409–418.

Sacks, Harvey 1992 [1964–1972]. Lectures on conversation. Ed by G. Jefferson. Cambridge & Oxford: Blackwell.

Schegloff, Emanuel 1967. The first five seconds: the order of conversational openings. Unpublished PhD dissertation, University of California, Berkeley.

Schegloff, Emanuel 1968. Sequencing in conversational openings. – American Anthropologist 70, 1075–1095.

Schegloff, Emanuel 1979. Identification and recognition in telephone conversation openings. – Everyday Language: Studies in Ethnomethodology. Ed by G.Psathas. New York: Irvington, 23–78.

Schegloff, Emanuel 1986. The routine as achievement. – Human Studies 9, 111–152.

Whalen, Marilyn R., Zimmerman, Don H. 1987. Sequential and institutional contexts in calls for help. – Social Psychology Quarterly 50, 172–185.

Algusrituaalid müügikõnedes[1]

Riina Kasterpalu
Tartu Ülikool

1. Sissejuhatus

Käesolev artikkel kasvas välja ühe Eestis tegutseva koolitusfirma telefoni teel peetud müügivestluste märgendamisel üles kerkinud probleemidest ja neile leitud lahendustest dialoogikorpuse töörühma koosolekutel. Eesti dialoogikorpusesse (Hennoste, Rääbis 2004: 11; vt ka http://math.ut.ee/~koit/Dialoog/EDiC) on kuulunud käesoleva aastani (väheste eranditega) peamiselt sellised dialoogid, kus helistajaks on eraisik ja vastaja esindab ametiasutust. Praeguseks on dialoogikorpuse koosseisu lisandunud 44 müügivestlust (27776 sõna), mis erinevad mitme tunnuse poolest senistest dialoogidest, moodustades dialoogikorpuses omanäolise allkorpuse. Üks tunnuseid, mis eristab müügivestlusi teistest Eesti dialoogikorpusse seni kuulunud institutsionaalsetest dialoogidest, on see, et mõlemad suhtlejad on ametiisikud.

Märgendades müügivestlusi Eesti dialoogikorpusesse, ilmnes vajadus täiendada olemasolevat märgendustüpoloogiat EDiT, mis on loodud arvestades dialoogikorpusesse seni kuulunud dialoogide struktuuri (Hennoste, Rääbis 2004: 13–14), ja lisada telefonivestluse algusrituaalide rühma dialoogiaktide märgendid, mis kajastaksid algusrituaale teisetüübilistes vestlustes, kui seda on kõned infotelefonile (vrd Rääbis 2006). Tartu Ülikooli suulise kõne korpusesse (Hennoste 2003) seni kuulunud müügivestlused on infokõnedest erineva struktuuriga (lähemalt vt Vellerind 1997). Müügivestluses tutvustavad mõlemad osapooled kõigepealt ennast ja ametiasutust, mida nad esindavad, vestlused on pikemad ja teemasid võib olla mitu. Kõned võivad olla formaalsemad, kui osalejad räägivad esmakordselt, või meenutada argisuhtlust, kui osalejad tunnevad teineteist hästi.

Infokõnedes on algusrituaal lühike ja ebasümmeetriline – helistaja jääb anonüümseks, tal ei ole vaja infopäringu tegemiseks ennast

[1] Artikli valmimist on toetanud ETF (grant 5685) ja HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

tutvustada, ennast tutvustab vaid infotöötaja. Teistsugune on olukord nt juhul, kui helistaja soovib rääkida konkreetse isikuga, aga kõne võtab vastu keegi teine. Klassikaline näide on kõne direktorile, mille võtab vastu sekretär. Seesugune suhtlus saab toimuda reeglina ainult läbi vahendaja (sekretär, juhiabi, keskjaama töötaja vmt). Müügivestluste hulgas on nii kõnesid, mis algavad sekretäriga suhtlemise sekventsist, kui ka selliseid, kus kõne võtab vastu soovitud isik ise.

Müügivestlused kuuluvad müügiläbirääkimiste alguse etappi, mis tähendab, et nende eesmärk on info kogumine kliendi kohta, pakutavate kursuste tutvustamine, ostusoovi korral ka argumenteerimine või kui see ei ole enam vajalik, siis ka ostuotsusele suunamine. Müüjad püüavad luua sundimatu õhkkonna ja seetõttu on müügivestlused pikemad ning sarnanevad kohati argivestlustega, esitades märgendajale mitmeid väljakutseid.[2]

2. Müügivestluste alustamine

Müügivestluste alustamiseks on põhimõtteliselt kaks võimalust:

1. telefonikõne võtab vastu soovitud isik. Kui kõne võtab vastu soovitud isik, on võimalusi kaks:

1.1. soovitud isik saab rääkida;

1.2. soovitud isik ei saa rääkida.

2. telefonikõne võtab vastu vahendaja (või kolleeg, st keegi teine). Kui kõne võtab vastu vahendaja, on kolm võimalust:

2.1. soovitud isik on kohal ja saab rääkida;

2.2. soovitud isik on kohal, aga rääkida ei saa (müügivestlustes ei esinenud);

2.3. soovitud isikut pole

a. kõne toimumise ajal majas,

b. enam (pikemat aega) tööl.

Helistaja võib vormistada oma rääkimissoovi kahel viisil:

a. küsimuse vormis (kas ma X-ga saan rääkida),

b. direktiivi vormis (palun X / soovin rääkida X-ga).[3]

---

[2] Kõik nimed on suhtlejate anonüümsuse huvides muudetud. M tähistab müüjat, S sekretäri, K1 vahendajat, K klienti. Näidetes on lausungid nummerdatud nende lausumise järjekorras.

[3] Dialoogikorpuse aktide tüpoloogiasse on müügivestluste algusrituaale arvestades juurde lisatud järgmised aktid:

Direktiividena käsitleb EDiT naaberpaariakte *soov, ettepanek, pakkumine* (Hennoste, Rääbis 2004: 72). Juhul kui rääkimissoov on vormistatud direktiivina, saab dialoogiakt ühe märgendi RIE: SOOV RÄÄKIDA, kui küsimuse vormis, siis kaks märgendit. Üks märgend kajastab küsimuse tüüpi, nt KYE: SULETUD KAS, teine on rituaali märgend RIE: SOOV RÄÄKIDA. Sama loogika kehtib ka parandussekventside ja vastuse tingimuste täpsustamise märgenduse puhul, kus dialoogiaktid saavad topeltmärgendid. Direktiivi vormis esitati 44 müügivestluses rääkimissoovi 21 korral, küsimuse vormis 9 korral.

Lisaks rituaalile *soov rääkida*, tuli lisada tüpoloogiasse ka rituaalsed aktid RIE: IDENTIFITSEERIMINE (esiliige) ja RIJ: IDENTIFITSEERIMINE (järelliige). Müügivestlustes esines sekventse, kus vestluspartnerit identifitseeritakse näiteks väitlausena vormistatud küsimusega. Selleks, et vältida segadusi kõnede analüüsimisel, märgendame sellisel juhul identifitseerimissekventsi topeltmärgenditega (nagu ka parandused ja vastuse tingimuste täpsustamised). Identifitseerimissekventse esines 8 korral.

Tüpoloogiasse tuli müügivestluste struktuuri arvestades tagasi võtta dialoogiaktid RIE: VIISAKUSKÜSIMUS (esiliige) ja RIJ: VIISAKUSVASTUS (järelliige). Viisakusküsimusi (ingl k *small-talk*, nt *Kuidas teil vahepeal on läinud? – Noh pole viga, oleme elus veel.*) Teisetüübilistes vestlustes esineb viisakusküsimusi üliharva, mistõttu jäeti need dialoogiaktid algul tüpoloogiast välja. Müügivestlustes, kus müüjad soovivad saavutada sundimatut õhkkonda, võivad esineda vestluse alguses viisakusküsimused. Eristamaks neid infoküsimustest, märgendame ka viisakusküsimused topeltmärgenditega, kus esimene märgend tähistab küsimuse tüüpi ja teine viitab selle küsimuse rituaalsele iseloomule. Viisakusküsimusi esitati 8 korral.

Teatud kindel muster institutsionaalse suhtluse läbiviimisel loob ühe konkreetse suhtlustüübi sõrmejälje (Drew, Heritage 1992: 26). Müügivestlust aitavad muuta iseloomulikuks suhtlustüübiks ja loovad tema sõrmejälje: identifitseerimine, viisakusküsimused, tea-

---

RIE: SOOV RÄÄKIDA

a) kui kõnele vastab soovitud isik, siis märgendame RIJ: NÕUSTUMINE / MITTENÕUSTUMINE

b) kui kõnele vastab vahendaja, siis RIJ: SUUNAMINE / MITTESUUNAMINE

tud kindla struktuuriga tutvustussekventsid, mida kasutatakse teatud kindlas järjestuses ja öeldakse teatud kindla intonatsiooniga (mis ei ole käesoleva artikli teema) ja rääkimissoovi esitamine *kas*-küsimuse või direktiivi vormis.

Järgnevalt (uute) märgendite selgitused näidete varal.

3. Müügivestluste algusrituaalid

3.1. Kõne võtab vastu soovitud isik

Kui kõne võtab vastu soovitud isik, on kaks põhimõttelist võimalust: soovitud isik saab rääkida või soovitud isik ei saa rääkida.

3.1.1. Soovitud isik saab rääkida. Rääkimissoov esitatakse müügivestlustes kahel viisil: a. küsimusena, b. direktiivina.

a. Rääkimissoov on esitatud küsimuse vormis

(1)
1. K: Nirk Mahtras. | RIJ: KUTSUNGI VASTUVÕTMINE || RY: TUTVUSTUS |
2. (0.8) ((telefonitoru tõstmise müra))
3. M: tere päevast. | RIE: TERVITUS |
4. K: tervist. | RIJ: VASTUTERVITUS |
5. M: siin on Tiit=Tukat=Tiritammest. | RY: TUTVUSTUS |
6. K: tere=`tere. | RIJ: VASTUTERVITUS |
7. M: kas on hetk aega rääkida. | KYE: SULETUD KAS || RIE: SOOV RÄÄKIDA |
8. K: on ikka. | KYJ: JAH | | RIJ: NÕUSTUMINE |

Kõne alguses (rida 1) võtab klient kõne vastu ja tutvustab ennast. Seejärel tuleb tervituste sekvents (read 3–4), siis tutvustab ennast müüja ja klient tervitab teda uuesti real 6. Müüja esitab oma rääkimissoovi küsimuse vormis ja saab jaatava vastuse. Kuivõrd soov rääkida on esitatud küsimusena, tuleb sellele dialoogiaktile panna topeltmärgendid. Samuti saab topeltmärgendid kliendi vastus real 8.

b. Rääkimissoov on esitatud direktiivi vormis

Müüja tutvustab ennast (näites 2, rida 6) ja esitab rääkimissoovi direktiivi vormis (rida 8). Klient nõustub müüjaga vestlema vastates selle peale *no ta kuuleb*. Direktiivi vormis esitatud soov rääkida märgendatakse ühe märgendiga – RIE: SOOV RÄÄKIDA.

(2)
6. M: olen Agu Rein Tiritamme koolits`firmast | RY: TUTVUSTUS |
7. K: jah | VR: NEUTRAALNE JÄTKAJA |
8. M: ja sooviks rääkida Tauno Laariga | RIE: SOOV RÄÄKIDA |
9. K: no ta kuuleb | RIJ: NÕUSTUMINE |

3.1.2. Soovitud isik võtab kõne vastu, aga rääkida ei saa. Telefonisuhtluse eripära on see, et puudub võimalus saada visuaalset informatsiooni (Schegloff 1986:5; Luke, Pavlidou 2002: 5). Kogu informatsioon tuleb edasi anda verbaalselt ja võimalikult otstarbekalt. Teatud situatsioonides, kus näiteks kõne vastuvõtjal on väga kiire, on info edastamise otstarbekuse huvides vaja rikkuda rituaalide järgnevuse skeemi ja algatada telefonikõne lõpetamine juba kõne alguses. Vaielda võib siinkohal selle üle, kas antud situatsioonis tuleks märgendada müüja enesetutvustuse voor (näide 3, rida 1) ka rääkimissoovina, nagu klient seda tõlgendab. Selle otsuse vastu räägib asjaolu, et normaalses olukorras, kus vastajal oleks võimalik rituaalide loomulikku järjestust välja pidada, järgneks tutvustusele ka verbaalselt vormistatud soov rääkida.

(3)
1. M: Einar=Mattias=ja=`Tiritamm=siinpol. | RY: TUTVUSTUS |
2. .hh | YA: MUU |
3. K: > ee `praegu on nagu `paha sin teemat arutada < | RIE: LÕPU SIGNAAL |
4. M: a`haa | VR: NEUTRAALNE INFO OSUTAMINE UUEKS |
5. K: > võtame kuskil esmaspäev=`teisipäev < | DIE: ETTEPANEK |

Näites 4 on kliendi vastus vasturääkiv – ta järgib rituaali, vastab soovile rääkida nõustumiseks kasutatava vormeliga (rida 7) ja ütleb seejärel, et ta ei saa hetkel edasi vestelda (rida 8).

(4)
1. ((kutsung)) | RIE: KUTSUNG |
2. ((maki urin))
3. K: Eras. | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
4. =tere=päevast. | RIE: TERVITUS |
5. M: e tere `päevast. | RIJ: VASTUTERVITUS |
6. soovin rääkida härra Ando Simmiga. | RIE: SOOV RÄÄKIDA |
7. K: `j:aa ma `kuulen. | RIJ: NÕUSTUMINE |
8. (0.8) aga $ ei ole moment aega `rääkida=küll $ | RIJ: MITTENÕUSTUMINE |

3.2. Kõne võtab vastu vahendaja

Kui kõne võtab vastu vahendaja, on põhimõttelisi võimalusi kaks: soovitud isik on majas või ei ole. Kui ta on majas, on omakorda kaks võimalust: ta saab või ei saa rääkida. Viimast varianti meie andmebaasis ei esinenud. Kui soovitud isikut majas ei ole, siis võib ta olla ajutiselt ära, või pole enam (pikemat aega) tööl.

3.2.1. Soovitud isik on kohal ja saab rääkida. Kui telefonikõne võtab vastu vahendaja, peab helistaja ennast tutvustama kõigepealt vahendajale ja esitama oma rääkimissoovi talle. Seejärel tegutseb vahendaja vastavalt olukorrale – kas soov on põhjendatud ja kõne on edasisuunamist väärt ning kas soovitud isik on majas või mitte.

Märgendatud müügivestlustes ei esine olukorda, kus sekretär leiaks, et soov rääkida ei oleks põhjendatud. Kõne suunatakse alati edasi, kui soovitud isik on kohal. Näites 5 esitab müüja oma soovi rääkida direktiivi vormis ja sekretäri vastusvoor koosneb kahest dialoogiaktist: vastuvõtuteade on vormistatud partikliga *jaa*, millele järgneb rituaalne vormel *üks hetk*, millega sekretär annab märku, et ta hakkab kõnet soovitud isikule edasi suunama. (See hetk võib kesta üsna kaua, näites 5 on see 6.2 sekundit, pikim suunamise aeg on senistest märgendatud kõnedes 32 sekundit)

Müügivestlustes esineb edasisuunamisi, mida on kuulda ootemuusikana, ja selliseid, mida ei täida muusika, vaid vaikus. Olenemata sellest, kas muusikat on kuulda või mitte, pannakse dialoogi märgendamisel suunamisele järgnevale pausile märgend RIE: KUTSUNG, sest kõne suunamine on ühtlasi kutsungiks soovitud isikule, kes selle peale telefonitoru tõstab ja kõne sekretärilt vastu võtab.

(5)
1. S: Autokere. | RY: TUTVUSTUS |
2. (.)
3. M: .h tere. | RIE: TERVITUS |
4. S: tervist? | RIJ: VASTUTERVITUS |
5. M: Rein=Kaltenberg=ja=`Tiritamm on siinpool, | RY: TUTVUSTUS |
6. ja: ma soovin rääkida `Rauno Rätik palun. | RIE: SOOV RÄÄKIDA |
7. S: jaa | VR: NEUTRAALNE VASTUVÕTUTEADE |
8. üks hetk. | RIJ: SUUNAMINE |
9. M: aitäh? | VR: MUU |
10. (6.2) ((ootemuusika)) | RIE: KUTSUNG |
11. K: jaa Rauno kuuleb. | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |

Sekretäri vastusele *üks hetk* võib järgneda müüjapoolne reaktsioon (näide 5 rida 9), aga ei pruugi. Kuivõrd kõne suunamine või mittesuunamine ei olene helistaja reaktsioonist, otsustasime märgendada helistaja vooru vabatahtliku reaktsioonina VR: MUU. Tüpoloogia varasema versiooni järgi tulnuks märgendada sekretäri vastus *üks hetk* edasilükkamisena, kuid sekretäri vastus on erinev infotelefoni infotöötaja vastusest *üks hetk*. Infotöötaja palub helistajal oodata, kuni ta soovitud numbri andmebaasist leiab, sekretär aga annab oma vastu-

sega teada, et ta suunab kõne edasi soovitud isikule, seetõttu otsustasime luua uue rituaalse dialoogiakti RIE: SUUNAMINE.

Näide 6 iseloomustab algusrituaalide kaotsiminekut tüpoloogia varasema versiooni märgendite hulka eriti hästi. Real 4 esitab müüja soovi vestelda isikuga (Hart Tort). Soov on esitatud *kas*-küsimusena ja oleks tulnud märgendada suletud *kas*-küsimuseks, kuna on sellisena sõnastatud ja vastus on jaatav. Siiski on see eestikeelses suhtluses lisaks direktiiviga vormistatud soovile rääkida teine võimalus vormistada oma rääkimissoovi. Seetõttu kasutame topeltmärgendeid, nagu ka paranduste ja vastuse tingimuste täpsustamiste puhul, kus märgendatakse nii küsimuse liik kui ka funktsioon, mida see küsimus antud sekventsis täidab.

Sekretäri vastus real 6 on adresseeritud küsimuse vormile *on `küll jah.* Real 7 vastab sekretär soovi sisule – vestelda Hart Tordiga.

(6)
1. M: Asta Topelt siinpol `Tiritammest. | RY: TUTVUSTUS |
2. (0.8) koolitus ja konsultatsiooni`firmast. .h | IL: TÄPSUSTAMINE |
3. S: * jah. * | VR: NEUTRAALNE VASTUVÕTUTEADE |
4. M: kas on Hart Tort seal. | KYE: SULETUD KAS | | RIE: SOOV RÄÄKIDA |
5. (0.8)
6. S: on `küll jah. | KYJ: JAH |
7. (.) kohe annan. | RIJ: SUUNAMINE |

Suhtluses vahendajaga võib ette tulla takistusi. Järgmises näites algab suhtlus sujuvalt, helistaja teab, et kõne võtab vastu sekretär. Müüja tutvustab ennast real 5 ja esitab real 7 oma soovi. Sekretär suunab kõne edasi öeldes *üks hetk* ja kõne suunamise ajal hakkab mängima muusika. Suunamist tuleb oodata 18 sekundit. Seejärel võtab vastu soovitud isiku asemel teine sekretär. Talle on informatsioon helistaja ja tema soovi kohta edasi öeldud, teine sekretär ei esita enam küsimusi (nagu võib ka juhtuda), ütleb real 11 *ühendan teid Virtin Mulderiga kohe.* Müüja vastuse *suur tänu* märgendamine tänamiseks läheks vastuollu tüpoloogia loogikaga, mille järgi tänamine on naaberpaariakt, esinedes olenevalt asukohast vestluses kas esi- või järelliikmena. Praeguse otsuse järgi märgendame helistaja reaktsiooni suunamisele vabatahtlikuks reaktsiooniks. Põhjusi selle otsuse tegemiseks oli mitu. Esiteks sellepärast, et vastuseks suunamisele võivad suhtlejad öelda väga erinevaid lausungeid, mis ei ole tänamised (*palun, jaa, jah, mhmh*), teiseks võib reaktsioon tulemata jääda ja kolmandaks ei oota helistaja sellele suunajapoolset vastu-

reaktsiooni, mis tähendab, et tegemist on helistajapoolse vabatahtliku reaktsiooniga (VR).

(7)
1. ((kutsung)) | RIE: KUTSUNG |
2. S1: Tasuja. | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
3. M: tere päevast. | RIE: TERVITUS |
4. S1: tere, | RIJ: VASTUTERVITUS |
5. M: olen Oona Rahumägi koolitusfirmast [Tiritamm] | RY: TUTVUSTUS |
6. S1: [jaa?] | VR: NEUTRAALNE JÄTKAJA |
7. M: ja Virtin Mulder palun | RIE: SOOV RÄÄKIDA|
8. S1: üks=`hetk. | RIJ: SUUNAMINE |
9. (18.0) ((ootemuusika)) | RIE: KUTSUNG |
10. S2: tere päevast | RIE: TERVITUS |
11. ühendan teid Virtin Mulderiga kohe | RIE: SUUNAMINE |
12. M: jaa suur tänu | VR: MUU |

Suunamiseks märgendame ka sekretäri sellise vastuse, mis koosneb ainult partiklist *jah*, nagu näiteks järgmises näitelõigus.

(8)
1. ((kutsung)) | RIE: KUTSUNG |
2. S: tere `päevast. | RIJ: KUTSUNGI VASTUVÕTMINE | | RIE: TERVITUS |
3. M: ee tere päevast. | RIJ: VASTUTERVITUS |
4. Elar Vigur Tiritammest siin | RY: TUTVUSTUS |
5. ma soovin rääk- rääkida härra Naan Kaudsepaga. | RIE: SOOV RÄÄKIDA |
6. S: * jah. * | RIJ: SUUNAMINE |
7. (32.0) ((ootemuusika)) | RIE: KUTSUNG |

Kõne võib vastu võtta ka alluv, nagu järgmises näites, kus kaupluse juhatajale mõeldud kõne võtab vastu sama kaupluse müüja ja annab telefonitoru edasi lähedal olevale juhatajale.

(9)
1. ((kutsung)) | RIE: KUTSUNG |
2. K1: tere päevast. | RIJ: KUTSUNGI VASTUVÕTMINE | | RIE: TERVITUS |
3. kauplus Ruloo kuuleb. | RY: TUTVUSTUS |
4. M: tere päevast. | RIJ: VASTUTERVITUS |
5. (.) Oona Rahumägi `Tiritammest on siinpool. | RY: TUTVUSTUS |
6. (0.2) ja: härra `Sten `Kalmer=palun. | RIE: SOOV RÄÄKIDA |
7. K1: jaa= | VR: NEUTRAALNE VASTUVÕTUTEADE |
8. üks=hetk? | RIJ: SUUNAMINE |
9. ((ütleb kõrvale)) * Sten (.) saad rääkida * | KYE: SULETUD KAS | | RIE: KUTSUNG |
10. K: allo. | KYJ: MUU | | RIJ: KUTSUNGI VASTUVÕTMINE |

Helistaja voor (algus real 4) koosneb kolmest üksteisele järgnevast aktist: tervitusest (rida 4), tutvustusest (rida 5) ja direktiivina esitatud soovist rääkida Sten Kalmeriga (rida 6). Kaupluse müüja, kes kõne

vastu võtab (K1), võtab soovi vastu partikliga *jaa*, mis saab märgendi VR: NEUTRAALNE VASTUVÕTUTEADE, mille järel ütleb *üks hetk*.

Kaupluse müüja küsib seejärel lähedal olevalt Sten Kalmerilt *Sten (.) saad rääkida*. Tema pöördumine on vormistatud küsimusena, pöördumise eesmärgiks on kõne edasi andmine kaupluse juhatajale. Kuigi kõne suunamist ei toimu, on telefonitoru ulatamine samal ajal ka kutsungiks kaupluse juhatajale, mistõttu tuleb tema voorule panna topeltmärgendid – nii küsimuse kui ka kutsungi märgendid. Kaupluse juhataja võtab kõne vastu. Selleks, et esitatud küsimus saaks vastuse, märgendame juhataja vooru (K, rida 10) nii vastusena küsimusele kui ka kutsungi vastuvõtmisena.

Põhjus, miks selles näites partikkel *jaa* (rida 7) saab teistsuguse märgendi kui näites 9, on järgmine. Infokõnedes vastavad infoametnikud helistaja soovi vastu võttes vastusekuju *jah/jaa üks hetk*, kus partikkel *jah/jaa* kannab signaali, et infotöötaja on helistaja soovist aru saanud ja vastuse teine osa *üks hetk* annab märku, et ta asub andmebaasist vastust otsima. Kuivõrd info vastuvõtu signaal partikliga *jah* ei ole infosoovi esitamise järel naabruspaari järelliige ega oota enda järel ka helistaja reaktsiooni, märgendame selle vabatahtliku reaktsioonina VR: NEUTRAALNE VASTUVÕTU-TEADE. Samakujulise sekretäri vastuse märgendamisel kasutame analoogilist skeemi – partikkel *jah* märgendatakse kui vabatahtlik reaktsioon ja vastuse teine osa *üks hetk* kui suunamine.

Näites 8 sekretär muud ei vastagi kui ainult *jah* (näide 8 rida 6), seetõttu tuleb sellise vastuse märgendamisel skeemi muuta ja valida kahe võimaluse vahel olulisem ning märgendada see. Käesoleval juhul on tähtsam, et rituaalne naabruspaar *soov rääkida* saaks järelliikme *suunamine* kui välja tuua võimalus, et sekretär kasutab partiklit *jah* vabatahtliku reaktsioonina. Seetõttu märgendame sellises positsioonis sekretäri (vahendaja) vastuse partikliga *jah/jaa* rituaalse direktiivi järelliikmena RIE: SUUNAMINE.

3.2.2. Soovitud isikut pole kõne toimumise ajal majas. Näites 10 helistab müüja firmasse ja soovib rääkida Madis Paljasega. Müüja soov on vormistatud küsimusena (rida 6). Kõne võtab vastu vahendaja ja vastab küsimusele *ee ei:=ole kahjuks* (rida 8). Kui märgendada need kaks naaberpaari ainult küsimuseks ja vastuseks, läheks kaduma nende dialoogiaktide tegelik sisu – rituaalne küsimuse vor-

mis esitatud soov rääkida ja vastus, mille sisu on, et vahendaja ei saa soovi täita, kuna soovitud isik ei viibi kohal. Vahendaja vastuse märgendame kõne mittesuunamiseks.

(10)
1. ((kutsung)) | RIE: KUTSUNG |
2. K: {-} | YA: PRAAK |
3. tere päevast. | RIE: TERVITUS |
4. (.)
5. M: tere | RIJ: VASTUTERVITUS |
6. (0.2) kas: Madis Paljas `ka on. | KYE: SULETUD KAS | | RIE: SOOV RÄÄKIDA |
7. (0.2)
8. K: ee ei:=ole kahjuks. | KYJ: EI | | RIJ: MITTESUUNAMINE |

Teine võimalus on, et soovitud isikut pole tööl olnud pikemat aega. Sellesse gruppi võiksid kuuluda ka kõned, kus helistaja ei tea, et soovitud isik on töölt lahkunud, üheski uuritud 44 müügivestluses sellist olukorda ette ei tulnud. Järgnev näitelõik (näide 11) on järjekorras teine kõne samasse firmasse, soovitud isik on pikemat aega olnud haige. Müüja esitab oma soovi rääkida Siret Nukiga direktiivi vormis (rida 6), vahendaja ei saa tema soovi täita, kuna soovitud isikut ei ole tööl. Vahendaja vastusvoor (rida 8) märgendatakse mittesuunamisena.

(11)
1. ((kutsung)) | RIE: KUTSUNG |
2. S: Videokandja=Eesti=Maarika=Täht=kuuleb= | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
3. tere | RIE: TERVITUS |
4. M: `te:re päevast. | RIJ: VASTUTERVITUS |
5. Elar=Vigur=`Tiritammest=sin= | RY: TUTVUSTUS |
6. ma=soovin=`rääkida `Siret Nukiga. | RIE: SOOV RÄÄKIDA |
7. (.)
8. S: Siret Nukk on `haige. | RIJ: MITTESUUNAMINE |

4. Identifitseerimine

Müügivestluste algused on üsna erinevad, vastajad kasutavad mitmesuguseid lausungeid kõne vastuvõtmiseks: *hallo* (8 korda), asutuse/firma vms nimi (6 korda), *tere (päevast)* (5 korda), enda nimi (4 korda), asutuse/firma nimi + enda nimi (3 korda), *ja(a)/jah* (3 korda), *kuulen/kuuleb* (2 korda). Lindistatud kõned pärinevad aastatest 1987/1988, kus telefonikõnede vastuvõtmise rituaalide muutused olid alles toimumas, seetõttu esinevad institutsionaalse kõne vastuvõtmisel partiklid *jah/jaa* või *aljo/hallo/alo:u*. Tänapäeval on saanud kom-

beks, et ametikõnede vastuvõtmisel öeldakse vähemalt asutuse/firma nimi, pikema variandi puhul ka vastaja nimi ja tervitus. Alati aga ei piisa sellest, et vastaja ennast tutvustab, helistaja võib ikkagi algatada identifitseerimise.

Järgmises näites võtab soovitud isik ise telefoni vastu, tutvustab ennast, seejärel järgneb tervituste sekvents. Pärast tervitusi on 0,8sekundiline paus. Selle järel esitab helistaja (müüja) real 7 kontrolliva vastustpakkuva küsimuse *ee Maia Sireda kuuleb*. Müüja küsimuse märgendame vastustpakkuvaks, kuna müüja eeldab, et see on just nimelt Maia Sireda, kes kõne vastu võttis. Lisaks küsimus–vastus sekventsi märgenditele vajab antud lõik esile tõstmist ka kui identifitseerimissekvents, seetõttu kasutame ka siin topeltmärgendeid, nagu ka parandussekventside, vastuse tingimuste täpsustamise ja küsimuse vormis esitatud rääkimissoovi puhul.

(12)
1. ((kutsung)) | RIE: KUTSUNG |
2. K: Aktsiisimaik=Maia kuuleb. | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
3. (0.5) ((telefonitoru tõstmise heli))
4. M: tere `päevast. | RIE: TERVITUS |
5. K: tere? | RIJ: VASTUTERVITUS |
6. (0.5)
7. M: ee Maia `Sireda kuuleb. | KYE: VASTUST PAKKUV | | RIE: IDENTIFITSEERIMINE |
8. K: jaa kuulen. | KYJ: JAH | | RIJ: IDENTIFITSEERIMINE |
9. (0.2)

5. Viisakusküsimused

Viisakusküsimustega soovib helistaja luua sundimatut ja meeldivat õhkkonda. Ühtlasi toimib esitatud küsimus sillana eelmise vestluse juurde. Müüja näitab, et ta mäletab eelmise kõne sisu ja väljendab huvi kliendi tegemiste vastu.

(13)
9. M: mt (.) kuidas on elu `vahepeal läinud, kõik kenad `reisid on `seljataha jäänud. | KYE: VASTUST PAKKUV | | RIE: VIISAKUSKÜSIMUS |
10. K: jajaa. | KYJ: JAH | | RIJ: VIISAKUSVASTUS |
11. seda küll. | KYJ: JAH | | RIJ: VIISAKUSVASTUS |
12. (0.5)
13. M: mt kuidas `oli (...) `Ameerika. | KYE: AVATUD | | RIE: VIISAKUSKÜSIMUS |
14. K: tore. | KYJ: INFO ANDMINE | | RIJ: VIISAKUSVASTUS |
15. M: mt talvel. | KYE: VASTUST PAKKUV | | RIE: VIISAKUSKÜSIMUS |
16. K: jah | KYJ: JAH | | RIJ: VIISAKUSVASTUS |

Alati ei pruugi viisakusküsimus kavandatud eesmärki teenida. Näites 14 ei hakka klient viisakusküsimusele vastama, jätab selle tähelepanuta (rida 12) ja esitab omakorda küsimuse müüjale (rida 13).

(14)
1. K: kuulen. | RIJ: KUTSUNGI VASTUVÕTMINE |
2. M: e tere `päevast. | RIE: TERVITUS |
3. K: tere. | RIJ: VASTUTERVITUS |
4. M: Einar=Mattias=ja=`Tiritamm=siinpool. | RY: TUTVUSTUS |
5. (0.8)
6. K: jah? | VR: NEUTRAALNE JÄTKAJA |
7. M: e `saate praegu `vestelda. | KYE: VASTUST PAKKUV | |RIE: SOOV RÄÄKIDA |
8. K: ee jaa? (.) ma `kuulen teid? | KYJ: JAH | | RIJ: NÕUSTUMINE |
9. M: j:aa. | VR: NEUTRAALNE VASTUVÕTUTEADE |
10. me `viimane kord me rääkisime sin `juunikuus=ja. | SEE: VÄIDE |
11. .h ja `vahepeal on nüd ` aasta `vahetunud=ja. | SEE: VÄIDE |
12. (1.2) ja tunnen nüd `huvi et kuidas on teil `läinud vahepeal, | KYE: AVATUD | | RIE: VIISAKUSKÜSIMUS |
13. K: meie `koolituse vastu jah? | KYE: VASTUST PAKKUV | | VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |

Müügivestlustes esineb veel üks tüüp rituaalseid tegevusi. Need on viited eelmisele kõnele, mille eesmärgiks on meenutada kliendile eelmist vestlust ja selle sisu. Võimalik, et eelmisel korral lepiti milleski kokku, sellisel juhul on viitega hea põhjendada uut helistamist. Praeguste otsuste järgi ei saa seesugused viited topeltmärgendeid, piirdutakse ainult ühe märgendiga SEE: VÄIDE (vastusvooruks võib olla kas SEE: NÕUSTUMINE või SEE: MITTENÕUSTUMINE). Näites 15 (rida 8) viitab müüja eelmisele kõnele, mis tema sõnul toimus detsembrikuus, klient nõustub müüjaga real 10. Real 11 meenutab müüja kliendile oma lubadust saata pakutavate kursuste kataloog, millega klient real 12 uuesti nõustub.

(15)
1. K: aljo. | RIJ: KUTSUNGI VASTUVÕTMINE |
2. M: et (.) t- tere `päevast. | RIE: TERVITUS |
3. K: tervist. | RIJ: VASTUTERVITUS |
4. M: mt härra Andrus Voit kuuleb. | KYE: VASTUST PAKKUV | | RIE: IDENTIFITSEERIMINE |
5. K: jah. | KYJ: JAH | | RIJ: IDENTIFITSEERIMINE |
6. M: mt Einar=Mattias=ja=`Tiritamm=siinpol. | RY: TUTVUSTUS |
7. K: jah. | VR: NEUTRAALNE VASTUVÕTUTEADE |
8. M: mt vestlesime teiega:: (...) mt tetsembri lõpul, mt | SEE: VÄIDE |
9. (.)
10. K: täiesti `õige. | SEJ: NÕUSTUMINE |
11. M: jaa: (.) lubasin teile saata ka:: (.) m:aterjalid, mt | SEE: VÄIDE |
12. K: jaa. | SEJ: NÕUSTUMINE |

13. tänan `väga. | RIE: TÄNAN |
14. (.) asi on nagu (.) selgeks (.) selgeks tehtud. | YA: INFO ANDMINE |

6. Kokkuvõtteks

Müügivestluste kui infokõnedest erineva struktuuriga telefonikõnede märgendamisel ilmnes vajadus täiendada olemasolevat dialoogiaktide märgendamise tüpoloogiat EDiT uute dialoogiaktidega, mis kajastaksid paremini müügivestluse struktuuri ja nendes kasutatavaid rituaale.

Seal, kus võimalik, püüti topeltmärgendite lisamist vältida, näiteks viidete puhul eelmisele kõnele, kus piirdutakse ühe märgendiga. Soov rääkida / kõne suunamine, identifitseerimised ja viisakusküsimused/viisakusvastused märgendatakse topeltmärgenditega – esimene märgend kajastab küsimuse/vastuse või direktiivi/direktiivi järelliikme tüüpi ja teine märgend rituaali, mida selle küsimuse või direktiiviga sooritatakse. Direktiivi vormis esitati 44 müügivestluses rääkimissoovi 21 korral, küsimuse vormis 9 korral. Viisakusküsimusi esitati 8 korral, identifitseerimissekventse esines samuti 8 korral.

Kirjandus

Drew, Paul, Heritage, John 1992. Talk at Work. Cambridge: Cambridge University Press.

Hennoste, Tiit 2003. Suulise eesti keele uurimine: korpus. – Keel ja Kirjandus 7, 481–500.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu Ülikooli arvutiteaduse instituut. Tartu: Tartu Ülikooli kirjastus.

Sacks, Harvey, Schegloff, Emanuel A., Jefferson Gail 1974. A simplest systematics for the organization of turn-taking for conversation. – Language 50, 696–735.

Luke, Kang Kwong, Pavlidou, Theodossia-Soula 2002. Telephone calls. Unity and diversity in conversational structure across languages and cultures. Ed by K. K. Luke, T.-S. Pavlidou. Amsterdam: John Benjamins.

Rääbis, Andriela 2006. Infodialoogide algusrituaalid. (Käesolevas kogumikus.)

Vellerind, Riina 1997. Suulise vestluse struktuur telefoni teel peetud müügiläbirääkimiste näitel. Bakalaureusetöö. Käsikiri Tartu Ülikooli eesti keele õppetoolis.

Schegloff, Emanuel A. 1986. The routine as achievement. – Human Studies 9, 111–151.

Kõneleja reaktsioon vestluskaaslase parandusalgatusele[1]

Krista Strandson
Tartu Ülikool

1. Sissejuhatus

Kõneleja võib mitmesugustel põhjustel teha keelelisi otsuseid, mida ta kas ise või tema kaasvestleja(d) tagantjärgi õigeks või sobivaks ei pea. Seega tuleb juba välja öeldud lausungid ümber teha, neid täpsustada, täiendada vms. Konversatsioonianalüüsi termineis on vahendeid, mille abil vestlejad lahendavad rääkimisel, kõne kuulamisel ja sellest arusaa-misel tekkinud probleeme, nimetatud parandusmehhanismiks (*repair organization*) (Schegloff, Jefferson, Sacks 1977). Probleemid ei pruugi alati lähtuda otsesest grammatilisest või semantilisest veast, piisab sellest, kui kõneleja ise või tema kaaslane leiab, et öeldus on midagi valesti, ebasobiv, kinnitamist ja/või täpsustamist vajavat (Schegloff 1979). Kõige üldisemalt võib eristada kõneleja enese ja kaaslase algatatud parandusi. Esimesel juhul muudab, täiendab, täpsustab või parandab teksti väljaütleja oma teksti ise, omal algatusel, tavaliselt samas kõnevoorus, kus probleem ilmnes. Vestluskaaslase parandusalgatuse põhiline ülesanne on tähistada probleemallika asukoht, et anda kõnelejale, s.o probleemvooru väljaütlejale, uus võimalus ise parandus lõpetada. Seega võtavad teiste algatatud parandused vähemalt kaks vooru: esimeses neist määrab kaaslane probleem-allika asukoha ning teises viib kõneleja parduse lõpuni.

Artikli eesmärk on vaadelda kõneleja reaktsiooni kaaslase parandusalgatustele: kas ja mil määral on kõneleja reaktsioon seotud kaaslase parandusalgatuse tüübiga. Analüüsi aluseks on 50 telefonikõnet infoliinile ja 50 argist telefonikõnet. Argilindistused on pärit Tartu Ülikooli Eesti suulise keele korpusest, infokõned Tartu Ülikooli Eesti dialoogikorpusest. Praktilises analüüsis on kasutatud konversatsioonianalüüsi meetodeid.

[1] Artikli valmimist on toetanud ETF (grant 5685) ja HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

2. Parandusalgatuse vahendeid: keeleline vorm

Kaaslase algatatud ja kõneleja läbiviidud paranduste tüpoloogia on eri autoritel erinev (vt Hennoste 2001; Hennoste, Rääbis 2004; Schegloff, Jefferson, Sacks 1977: 361; Tainio (toim) 1997). Käesolevas artiklis on parandusalgatusi liigitades lähtutud nende keelelisest vormist ja asjaolust, mil määral konktreetne parandusalgatuse tüüp raskusi tekitanud kohta lokaliseerib. Vormist lähtuvalt saab eristada järgmisi parandusalgatusi.

1. Avatud tüüpi parandusalgatus laieneb tervele probleemvoorule, raskuskohta täpsemalt lokaliseerimata. Selline algatus võib osutada kaaslase kuulmishäirele või mõistmisraskusele. Avatud tüüpi parandusalgatused on näiteks küsipartiklid *ah*, *mh*, tõusva intonatsiooniga hääldatud *jah*; küsisõnad *kuidas*, *mida*, *misasja*.

2. Raskuskohta lokaliseeriva küsimuse vorm ja semantika võimaldavad probleemallika asukohta täpsemalt määrata. Siia rühma kuuluvad üksikud küsisõnad (*kes, kus*), pikemad küsimused, kombinatsioon *mis* + probleemallika kordus ja lünkkordus. Lünkkorduse puhul kordab kaaslane osa probleemvoorust ja jätab raskuskoha välja ütlemata.

3. Probleemallika kordus on algatus, mille puhul kuulaja kordab täpselt või väikeste modifikatsioonidega mõnda kaaslase lausungit, fraasi või sõna, et saada kinnitust selle kohta, et see oli just selline. Kordusele võivad liituda küsipartiklid *või, vä, jah* (*seitse jah*).

4. Tõlgendusettepanek (*ymmärrystarjous, candidate understanding*) on kuulaja interpretatsioon eelnenud vooru kohta, kui ta ei ole kindel, et ta seda õigesti mõistis (Sorjonen 1997). Vormilt on tõlgendusettepanekud sõna, fraas või lausung, millele võib olla liitunud partikkel *või, võ, vä, jah*; samuti võib kaaslane kasutada tõlgenduse esitamiseks kombinatsiooni sa/te mõtled/mõtlete + tõlgendus *(te mõtlete raadio Elmar*).

3. Kõneleja reaktsioon kaaslase parandusalgatusele

Järgnevalt on kirjeldatud, kuidas kõneleja probleemi lahendab ning mil määral sõltub kolmanda vooru eneseparanduse vormistus vastavast kaaslase kasutatud parandusalgatuse tüübist. Parandusalgatused on järgnevas analüüsis jaotatud kolmeks suuremaks rühmaks vastavalt sellele, mil määral ja kuidas nad raskuskoha lokaliseerivad.

1. Algatused, mis laienevad tervele probleemvoorule.
2. Algatused, mis lokaliseerivad raskuskoha asukoha.
3. Algatused, kus vestluskaaslane ütleb ise probleemallika välja: raskuskoha kordused ja kaaslase pakutud tõlgendused.

3.1. Kõneleja reaktsioon avatud tüüpi parandusalgatusele

3.1.1. Kõneleja kordab eelpool öeldut. Võimalikke kõneleja reaktsioone kaaslase parandusalgatusele, mis laieneb tervele probleemvoorule, on oma eelmise vooru kordus. Kõneleja kordab varem öeldut eelkõige siis, kui ta arvab või näitab end arvavat, et kaaslase parandusalgatuse põhjuseks oli kuulmishäire. Vormiliselt saab kordused jagada kahte suuremasse rühma: puhtad kordused ja modifitseeritud kordused.

1. Puhta ehk täpse eelmise vooru korduse puhul kordab kõneleja kaaslase parandusalgatuse tulemusel oma eelmist vooru muutmata kujul nagu näites (1).

(1) 237a9
1.→ M: keegi elistas ´ka=vä.
2. V: ah
3. → M: keegi elistas ´ka=vä. (1.0)
4. V: ei?

2. Modifitseeritud eelmise vooru kordustes teeb kõneleja grammatilisi ja/või leksikaalseid muudatusi. Kõneleja võib näiteks lisada oma eelmise vooru mõistmiseks olulise fraasi või vastupidi, lühendada eelmist vooru, korrates vaid kõige olulisemat osa sellest; muuta oma eelmise vooru sõnajärge, aega. Kordusele on lisatud öeldu mõistmiseks oluline fraas näites (2).

(2) 98a3
1. →V: kas seal on `hästi.
2. H: mida?
3.→ V: kas seal on `hästi meie jaoks.
4. H: ma ei tea,

Sõbrannad räägivad parajasti õppejõu soovitatud raamatust. Esimeses reas on V küsimus: *kas seal on `hästi*, millele järgneb H parandusalgatus *mida*. Seepeale kordab V oma küsimust täpselt ning lisab sellele fraasi *meie jaoks* ja saab seejärel küsimusele vastuse.

3.1.2. Probleemvooru muutmine. Mõnikord võib tervele voorule laienev parandusalgatus viia eelmise vooru ümbertegemiseni. Seega

ei täida avatud parandusalgatused vaid üht ülesannet: nad võivad osutada lisaks kaaslase kuulmisprobleemile ka arusaamisraskusele. Võimalik on, et eelmine voor oli kaaslase seisukohalt problemaatiline: eelpool kõnelnu võis öelda midagi ebatavalist, arusaamatut või oli tema voorus mingisugune viga, mis vallandas parandusprotsessi.

(3) 98a1
1. V: nojah ˈimelik. (0.5) .hh noja=et tähendab ˈhalb selline˴ poolik aeg. (.)
2.→ aga kus ta ˈelab sis.
3. H: mida?
4.→ V: e kus te=sis ˈedasi lähete kui=te=
5. H: =ˈEve poole lähme õhtul.

Näide (3) on katke kahe sõbranna vestlusest. Helistaja H peab mõne tunni pärast klassikaaslastega kohtuma ja soovib kohtumisele eelnevat aega V-ga veeta. Tüdrukud rääkisid näitelõigule eelnevas osas klassikaaslastega kohtumise ajast. Näite 1. reas kommenteerib V kellaaega ja jätkab küsimusega *aga kus ta elab sis*. Seepeale algatab H tervele voorule laieneva küsisõnaga *mida* paranduse, mispeale V sõnastab oma küsimuse ümber. V reaktsioon kaaslase parandusalgatusele näitab, et ta ei tõlgendanud H parandusalgatuse põhjust kuulmishäirena, vaid pigem mõistmisprobleemina.

3.1.3. Kõneleja kinnitab eelpool öeldut. Avatud tüüpi parandusalgatuste seas moodustavad omaette rühma partikliga *jah* parandusalgatused. Seda tüüpi algatusi oli vaadeldud materjali hulgas vaid kaks. Mõlemal juhul väljendas *jah* kaaslase imestust, üllatust ja sellele järgnes kõneleja kinnitus. Kõneleja reaktsiooni järgi kuuluvad nad ühte rühma korduste ja tõlgendustega (vt osa 3.3).

(4) 242
1. H: eg=aga mulle ˈtundus=et sa ˈolla: ˈnäinud neid Välkusid üksõhtu
2. ˈkau[em.]
3. V: [aga] (.) no (.) seda
4. ˈküll=aga ega=ma=nüd sis ˈseda ei=ˈrääkind kah,
5. H: jah?
6. →V: jah.

Sõbrannad ei ole kaua kohtunud, üks neist, H, on Tartust Tallinna kolinud. V küsis temalt Tallinnasse kolimise põhjuse kohta ja H väitis seepeale, et V peaks ju seda nende ühiste tuttavate Välkude kaudu teadma. H väite põhjendus jätkub real 1. Ridades 3–4 teatab V, et ta sellest Välkudega ei rääkinud, mispeale algatab H tõusva intonat-

siooniga *jah*-iga paranduse. Parandusalgatusele järgneb V lühike kinnitus (real 6).

3.1.4. Probleemvooru korduste ja muutmiste vahekord avatud parandusalgatuste puhul. Tabelisse 1 on koondatud kõneleja reaktsioon kaaslase avatud tüüpi parandusalgatustele.

Tabel 1. Kõneleja reaktsioon kaaslase avatud tüüpi parandusalgatusele

Kõneleja reaktsioon		Arv
Probleemvooru kordus	Puhas kordus	10
	Modifitseeritud kordus	16
Probleemvooru muutmine		5

Tabel 1 näitab, et kõneleja reageeris 84 % juhtudest (26 korral 31-st) kaaslase avatud tüüpi parandusalgatusele eelmise vooru kordusega. Sellisel juhul osutab ta, et tõlgendas parandusalgatuse põhjusena kuulmishäiret. Viiel korral sõnastas kõneleja oma eelmise vooru ümber, tõlgendades parandusalgatuse põhjuseks arusaamisprobleemi.

3.2. Kõneleja reaktsioon raskuskohta lokaliseerivatele küsimustele

Kaaslase tõlgendusraskuse tõttu välja öeldud küsisõnad ja pikemad küsimused, samuti konstruktsioon küsisõna+eelmise vooru osaline kordus võimaldavad probleemallikat täpsemalt lokaliseerida kui avatud tüüpi parandusalgatused. Siis ei pruugi kõneleja tervet eelmist vooru korrata: piisab, kui formuleerida uuesti kaaslase määratletud raskuskoht.

3.2.1. Kõneleja täpsustab raskuskohta. Sageli on kaaslase tõlgendusraskuste ajendiks kõneleja eelmise vooru ebamäärasus. Nii võib parandusalgatuse põhjustada mõnda pronoomenit sisaldav lausung, kui asesõna täpsem tähendus määratlemata jääb; samuti lausung, kust puudub mõni info vastuvõtu seisukohast oluline osa. Sel juhul asendab kõneleja kolmanda vooru eneseparanduses varem öeldud pronoomeni vastava täistähendusliku sõnaga või lisab puudunud fraasi. Kaaslase seisukohalt lähtudes võib sellised parandusalgatused jagada kaheks. Mõlemal juhul algatas paranduse raskuskohta semantiliselt/grammatiliselt lokaliseeriv küsisõna.

1. Algatused, mille põhjustas ebamäärane viiteseos. Sel juhul asendab kõneleja kaaslasele mõistmisraskusi tekitanud pronoomeni vastava täistähendusliku sõnaga.

(5) 8a1
1. →V: .hh kuidas sul `seal läks.
2. H: kus.
3. →V: .hh `Eve juures.
4. H: aa `Eve juures oli `kihvt

Näites (5) asendab kõneleja V kaaslase küsisõnaga kus parandusalgatusele reageerides asesõna seal täistähendusliku fraasiga (rida 3).

2. Algatused, mille põhjustas elliptilisus. Sel juhul puudub eelmise kõneleja lausungist obligatoorne osa, mis tekitas kaaslasele mõistmisraskusi. Sellisele parandusalgatusele reageerides ütleb kõneleja välja oma eelmisest voorust puudunud elemendi.

(6) 237
1. H: $ no mis sa ´teed. $
2. V: ee kirjutan ´referaati ümber.
3. →H: ega sa=i=´viitsi=ju tegelikult ´präegu.
4. V: mida.
5. →H: kirjutada.
6. V: ma ´pean.

Näite (6) 3. reas on H voor *ega sa=i `viitsi=ju tegelikult `präegu*, mis osutab ilmselt V eelmises voorus sõnastatud infole *kirjutan referaati ümber*. Neljandas reas algatab V aga paranduse küsisõnaga *mida*: küsisõna vorm ja semantika viitavad, et kaaslase tõlgendusraskuse põhjustas eelpool kõnelnu välja ütlemata lausungiosa, mis kuulub mõtteliselt fraasi *ega sa=i `viitsi=ju tegelikult `präegu* juurde. V parandusalgatusele järgnebki 3. vooru eneseparandus, kus H ütleb oma eelnenud voorust puudunud lausungiosa välja.

3.2.2. Kõneleja kordab kaaslase lokaliseeritud probleemallikat. Selliseid kolmanda vooru eneseparandusi põhjustavad nagu eelmisegi alapunkti parandusalgatusi mitmesugused küsisõnad, mis võimaldavad raskusi tekitanud kohta täpselt määratleda, samuti lünkkordusega parandusalgatused. Erinevalt eelmises alapunktis kirjeldatud juhtudest ei ole kaaslase parandusalgatuse põhjus eelmise vooru ebamäärasus. Pigem on võimalik, et vestluskaaslane ei kuulnud päris täpselt osa eelpool kõnelnu voorust või ütles kõneleja midagi kuulaja meelest täpsustamist ja/või kinnitamist vajavat. Järgmine näide on analüüsi aluseks olnud materjali hulgas ainuke siia rühma kuuluv kü-

sisõnaga parandusalgatus. Näite esimese rea algus ei olnud ka lindil kuuldav. Kaaslase küsimus *kes* lokaliseerib selgelt raskuskohaks mittekuuldava noomenifraasi V esimesest voorust. Küsimusele vastates ütlebki V raskusi tekitanud noomenifraasi uuesti: *vint*.

(7) 99a7
1. → V: {-} on praegu ´akna all.
2. H: kes (0.5)
3. → V: *vint*
4. H: vint .hh mina=k (0.4) ´mina määrasin ´ka eile ´vindi

Lünkkordusega parandusalgatus on näites (8). Lünkkorduse puhul lokaliseerib kaaslane raskuskohale eenenud lausungiosa kordusega, jättes lausungi lõpetamise õiguse selle väljaütlejale. Näites kordab H osa V öeldud telefoninumbrist: *null null*. Neljandas reas viib V paranduse lõpuni, korrates oma eelmises voorus öeldud numbri lõppu: *null üks*.

(8) 456a12
1. H: .hh e palun Salu´taarise telefon. (0.8)
2.→V: jaa= üks=´hetk?(...) .hh ´seitse ´kolm ´neli, (1.0) ´null ´null, (0.8) ´null= üks. (.)
3. H: * null null, * (0.5)
4. →V: ´null, ´üks.
5. H: aitäh.

3.2.3. Kõneleja lisab eelmisele voorule olulist infot. Kombinatsiooniga *mis* + probleemallikas ja pikemate arusaamisraskustele osutavate küsimustega soovib kaaslane saada lisainfot, mis aitaks tal kõneleja öeldut tõlgendada. Nii järgnebki sellistele parandusalgatustele kõneleja pikem või lühem selgitus või täpsustus.

(9) 226 a3
1. →V: ´nii tehaksegi ´kübarat. (0.8)
2. H: ´mis kübarat. (0.8)
3. → V: me peame ´koolis ´tegema. (0.5)
4. H: mm
5. V: tööõpetuses. (0.5) ´lõputööks.

Näites (9) selgitab V kaaslase kombinatsiooniga *mis* + probleemallikas reageerides oma 1. reas sõnastatud vooru.

3.2.4. Kõneleja reaktsioon raskuskohta lokaliseerivale küsimusele. Kuigi kõikidele ülalkirjeldatud kolmanda vooru eneseparandustele eelneb parandusalgatus, mis võimaldab kõnelejal kaaslasele tõl-

gendusraskusi tekitanud kohta üsna konkreetselt määratleda, reageerib kõneleja erinevaile parandusalgatustele erinevalt.

Tabel 2. Kõneleja reaktsioon raskuskohta lokaliseerivale küsimusele

Parandusalgatuse vorm	Kõneleja reaktsioon	Arv
Raskuskohta lokaliseeriv küsisõna	Raskuskoha täpsustus: • referentsiseose täpsustus • puudunud fraasi lisamine	7 3 4
	Raskuskoha kordus	1
	Info lisamine	1
Lünkkordus	Raskuskoha kordus	3
Pikem raskuskohta lokaliseeriv küsimus, kombinatsioon *mis* + raskuskoha kordus	Info lisamine	13

Kui kaaslasele tekitab raskusi viiteseose tõlgendamine või puudub eelmisest voorust mõni info vastuvõtu seisukohalt oluline osa, piisab sellest, kui kõneleja vastavat viiteseost täpsustab või puuduva lausungiosa välja ütleb. Sellisel juhul ei lisandu kolmanda vooru eneseparandusele enamasti selgitust, vaid vestlejad lähevad jutuga sujuvalt edasi. Selliseid parandusi algatab kaaslane mitmesuguste raskuskohta lokaliseerivate küsisõnadega. Vaadeldud 9 küsisõnaga algatatud parandusest 7 piirdusid raskuskoha täpsustusega.

Kui kaaslasel on probleeme eelmise vooru sisu mõistmisega ja ta soovib, et eelpool kõnelnu kinnitaks probleeme tekitanud osa veelkord, võib kaaslane raskusi tekitanud koha määratleda mõne küsisõnaga või lünkkordusega. Ka sel juhul lokaliseerib kaaslane raskuskoha üsna täpselt ning piisab, kui kõneleja juba määratletud probleemallikat kordab. Siia rühma kuulub üks küsisõnaga parandusalgatus ja kolm lünkkordusega parandusalgatust.

Kombinatsiooniga *mis* + probleemallikas; pikemate arusaamisraskustele osutavate küsimustega, nagu *mis see on* soovib kaaslane saada lisainformatsiooni, mis aitaks tal kõneleja öeldut tõlgendada. Nii järgnebki sellistele parandusalgatustele kõneleja selgitus või täpsustus. Lisainfo andmine järgnes ka ühele küsisõnaga algatatud parandusele.

3.3. Kõneleja reaktsioon raskuskoha kordusele ja tõlgendusettepanekule

Et määratleda raskuskohta, võib vestluskaaslane seda järgmises voorus korrata. Sellisel juhul ütleb kaaslane probleemallika ise välja, nii et kõnelejal jääb üle seda kas kinnitada või see tagasi lükata ja/või parandada. Samalaadse kõneleja reaktsiooni kutsuvad esile kaaslase tõlgendusettepanekud.

3.3.1. Kõneleja kinnitab kuulaja lokaliseeritud probleemallikat või pakutud tõlgendust. Sellisel juhul on kõneleja roll lisada oma heakskiit. Kaaslane võib eelmise vooru osalise kordusega väljendada imestust, kahtlust, huvi või muud varem kõnelnu vooru tõlgendamisega seotud emotsiooni, samuti soovida saada kõneleja veelkordset kinnitust kahtlusi tekitanud kohale. Juhul kui kaaslase esile tõstetud koht kõneleja jaoks probleemne ei ole, kinnitab ta varem öeldut veel kord. Samu vahendeid kasutab kõneleja ka kaaslase pakutud interpretatsiooni heaks kiites. Kõneleja kinnitus järgnes ka kahele materjali hulgas leidunud partikliga *jah?* parandusalgatusele.

Eelmise vooru kinnituseks kasutas kõneleja järgmisi võimalusi.

1. Lühike kinnitus. Sel juhul reageerib kõneleja kaaslase kordusele/tõlgendusele heakskiitvalt, sõltuvalt eelnenud lausungi kõneliigist kas partikliga *jah* või selle variantidega; tagasisideüneemidega *mhmh, mqm.*

2. Kõneleja kordab kaaslase sõnastatud probleemallikat. Et eelmist vooru kinnitada, võib kõneleja kaaslase määratletud raskuskohta korrata ja lisada kordusele heakskiitva ja kinnitava partikli *jah*, kusjuures partikkel võib kordusele kas eelneda või järgneda.

Eesmärk kinnitada kaaslasele raskusi tekitanud kohta on kombinatsioonil probleemallika kordus + *jah* ka alltoodud näites.

(10) 242
1. H: mina külastasingi: (.) ˈV-Välke.
2. V: jah.
3. H: kahe ˈTimmoga.
4. → V: ˈkahega=või.
5. → H: kahega ˈjah.
6. V: a neid pidi olema oma ma=i=tea, (.) ˈkaheksa=või.

3.3.2. Kõneleja annab öeldu kohta lisateavet. Reageerides vestluskaaslase kordusele, lisab kõneleja mõnikord eelmise vooru kinnitusele asjakohast teavet. Märkimisväärne on, et osal juhtudest lisainfo

andmisele eelmise vooru kinnitust ei eelne. Näites (11) järgneb kinnitavale kordusele täpsustus (rida 8).

(11) 99 a9
1. H: $ kui 'sul on täna 'aega kell 'neli? (0.8) siis ermi 'näituste
2. majas on üks hästi
3. 'kortsuline (.) neenetsi 'ätt (0.5) kes räägib muuhulgas 'neenetsi
4. 'mütoloogiast. $
---
5. V: ota mis 'kell see 'oli.
6. H: kell 'neli kell 'kuusteist.
7. →V: ermis
8. →H: ermis seal 'näituste majas.
9. V: ahah

3.3.3. Kõneleja ei kinnita kaaslase korratud raskuskohta/tõlgendust. Kuna kaaslane omistab kordusega/tõlgendusettepanekuga alati mingisuguse propositsiooni eelpool kõnelnule, jääb alati võimalus, et viimane kaaslase tõlgendust heaks ei kiida. Vaadeldud 100 vestluse hulgas oli vaid 3 kordusega parandusalgatust ja 9 tõlgendusettepanekut, mida kõneleja ei kinnitanud. Kõigil juhtudel asendas kõneleja väära tõlgenduse õigega. Tõlgenduste/korduste kummutamise vormistas kõneleja järgnevalt.

1. *Ei* + parandus. Kõneleja kummutab kaaslase tõlgenduse partikliga *ei* ning lisab sellele paranduse.

2. Kõneleja parandab väära korduse/tõlgenduse, kasutamata parandusele osutavaid vahendeid. Selline võimalus on näites (12) reas 3.

(12) 455b17
1. V: ja 'täpne 'aadress on 'Turu kolmkend kaks 'dee. (1.5)
2. →H: kolmkend kaks 'bee või.
3. →V: 'dee nagu 'doomino.
4. (0.5) j:ah.(.) < seitse kolm kuus? (1.5) kuus seitse üheksa null.

3.3.4. Kõneleja reaktsioon kordusega parandusalgatusele ja tõlgendusettepanekule. Tabel 3 võtab kokku kõneleja reaktsiooni kaaslase kordusega parandusalgatustele ja tõlgendusettepanekutele.

Nii korduste kui tõlgenduste kinnitamiseks ja kummutamiseks kasutas kõneleja samu keelelisi vahendeid. Eelmise vooru minimaalseks kinnitamiseks kasutas kõneleja partikleid *jah, ei, mhmh, ähäh, mqm*; kombinatsiooni kordus + *jah*. Tähelepanu äratab asjaolu, et korduste puhul kasutas kõneleja *jah*-i lühikese kinnitava tagasisidena 36 korral ja *mhmh*-i 12 korral, tõlgenduste kinnitamiseks kasutas

kõneleja *mhmh*-i vaid 1 korral, ülejäänud juhtudel oli kinnitavaks keelendiks *jah* või kombinatsioon kordus + *jah*.

Tabel 3. Kõneleja reaktsioon kordusele ja tõlgendusettepanekule

Parandusalgatuse tüüp	Kõneleja reaktsioon	Arv
Raskuskoha kordus	Lühike kinnitus:	**58**
	• *jah*	36
	• *ei*	1
	• *mhmh*	12
	• *mqm*	3
	• *ähäh, õhõh*	2
	• kordus + (*jah*)	4
	(Kinnitus) + lisainfo	**16**
	Parandab eelmise vooru	**3**
Tõlgendusettepanek	Lühike kinnitus:	**26**
	• *jah*	19
	• *nojah*	1
	• *mhmh*	1
	• *ei*	1
	• kordus + *jah*	4
	(Kinnitus) + lisainfo	**7**
	Parandab tõlgenduse:	**9**
	• parandab tõlgenduse	2
	• *ei* + tõlgenduse parandus	7

Kõneleja lisab omapoolse kinnituse ka juhul, kui kaaslane algatab paranduse partikliga *jah?*. Reageerides erinevatele kordusekombinatsioonidega parandusalgatustele ja tõlgendustele, võib kõneleja lisaks kinnitavale tagasisidele anda eelmise vooru kohta lisateavet.

Kui kõneleja soovib vestluskaaslase tõlgendust kummutada, võib ta selle tühistada keelendi *ei* abil. Varem kõnelnu lisas alati pärast kaaslase interpretatsiooni kummutamist ka omapoolse paranduse. Võimalik on ka see, et kõneleja parandab kaaslase väära tõlgenduse seda eelnevalt tagasi lükkamata.

4. Kokkuvõte

Parandusalgatuste tüübistik on astmeline: eri rühmadesse kuuluvad algatused määratlevad raskuskoha asukohta eri määral. Tüübistiku ühes otsas paiknevad tervele voorule laienevad parandusalgatused; teises otsas algatused, kus kaaslane ise raskuskoha välja ütleb: kordused ja tõlgendused. Analüüs näitas, et 84% avatud tüüpi parandus-

algatustest reageeris kõneleja probleemvooru kordusega, ülejäänud 16% reaktsioon oli probleemvooru muutmine. Raskuskohta lokaliseerivatest küsimustest 78% päädisid raskuskoha täpsustamisega, 11% raskuskoha kordusega ja 11% info lisamisega. Lünkkordused viisid kõikidel juhtudel raskuskoha korduseni. Pikematele raskuskohta lokaliseerivatele küsimustele ja kombinatsioonile *mis* + raskuskoha kordus reageeris kõneleja alati lisainfot andes. Reageerides kordusega parandusalgatustele ja pakutud tõlgendustele piisas minimaalsest tagasisidest: 75% kordustest ja 62% tõlgendustest kinnitas kõneleja lühikese tagasisidega. Kõneleja kinnitus järgnes ka kahele materjali hulgas leidunud partikliga *jah?* parandusalgatusele. Kordust ei kinnitanud kõneleja 4% juhtudest ja tõlgendust 21% juhtudest. Lisaks kinnitusele andis kõneleja lisainfot tõlgenduste puhul 17% juhtudest ja korduste puhul 21% juhtudest.

Kõneleja reaktsioon kaaslase parandusalgatusele sõltub parandusalgatuse tüübist, kuid ei ole sellega alati üksüheselt seotud. Siiski on võimalik välja tuua peamised tendentsid, kuidas kõneleja igale parandusalgatuse tüübile tavaliselt reageerib. Analüüs kinnitab ja täiendab Tartu Ülikooli dialoogi modelleerimise töörühma tehtud uurimust vestluskaaslase algatatud paranduste kohta (Gerassimenko jt 2004).

Kirjandus

Gerassimenko, Olga, Hennoste, Tiit, Koit, Mare, Rääbis, Andriela 2004. Other-initiated self-repairs in Estonian information dialogues: solving communication problems in cooperation. – Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, April 30 – May 1, 2004. Ed by M. Strube, C. Sidner. Cambridge, 39–42.

Hennoste, Tiit 2001. Sissejuhatus suulisesse eesti keelde. Lausung suulises kõnes IV. – Akadeemia 13, 1, 179–206.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli kirjastus.

Schegloff, Emanuel 1979. The relevance of repair to syntax-for-conversation. – Syntax and Semantics, Volume 12: Discourse and Syntax. Ed by T. Givon. N.Y: Academic Press, 261–288.

Schegloff, Emanuel, Jefferson, Gail, Sacks, Harvey 1977. The preference for self-correction in the organization of repair in conversation. – Language 52(2), 361–382.

Sorjonen, Marja-Leena 1997. Korjausjäsennys. – Keskustelunanalysin perusteet. Toim L. Tainio. Tampere: Vastapaino, 111–137.

Tainio, Liisa (toim) 1997. Keskusteluanalyysin perusteet. Tampere: Vastapaino.

Suhtlusstrateegiad infodialoogides[1]

Liina Eskor
Tartu Ülikool

Käesolev artikkel analüüsib suhtlusstrateegiaid 20 infotelefonidialoogis. Kui seni olen dialooge vaadelnud Soome keeleteadlase K. Jokineni suhtlusstrateegiate seisukohalt (Eskor 2004, 2005), siis nüüd on ülesandeks välja selgitada, missuguseid strateegiate järjendeid ehk laiemas mõistes strateegiaid vestluses osalejad kasutavad. Suhtlusstrateegiate leidmine ja analüüsimine on osa Eesti dialoogikorpuse[2] projektist. Dialoogikorpuse põhjal luuakse eksperimentaalne dialoogsüsteem, kus arvuti suudaks võimalikult hästi inimkäitumist jäljendada. Süsteem suhtleks kasutajaga eesti keeles ja selle abil saaks kasutaja infot erinevate ainevaldkondade kohta (ostu-müügiinfo, reisi planeerimine, liiklusinfo, teejuhatamine jms). Suhtlusstrateegiad võiksid sellises süsteemis olla arvutile abiks küsimuste mõistmisel ja täpsustavate küsimuste ning vastuste genereerimisel.

1. Mis on suhtlusstrateegia?

Suhtlusstrateegia mõistet on käsitletud erinevates valdkondades:

- intellektitehnikas uuritakse strateegiaid seoses suhtluse planeerimise ja suhtluspartneri plaani tuvastamisega (Litman, Allen 1987);
- diskursuseanalüüsis tähendab strateegia viisi, kuidas teksti autor viib ellu oma üldisi kavatsusi – kutsuda esile muudatusi vastuvõtjas (lugejas) (Van Dijk 1983; McKeown 1982);
- suhtlusanalüüsis mõistetakse strateegia all korrapära, milles üks osaleja genereerib korduvalt oma suhtlusakte, saavutamaks teatud suhtluseesmärki (Heritage 1991);
- võõrkeele õppimisel käsitletakse strateegiat kui mentaalset protsessi, mida õppur pidevalt kordab seoses keeleinfo meeldejätmise, taaskasutuse ja rakendamisega (Cohen 1998).

---

[1] Tööd on toetanud Eesti Teadusfond (grant nr 5685) ning HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

[2] Vt dialoogikorpuse kohta pikemalt (Hennoste, Rääbis 2004: 13–14) ja http://math.ut.ee/~koit/Dialoog/EDiC.

Huvi koostöödialoogi vastu, kus osalejad tegutsevad ühise eesmärgi nimel, on eriti tõusnud seoses veebiteenuste laia levikuga. Enamasti käsitletakse just kooperatiivseid probleemilahendusdialooge. Paljud uurijad on modelleerinud nõustumisprotsessi kooperatiivses dialoogis, s.o situatsiooni, kus üks osaleja teeb partnerile ettepaneku ja partner kas aktsepteerib seda või mitte. Chu-Carroll ja Carberry (1998) esitavad kooperatiivse vastuste genereerimise mudeli tsüklina 'ettepanek-hindamine-modifitseerimine', keskendudes infoandmis- ja läbirääkimisdialoogidele. Heeman ja Hirst (1995) modelleerivad koostööd tsükli 'esitamine–arvustamine–ümberkujundamine' abil. Lochbaum (1998) esitab ühistegevuse mudeli, mis käsitleb planeerimisprotsessi, kus osaleb mitu agenti. Kesksel kohal on agentide kavatsuste tuvastamine ja nende koordineerimine ühise eesmärgi saavutamiseks. Di Eugenio jt (2000) esitavad mudeli 'kaalumine–ettepanek–seisukohavõtt': esmalt kaalutakse infot, arutatakse, peetakse nõu, siis tehakse ettepanek ja lõpuks võetakse seisukoht – kas aktsepteerida ettepanekut või mitte.

Jokinen kasutab suhtlusstrateegiate mõistet (ingl *communicative strategy*) konstruktiivse dialoogi juhtimise mudeli (ingl *Constructive Dialogue Management* ehk CDM) juures (Jokinen 1996a). Mudel lähtub üldistest suhtlusprintsiipidest, mis piiritlevad kooperatiivset ja sidusat suhtlust. Mudelis käsitletud suhtlusprintsiibid on H. Paul Grice'i maksiimide (vt nt Jurafsky, Martin 2000: 727) täpsustus. Need on üldised põhimõtted, mille abil suhtlejad vestluse käigus üksteise lausungeid tõlgendavad. Grice toob välja neli maksiimi: kvantiteedimaksiim (anna nii palju infot kui tarvis ja mitte rohkem), kvaliteedimaksiim (räägi tõde: ära ütle midagi, mille kohta tead, et see on vale, või mille kohta sul ei ole adekvaatset infot), relevantsusmaksiim (ole relevantne) ja meetodimaksiim (ole arusaadav: väldi segasust ja mitmemõttelisust, räägi lühidalt).

Jokineni mudelis saab arvutiprogramm suhtlusprintsiipide abil valida sobiva suhtlusstrateegia ja lahendada kasutaja kaasabil nii oma andmebaasi puudujääke. Sobiva suhtlusstrateegia valik tähendab seda, et süsteem valib erinevate dialoogi jätkamise võimaluste vahel (küsida kasutajalt infot juurde, soovitada muid võimalusi, paluda küsimust korrata jms). Formaalsemalt öeldes on suhtlusstrateegia võimalus ühiste teadmiste loomiseks, säilitamiseks, muutmiseks ja rakendamiseks. Strateegiad põhinevad suhtleja ratsionaalsusel:

suhtleja tegevused/lausungid valitakse nii, et ühised eeldused koostöövalmis käitumisest oleksid antud olukorras samad.

Mõned strateegiad, mida kõnelejad kasutavad ühise info töötlemisel, osutuvad edukamateks kui teised. Selgub, et ei piisa ainult sobiva faktilise informatsiooni andmisest partnerile. Samuti on oluline märkida, et mida paremini suudavad suhtlejad probleeme ja arusaamatusi lahendada, seda täpsemalt ja kiiremini saavutatakse oma eesmärk. (Jokinen 1996a).

Suhtlusstrateegia valimisel arvestatakse iga lausungi puhul nelja faktorit (Jokinen 1996b).

• Ootused/reaktsioon – kas lausung on eelmist lausungit arvestades ootuspärane (kas lausungiga vastatakse eelmisele lausungile või mitte)? Reaktsioon võib olla oodatud ja mitteoodatud.

• Teema – kas lausung on seotud vestluse senise teemaga? Lausung võib eelmise teemaga olla seotud või mitteseotud.

• Initsiatiiv – kas initsiatiiv on kõnelejal või partneril? Kui üks suhtlejatest on algatanud uue suhtluseesmärgi, siis on suhtluse initsiatiiv parajasti tema käes. Agendil on õigus püüda oma eesmärki täita kuni see on saavutatud või pole enam relevantne. Samuti on agendil õigus oodata suhtluspartneri koostööd või vähemalt seda, et ta ei takistaks agendi eesmärkide saavutamist.

• Eesmärk – võib märgendatava lausungi ajal olla täidetud või täitmata. Kui partner vastab ammendavalt agendi küsimusele või soovile, täitub sellega agendi eesmärk. Täitmata eesmärgi korral jäetakse see kõrvale või hoitakse hilisemaks lahendamiseks. Kui agendil on uus võimalus initsiatiiv haarata, võib ta uuesti püüda oma täitmata eemärki saavutada.

Ülaltoodud nelja faktorit kombineerides moodustub 16 suhtlusstrateegiat (tabel 1).[3]

[3] Strateegiate eestikeelsed nimetused: *backto* = tagasi eelmise juurde; *follow-up-old* = jätka eelmisega; *finish/start* = eelmine lõpetatud, alusta uuega; *follow-up-new* = jätka uuega; *repeat-new* = korda uut; *new question* = uus küsimus; *specify* = täpsusta eelmist; *new request* = uus palve; *subquestion, X* = lisaküsimus; *continue* = jätka; *new dialogue* = uus dialoog; *somethingelse* = muu; *object, X* = vaidle vastu; *notrelated* = teemaga sidumata; *specify-new* = paku jätkamist; *new st-request* = uus kaudne küsimus.

Tabel 1. 16 võimalikku suhtlusstrateegiat K. Jokineni järgi

Vastus	Teema	Eesmärk	Kõneleja initsiatiiv	Partneri initsiatiiv
oodatud (expected)	seotud (related)	täitmata (unfulfilled)	backto	follow-up-old
		täidetud (fulfilled)	finish/start	follow-up-new
	mitteseotud (unrelated)	täitmata	repeat-new	new question
		täidetud	specify	new request
mitte-oodatud (non-expected)	seotud	täitmata	subquestion, X	continue
		täidetud	new dialogue	somethingelse
	mitteseotud	täitmata	object, X	notrelated
		täidetud	specify-new	new st-request

Suhtlusstrateegiad Jokineni mõistes tähendavad iga lausungi strateegia määramist. Dialoog, kus info küsijal on laiemas mõistes üks eesmärk (nt telefoninumbri saamine), sisaldab endas ka vastaja poolt tekitatud väiksemaid eesmärke (firma nime, asukoha täpsustamine jms). Nende kõigi täitmine on põhieesmärgi saavutamise eelduseks.

2. Infodialoogide analüüs

Järgnevalt analüüsitakse dialooge lähtudes suhtlusstrateegiatest laiemas mõttes. Selleks on tarvis vaadelda, mis on helistaja konkreetne eesmärk ja kas selle saavutamiseks piisab ühest küsimusest või soovist, millisel juhul helistaja täpsustab oma soovi või sõnastab selle ümber. Vastaja (infotelefoni töötaja) seisukohalt on strateegia leidmiseks vaja analüüsida vastaja kooperatiivsust (kas ta võtab helistaja eesmärgi oma eesmärgiks). Oluline on ka eristada, kas vastaja vastab helistaja küsimusele kohe või täpsustab küsimust, kas vastus on piisav või tuleb seda täiendada. Lisaks kirjeldatakse küsimuste ja vastuse tingimuste täpsustamise keelelist formuleerimist. Analüüsi käigus vaadeldakse, milliseid dialoogiaktide[4] ja (Jokineni mõttes) suhtlusstrateegiate järjendeid kasutavad suhtlejad eesmärkide saavutamiseks.

[4] Dialoogiaktide tüpoloogia kohta vt Lisa 2 ja Hennoste, Rääbis 2004. Dialoogiaktid on tegevused, mida inimene kõne abil teeb. Aktid jagunevad kahte suurde rühma: naaberpaariaktid (nt küsimus–vastus, direktiiv – direktiivi täitmine) ja üksikaktid (ei vaja partneri tagasisidet, nt jätkaja *mhmh*).

2.1. Helistaja eesmärk ja selle saavutamine

Infodialoogides algatab infovahetuse helistaja oma küsimuse või direktiiviga. Analüüsitud dialoogides soovitakse teada saada peamiselt telefoninumbreid (17 dialoogis), aga ka busside väljumisaegu (2 dialoogis), infot kinos linastuva filmi (1 dialoogis) ja tänavate asukoha kohta (1 dialoogis). Helistaja võib oma eesmärgi (telefoninumbri, bussiinfo vms saamine) saavutada ühe direktiivi nt DIE: SOOV (7 dialoogis, näide 1) või küsimusega, nt KYE: AVATUD (2 dialoogis, näide 2), KYE: JUTUSTAV KAS (1 dialoogis), KYE: SULETUD KAS (1 dialoogis), KYE: ALTERNATIIV (2 dialoogis, näide 3), st see on vastuse andmiseks piisav. Nimetatud dialoogiaktidele vastab strateegia finish/start.[5]

(1)

H: tere *='öelge palun: 'linnaliini 'bussijaama 'infotelefoni 'number.*	RIJ: VASTUTERVITUS *DIE: SOOV*	– *finish/start*
V: kolm kuus kaks	DIJ: INFO ANDMINE	follow-up-old

(2)

H: tere. *ma sooviksin 'teada 'mis 'film jookseb 'hetkel Rakvere 'kinos.*	RIJ: VASTUTERVITUS *KYE: AVATUD*	– *finish/start*
V: üks hetk mm se=on=sis: 'Mina {ma 'ise} ja l'reen.	KYJ: EDASILÜKKAMINE KYJ: INFO ANDMINE	notrelated follow-up-old

(3)

H: tere *oskate 'öelda, 'kummalt poolt akkab Alek'sandri 'tänav, kas 'kesklinna poolt või sealt (.) {-} juurest.*	RIJ: VASTUTERVITUS *KYE: ALTERNATIIV*	– *finish/start*
V: 'kesklinna=poolt.	KYJ: ALTERNATIIV: ÜKS	follow-up-old

Tüüpiliselt sisaldab dialoog infovahetust ühel teemal (nt ühe firma telefoninumber), kuid esineb ka dialooge, kus helistaja küsib mitut telefoninumbrit (näide 4) või infot mitme valdkonna kohta. Sellisel juhul kannab iga uus küsimus/direktiiv samuti strateegiat finish/ start.

[5] Vaadeldav nähtus on näidetes tähistatud kursiiviga. Dialoogide liitereerimisel on kasutatud konversatsioonianalüüsi transkriptsiooni. Vt Lisa 1.

(4)

H: tere päevast. *sooviksin: (.) Ekspress=reiside telefoni'numbrit.*	RIJ: VASTUTERVITUS *DIE: SOOV*	– *finish/start*
V: neli neli üks, (.) üheksa kuus kaheksa.	DIJ: INFO ANDMINE	follow-up-old
H: ja=ned 'Partner=reisid?	*DIE: SOOV*	*finish/start*
V: kolm üheksa null, viis üheksa kolm.	DIJ: INFO ANDMINE	follow-up-old

Helistaja võib oma direktiivi/küsimust ka täpsustada (näide 5). Sellisel juhul väljendab helistaja esimene küsimus/direktiiv strateegiat finish/start ning teine strateegiat backto (st jätkab sama küsimust).

(5)

H: tere. *.hh ee mul siuke 'küsimus tähendab=et=e see (.) 'Tartu 'maratoni 'korraldus*	RIJ: VASTUTERVITUS *DIE: SOOV*	– *finish/start*
V: jah?	VR: NEUTRAALNE JÄTKAJA	continue
H: e (0.5) ma=i=tea=kas se 'on mingi 'koht, (.) komitee {-} selle telefoni'numbrit oleks võimalik. 'leida kuskilt.	*KYE: JUTUSTAV KAS* IL: TÄPSUSTAMINE	*backto* backto

Sageli esineb ka olukordi, kus helistaja esialgne küsimus ei ole vastaja jaoks piisavalt konkreetne ning seetõttu tuleb seda kõigepealt täpsustada. Kui infotelefoni andmebaasis on mitu sobivat vastust, soovib vastaja teada, milline variant küsijat huvitab. Vastaja võib ka kontrollida, kas ta on küsimust õigesti kuulnud, täpsustada kohta või aega. Selleks täpsustab ta vastuse tingimusi, kasutades erinevat tüüpi küsimusi (koos aktiga VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE akte KYE: VASTUST PAKKUV (näide 6), KYE: ALTERNATIIV (näide 7) või KYE: AVATUD (näide 8)). Vastaja võib ka küsimuse üle küsida aktidega PPE: ÜLEKÜSIMINE, KYE: VASTUST PAKKUV (näide 9). Vastuse tingimuste täpsustamine ja üleküsimine kannavad strateegiat subquestion, X. Vastuse tingimuste täpsustamisele järgneb helistaja täpsustus, mis on tähistatud nt aktidega KYJ: JAH, KYJ: ALTERNATIIV (tähistatud strateegiaga follow-up-old) (näited 6–9) või uus helistajapoolne küsimus, nt KYE: AVATUD (näide 10) strateegiaga backto. Seega tähistab vastuste tingimuste täpsustamist strateegiate finish/start – subquestion, X – follow-up-old/backto järjend.

(6)

H: halloo? (.) tere (.) ee ´Ilmatsalu tänaval on mingi ´ilusalong.	KKE: ALGATUS RIJ: VASTUTERVITUS DIE: SOOV	specify-new – finish/start
V: ee kauplus ´Kimsto juures.	*KYE: VASTUST PAKKUV, VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE*	*subquestion, X*
H: jaa jaa ja.=	*KYJ: JAH*	*follow-up-old*

(7)

H: tere ´õhtust, ma paluks ´Maarjamõisa ´kööki.	RIJ: VASTUTERVITUS DIE: SOOV	– finish/start
V: ´haigla juures või poli´kliinikus.	*KYE: ALTERNATIIV, VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE*	*subquestion, X*
H: ´haigla.	*KYJ: ALTERNATIIV: ÜKS*	*follow-up-old*

(8)

H: tere. öelge=palun eraarst hambaarst {-} Vigorooviti numbrit teil ´on.	RIJ: VASTUTERVITUS KYE: JUTUSTAV KAS	– finish/start
V: e kus see arst ´asub.	*KYE: AVATUD, VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE*	*subquestion, X*
H: Tuglase kaks.	*KYJ: INFO ANDMINE*	*follow-up-old*

(9)

H: tere. (.) ma paluks Elvas ´Teetorni baari ´numbrit.	RIJ: VASTUTERVITUS DIE: SOOV	– finish/start
V: Elvas.	*KYE: VASTUST PAKKUV, PPE: ÜLEKÜSIMINE*	*subquestion, X*
H: jah.	*KYJ: JAH, PPJ: LÄBIVIIMINE*	*follow-up-old*

(10)

H: ee tervist. ega ´teie ei oska kogemata öelda Tallinna ´busside aegu.	RIJ: VASTUTERVITUS KYE: JUTUSTAV KAS	– finish/start
V: ja ´milliseid busse.	KYE: AVATUD, VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE	subquestion, X
H: ee (.) noh millal lähevad ´viimased ´kiirbussid täna ´Tallinna.	*KYJ: INFO ANDMINE, KYE: AVATUD*	*backto*

2.2. Kas vastus on piisav või tuleb seda täiendada?

Üldjuhul suudab vastaja peale vastuse tingimuste täpsustamist anda info, mis rahuldab helistajat ning helistaja ei palu enam täiendamist. Erandiks on olukorrad, kus vastaja ei ole oma vastuses päris kindel ja helistaja püüab omalt poolt mingit lahendust leida, oodates seejuures ka vastaja tagasisidet sellele (näide 11). Helistaja püüab vastust täpsustada ka siis, kui ta hakkab saadud info õigsuses kahtlema. Sellisel juhul soovib ta (aktiga PA: PARTNERI PARANDUS) partnerit parandades vestlust jätkata (näide 12). Vastuse täiendamist ootab ka helistaja akt PPE: ÜLEKÜSIMINE (näide 13). Helistaja kasutab seejuures strateegiat new dialogue.

(11)

H: kas ˈtelefoninumbrid mis algavad numbritega viis ˈkaks, (.) kas need on ˈRakvere või: ˈPajusti omad.	KYE: ALTERNATIIV, TVE: PAKKUMINE	finish/start
V: {-} (.) tähendab (.) vaatan? (22.0) no ˈPajustis akkavad numbrid viis ˈseitse=aga: (0.5) Rakvere::s (.) ma nüüd ei ˈnäe=et viis ˈkaks päris oleksid=aga	KYJ: EDASILÜKKAMINE, TVJ: VASTUVÕTMINE KYJ: ALTERNATIIV: EITAV	notrelated follow-up-old
H: ahah *aga võibola see siin on mingi ˈmuu koha number lihtsalt.*	VR: NEUTRAALNE INFO OSUTAMINE UUEKS *SEE: ARVAMUS*	something-else *new dialogue*

(12)

H: öelge mis=se ˈnumber ˈon.	KYE: AVATUD	backto
V: m ˈkinnitamata andmetel neli kaks kaks, kaks seitse kuus.	KYJ: INFO ANDMINE	follow-up-old
H: ee ˈPõlvas.	*PA: PARTNERI PARANDUS*	*new dialogue*
V: ˈPõlvas.	KYE: VASTUST PAKKUV, PPE: ÜLEKÜSIMINE	subquestion, X
H: jah.	KYJ: JAH, PPJ: LÄBIVIIMINE	follow-up-old
V: Põlvas ˈTuglase tänaval ei ole antud ˈühtegi hambaravi meile.	KYJ: INFO PUUDUMINE	continue

(13)

H: ahah (.) no vast on ikka õige. *viis=viis=õheksa oli viimane jah?*	VR: NEUTRAALNE INFO OSUTAMINE UUEKS, VR: PARANDUSE HINDAMINE VR: HINNANGULINE VASTUVÕTU-TEADE *KYE: VASTUST PAKKUV, PPE: ÜLEKÜSIMINE*	somethingelse somethingelse *new dialogue*
V: jah	KYJ: JAH, PPJ: LÄBIVIIMINE	follow-up-old

2.3. Küsimuste ja vastuste keeleline formuleerimine

Helistaja kasutab küsimuse formuleerimisel avatud küsimust (näide 2), alternatiivküsimust (st pakub küsimusega välja variandid ning soovib, et vastaja valiks mõne neist, näited 3 ja 11), suletud *kas*-küsimust (st soovib jah- või ei-vastust, näide 16), jutustavat *kas*-küsimust (st soovib rohkem infot kui "jah" või "ei", näited 5, 8 ja 10). Direktiivid (nt soov) sisaldavad tavaliselt viisakusvormeleid "öelge palun" (näide 1), "ma paluks" (näide 9), "(ma) sooviksin" (näited 2 ja 4) jms, kuid võivad esineda ka ilma viisakusvormeliteta (näited 5 ja 6).

Vastaja kasutab vastuse tingimuste täpsustamiseks erinevat tüüpi küsimusi, sõltuvalt sellest, millist infot ta helistajalt ootab, nt avatud küsimus (näide 8), alternatiivküsimus (näide 7). Vastust pakkuv küsimus esineb elliptilisena (näited 6, 9 ja 12).

2.4. Kooperatiivsus

Vestluses on eesmärgi saavutamiseks vajalik suhtlejate koostöö, st vastaja peaks olema kooperatiivne ning võtma helistaja eesmärgi oma eesmärgiks (st andma lõpuks soovitud infot). Kuna analüüsitud infotelefoni dialoogides on eesmärgiks info, siis võib eeldada, et sellise suhtluse puhul on tegemist kooperatiivsete suhtlejatega. Vaadeldud 20 dialoogist sai klient sobiva info (või nn asendusinfo) 18 dialoogis. Vastaja koostöövalmidus tuleb hästi esile olukordades, kus andmebaasis soovitud info puudub. Sellisel juhul pakutakse mingeid alternatiivseid vastuseid aktiga DIE: PAKKUMINE (näide 14) ja/või DIE: ETTEPANEK (näide 15) ning sellega kasutab vastaja strateegiat new dialogue. Samuti võib vastaja oma eitavat vastust põhjendada ning niimoodi siiski kooperatiivsust säilitada, kasutades sealjuures strateegiat follow-up-new (näide 16).

(14)

H: tere ma paluks 'Karlova Güm'naasiumi 'õpetajate 'tuba.	RIJ: VASTUTERVITUS DIE: SOOV	– finish/start
V: mul on 'välja pakkuda di'rektor, kantse'lei ja õppeala'juhataja.	DIJ: INFO PUUDUMINE, *DIE: PAKKUMINE*	continue *new dialogue*

(15)

H: ää kas te oskate öelda kui palju se 'pilet maksab.	KYE: AVATUD	finish/start
V: kahjuks 'piletite=inda meil ei=ole. *te peate sealt küsima=* *ma=võin 'numbri anda kui 'soovite.*	KYJ: INFO PUUDUMINE *DIE: ETTEPANEK* *DIE: PAKKUMINE*	continue *new dialogue* *new dialogue*

(16)

H: tere. palun 'öelge mulle kas 'teie ütlete ka era'isikute korteri'telefone.	RIJ: VASTUTERVITUS KYE: SULETUD KAS	– finish/start
V: ei, *era'isikuid annab ainult tasuline 'infoliin.*	KYJ: EI *IL: TÄPSUSTAMINE*	continue *follow-up-new*

Kooperatiivsust väljendavad ka vastaja vabatahtlikud reaktsioonid, mis tähistavad kuuldel olekut, nt VR: NEUTRAALNE JÄTKAJA, millele vastab strateegia continue (näide 17).

(17)

H: tere. .hh ee mul siuke 'küsimus tähendab= et=e see (.) 'Tartu 'maratoni 'korraldus	RIJ: VASTUTERVITUS DIE: SOOV	– finish/start
V: jah?	*VR: NEUTRAALNE JÄTKAJA*	*continue*
H: e (0.5) ma=i=tea=kas se 'on mingi 'koht, (.) komitee {-} selle telefoni'-numbrit oleks 'võimalik. 'leida kuskilt.	KYE: JUTUSTAV KAS IL: TÄPSUSTAMINE	backto backto

3. Suhtlusstrateegiad infodialoogides

Eelnenud analüüsist lähtudes on võimalik formaliseerida küsimise ja vastamise strateegiad. Kuna tegemist on üpris lihtsa struktuuriga dialoogidega, siis on põhiliseks strateegiaks *küsin infot → annan infot* koos neile liituda võivate strateegiatega. Vaadeldes eraldi küsimise ja vastamise strateegiaid, saab välja tuua järgmised strateegiate jär-

jestused ehk strateegiad laiemas mõttes (lainelistes sulgudes on strateegiad, mis võivad osaliselt või täielikult puududa):

Küsimise strateegia:

{sissejuhatus = YA: EELTEADE = finish/start}
↓
küsin infot = KYE/DIE = finish/start
↓
{parandan partneri palvel = KYJ/DIJ + PPJ: LÄBIVIIMINE = follow-up-old}
↓
{täpsustan omal algatusel = IL: TÄPSUSTAMINE = subquestion, X}
↓
{jätkan küsimust järgmises voorus = KYE/DIE = backto}

Vastamise strateegia:

{annan tagasisidet = VR: NEUTRAALNE JÄTKAJA = continue}
↓
{palun oodata = KYJ: EDASILÜKKAMINE = notrelated}
↓
{täpsustan küsimust = VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE / PPE: ÜLEKÜSIMINE = subquestion, X}
↓
annan infot = KYJ/DIJ = follow-up-old/continue
↓
{annan lisainfot = IL: TÄPSUSTAMINE = follow-up-new}

Kuna tegemist on dialoogidega, kus üks suhtlejatest soovib saada infot ning teise suhtleja ülesandeks on seda talle edastada, siis on sellised strateegiate järjestused üpris ootuspärased. Edaspidi on kavas uurida keerulisema struktuuriga dialooge (nt müügivestlused), kus ka strateegiate valik on tõenäoliselt teistsugune (mh kasutatakse veenmist, mida infodialoogides tavaliselt ette ei tule).

4. Kokkuvõte

Analüüsitud infotelefoni dialoogid on suhteliselt lihtsa struktuuriga (küsimus/direktiiv ning sellele järgnev vastus) ning seetõttu ei ole ka sellist tüüpi dialoogides kasutatavad suhtlusstrateegiad keerulised. Tavapärasele küsimus/direktiiv → vastus järjendile võib ette, vahele või järele liituda muid strateegiaid, kui see on suhtluse eesmärgi saavutamiseks vajalik. Selline dialoogide sisu üldistamine ja strateegiate vähesus võiks aga dialoogsüsteemis arvutile abiks olla kasutaja mõistmisel ja tema suhtluseesmärkide täitmisel.

Kirjandus

Chu-Carroll, Jennifer, Carberry, Sandra 1998. Collaborative response generation in planning dialogues. – Computational Linguistics, 24(3), 355–400.

Cohen, Andrew D. 1998. Strategies in Learning and Using a Second Language, London: Longman.

Di Eugenio, Barbara, Jordan, Pamela W., Thomason, Richmond H., Moore, Johanna D. 2000. The agreement process: an empirical investigation of human–human computer-mediated collaborative dialogs. – International Journal of Human Computer Studies, 53(6), 1017–1076.

Eskor, Liina 2004. Dialoogiaktid ja suhtlusstrateegiad: eesti dialoogikorpuse analüüs. Magistritöö. Tartu Ülikool, üldkeeleteaduse õppetool.

Eskor, Liina 2005. Dialoogiaktid ja suhtlusstrateegiad: eesti dialoogikorpuse analüüs. – Keel ja Kirjandus 9, 711–727.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Heritage, John 1991. Intention, Meaning, and Strategy: Observations on Constraints on Interaction Analysis. – Cognitive Sciences 24, 311–332.

Jokinen, Kristiina 1996a. Cooperative response planning in CDM: Reasoning about Communicative Strategies. – Proceedings of the 10th COLING-96, Stanford, 444–447 http://www.mlab.uiah.fi/~kjokinen/papers/199606TWEN.PDF (kasutatud 13.10.2005)

Jokinen, Kristiina 1996b. Goal Formulation based on Communicative Principles. – Proceedings of The 16th International Conference on Computational Linguistics (COLING-96), Copenhagen, Denmark, 598–603 http://www.mlab.uiah.fi/~kjokinen/papers/199608COLI.PDF (kasutatud 13.10.2005)

Jurafsky, Daniel, Martin, James H. 2000. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River (N.J.): Prentice Hall.

Litman, Diane J., Allen, James F. 1987. Discourse processing and commonsense plans. – Intentions in Communication. Ed by P. R. Cohen, J. Morgan, M. E. Pollack. Cambridge, Massachusetts: MIT Press, 1990, 365–388.

Lochbaum, Karen E. 1998. A collaborative planning model of intentional structure. – Computational Linguistics 24 (4), 525–572.

McKeown, Kathleen R. 1982. Generating Natural Language Responses to Questions about Data Base Structure. Ph.D. Dissertation, University of Pennsylvania.

van Dijk, Teun 1983. Cognitive and Conversational Strategies in the Expression of Prejudice. – Text, 3–4, 375–404.

Loomulik infodialoog ja infodialoogi simulatsioon: infoandja strateegiad[1]

Olga Gerassimenko, Maret Valdisoo
Tartu Ülikool

1. Sissejuhatus

Eestikeelse dialoogsüsteemi loomise eesmärgil on Tartu Ülikoolis kogutud Eesti dialoogikorpus (Hennoste, Rääbis 2004: 13, vt ka http://math.ut.ee/~koit/Dialoog/EDiC), kuhu kuulub 2005. a detsembri seisuga 871 loomulike inimestevaheliste infodialoogide lindistust ning 22 "võlur Ozi" tehnikaga kogutud arvuti ja inimese vahelise infodialoogi simulatsiooni (milles tegelikult on infoandjaks teine inimene ehk "võlur"). Loomulikest dialoogidest saab õppida inimestevahelise suhtluse seaduspärasusi, simuleeritud dialoogid aga on vajalikud aru saamaks, kuidas mõjutab kasutaja käitumist teadmine, et ta suhtleb arvutiga. Kasutaja kõrval on oluline ka arvuti rolli täitva "võluri" käitumine: kuivõrd vastab dialoogsüsteemi loojate ettekujutus infoandja modelleerimisest infoandja tegelikule käitumisele ja vestluses rakendatavatele strateegiatele. Selles artiklis oleme üritanud võrrelda infoandja algatatud alamdialooge vastuse tingimuste täpsustamise ja paranduse sekventside põhjal, samuti eeldasime erinevuste võimalikkust rituaalisekventsides.

2. Materjal

2.1. Dialoogid

Eesti dialoogikorpus sisaldab 22 "võlur Ozi" tehnikaga kogutud dialoogi, millest üks on jäetud analüüsist välja (katseisik üritas programmi n-ö kokku jooksutada ja esitas keelelises mõttes ebaselgeid küsimusi). 80 suulist infodialoogi oleme samuti valinud Eesti dialoogikorpusest, kuhu kuuluvad telefoni-, bussi-, reisiinfovestlused ja mõned teised institutsionaalse vestluse alaliigid. Infotelefoni vestlused on valitud võrdluse ühtluse huvides: need on pikkuselt ja vestle-

[1] Artikli valmimist on toetanud ETF (grant 5685) ning HTM (riiklik sihtprogramm "Eesti keel ja rahvuslik mälu").

jate rollijaotuselt võrreldavad simuleeritud dialoogidega. Suhtlejateks on infoandja ning infoküsija, suhtluse keskmes on infopäring, suhtlus on ametlik ja teemakeskne. Erinevalt "võlur Ozi" dialoogidest, kus infoandjale oli oluline suhtlust üleval hoida, ei paku infotelefoni töötaja uut teemat ning enamik kõnesid piirdub ühe päringuga. Infotelefoni dialoogid on kogutud telefonikõnede lindistamise teel. "Võlur Ozi" tehnikat tutvustame üksikasjalikumalt allpool. Suulised dialoogid on kirja pandud Jeffersoni transkriptsioonis, kirjalikes "võlur Ozi" dialoogides oleme säilitanud osalejate ortograafia.

2.2. "Võlur Ozi" tehnika

Paljud teadlased on seisukohal, et suhtlemine inimese ja arvuti vahel peab olema võimalikult sarnane inimestevahelise suhtlusega. Mitmete katsete põhjal on aga järeldatud, et inimene suhtleb arvutiga teisiti kui inimesega. Need katsed olid viidud läbi nn "võlur Ozi" meetodil (lühend WOZ – "Wizard of Oz", vt Dahlbäck jt 1998, Valdisoo, Vutt 2002). Teatud isikutel palutakse justkui testida programmi, millele saab esitada loomulikus keeles küsimusi ning mis annab õigeid vastuseid. Tegelikkuses on aga suhtluspartneriks arvutivõrgu kaudu teine inimene ("võlur"). Sellisel teel kogutakse kindla valdkonna dialooge ning moodustatakse neist uuringuteks kasutatav dialoogikorpus.

Katsete läbiviimisel on oluline silmas pidada mõningaid asjaolusid, mille poolest arvuti erineb inimesest ning mis on eksperimentide käigus tekitanud probleeme:

a) inimesed loovad kõnelemise käigus uusi kõneüksusi, arvuti suhtleb valmislausete ja -vormelitega;
b) inimesed trükivad aeglaselt, arvuti aga väljastab teksti kiiresti;
c) arvutil ei teki iialgi väikesi eksimusi (nt juhuslik õigekirjaviga), inimesed teevad neid pidevalt.

Olenemata katseisikutest, tekib "võlur Ozi" tehnika kasutamisel probleem veel sellest, et kui inimest on palutud midagi testima ja talle isiklikult pole sellest mingit kasu, siis ta ei tarvitse suhtuda katsesse täie tõsidusega. Selle probleemi lahendamiseks soovitatakse valida katseisikuteks inimesi, kellele antud valdkond huvi pakub. Seetõttu oli meil ainevaldkonnaks valitud reisiinfo: reisimine on meeldiv tegevus, mille vastu tunneb inimene tõenäoliselt vähemalt mingil määral huvi.

Võrreldes loomulike suuliste dialoogidega, on WOZ dialoogidel iseärasused, millest kahe esimese põhjuseks on simuleeritud, kunstlik dialoog ning kahe järgneva põhjuseks asjaolu, et suhtlus toimub kirjalikult ja reaalajas:

- ühe dialoogi jooksul algatatakse mitu teemat, loomulikes dialoogides on üldreeglina vaid üks teema;
- arvuti pakub uut teemat, loomulikes dialoogides ei ürita infoandja vestlust üleval hoida uue teema pakkumisega;
- arvuti annab infot portsjonitena, infoüksuste vahel ei esine jätkajaid (*mhmm*, *jah*, *jaa* jms);
- ei esine mittekuulmist, küll aga üheaegset trükkimist (arvuti ja infoklient trükivad samaaegselt kumbki oma teksti, vt näide 1).

(1)
Arvuti: Kas Teid huvitavad ka kohalejõudmise ajad | KYE: SULETUD KAS | | VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |
Infoklient: ega te ei oska öelda millal buss võrru jõuab | KYE: AVATUD |
Infoklient: huvitab küll | KYJ: JAH | | VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE |

Seni oleme WOZ dialooge kogunud telneti-laadse programmi abil, mis oli kasutajale lihtne, kuid nõudis äärmist tähelepanu "võlurilt". Tartu Ülikoolis loodi 2004. a kevadel õppeaine "Tarkvaraprojekt" raames rühma üliõpilaste poolt uus eestikeelne programm WOZ dialoogide kogumiseks. Programm on lihtsasti käsitsetav ja võlurisõbralik – mitmete kiirotsingute, tüüpvastuste ning -šabloonide abil on vastuste moodustamine oluliselt hõlpsam ja kiirem. Hetkel pole seda programmi veel kasutatud.

3. Loomulike ja WOZ dialoogide võrdlus

Järgnevas vaatame, missugused erinevused on rituaalides ning kuidas käitub infoandja vastuse tingimuste täpsustamise ja paranduse sekventsides.

3.1. Rituaalid

Rituaalid esinevad tavaliselt dialoogi alguses ja lõpus. Dialoogikorpuse märgendamisel tähistatakse rituaalse naaberpaari esiliiget tähekombinatsiooniga RIE ning järelliiget RIJ (märgendite tüpoloogia vt Lisa 2 ja Hennoste, Rääbis 2004: 42–50).

21 WOZ dialoogis esines 39 rituaali: RIE: TÄNAN 15, RIJ: PALUN 14, RIE: TERVITUS 3, RIJ: VASTUTERVITUS 3, RIE: LÕPUSIGNAAL 3, RIJ: LÕPETAMISE VASTUVÕTMINE 1.

Enamikus WOZ dialoogides alustati vestlust ilma rituaalideta, küsimuse või sooviga:

(2)
Infoklient: Mis kell väljub buss Ardust Tallinnasse, et jõuaks laevaga Helsinkisse kell 12? | KYE: AVATUD |
Infoklient: Tahaks sõita Võrust tallinna | DIE: SOOV |

Märkimist väärib asjaolu, et mitte üheski meie korpusesse kuuluvas WOZ dialoogis ei esine hüvastijätte. Põhiline vestluse lõpetamise viis oli lihtsalt tänamine:

(3)
Infoklient: tänan | RIE: TÄNAN |
Arvuti: Palun. | RIJ: PALUN |

80 loomulikus dialoogis esines rituaale aga tunduvalt enam – 499, sh RIE: TERVITUS 80, RIE: KUTSUNG 79, RIJ: KUTSUNGI VASTUVÕTMINE 79, RIE: TÄNAN 79, RIJ: PALUN 71, RIJ: VASTUTERVITUS 74, RIE: HÜVASTIJÄTT 15, RIE: SOOVIMINE 6, RIJ: VASTUHÜVASTIJÄTT 5, RIE: LÕPUSIGNAAL 5, RIJ: LÕPETAMISE VASTUVÕTMINE 5, RIJ: VASTUSOOVIMINE 1.

Dialoogid algavad harilikult kutsungi, selle vastuvõtmise, vastuvõtja tervituse ja enesetutvustuse ning helistajapoolse tervituse ja küsimuse või sooviga.

(4)
((kutsung)) | RIE: KUTSUNG |
V: `info`telefon= | RIJ: KUTSUNGI VASTUVÕTMINE | | RY: TUTVUSTUS |
Kersti= | RY: TUTVUSTUS |
tere? | RIE: TERVITUS |
H: tere=päevast, | RIJ: VASTUTERVITUS |
ma soovin `Dentese ambaravi Rop (.) Mõ- `Ropka=`Mõisa. | DIE: SOOV |

Sarnaselt WOZ dialoogidega lõpetatakse loomulikes suulistes dialoogides vestlus enamasti vaid tänamisega.

(5)
H: aitäh? | RIE: TÄNAN |
V: palun | RIJ: PALUN |

Analüüsitud loomulike ja WOZ dialoogide põhjal võib öelda, et info küsimisel käitutakse suhtluse alguses ning lõpus üsna sarnaselt, sõl-

tumata sellest, kas vestlus toimub suuliselt (telefoni teel) või kirjalikult ning kas vastajaks on inimene või "arvuti".

3.2. Vastuse tingimuste täpsustamine

Infodialoogide oluliseks iseloomustajaks on kliendi ja infoandja roll küsimuste esitamisel – dialoogi käigus esitab enamiku küsimusi/soove klient. Ent kui esitatud küsimus on liiga üldine või ebatäpne, võtab vastaja initsiatiivi üle ja esitab täpsustavaid küsimusi, mida nimetatakse vastuse tingimuste täpsustamiseks ning kus naaberpaari esiliiget tähistab tähekombinatsioon VTE ja järelliiget VTJ (Hennoste, Rääbis 2004: 65).

WOZ dialoogides esines vastuse tingimuste täpsustamist 76 korral. Kuna VTE ja VTJ märgivad küsimuse sisulist poolt, siis märgendamisjuhendi kohaselt (Hennoste, Rääbis 2004) tuleb sellistele lausungitele panna kaks märgendit – sisuline ja vormiline (vt tabel 1).

Tabel 1. WOZ dialoogide vastuse tingimuste täpsustamise vormilised märgendid

VTE vormiline märgend	Arv	VTJ vormiline märgend	Arv
KYE: JUTUSTAV KAS	28	KYJ: INFO ANDMINE	36
KYE: SULETUD KAS	28	KYJ: JAH + DIE: SOOV	15
KYE: AVATUD	9	KYJ: JAH	9
KYE: JUTUSTAV KAS + TVJ: VASTUVÕTMINE	5	KYJ: EI	7
DIE: SOOV	2	DIJ: INFO ANDMINE	2
KYE: ALTERNATIIV	1	KYJ: ALTERNATIIV: ÜKS	1
DIE: SOOV + TVJ: VASTUVÕTMINE	1	DIJ: INFO ANDMINE + KYE: JUTUSTAV KAS	1
KYE: AVATUD + TVJ: VASTUVÕTMINE	1	DIJ: MUU	1
KYE: SULETUD + TVJ: VASTUVÕTMINE	1		

WOZ dialoogides täpsustas arvuti enamikel juhtudel reisi algusaega ja nädalapäeva.

(6)
Arvuti: Kas Teid huvitab mingi konkreetne ajavahemik? | KYE: JUTUSTAV KAS | | VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |
Infoklient: 10 ja 15 vahel | KYJ: INFO ANDMINE | | VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE |

Loomulikes dialoogides ollakse vastuse tingimuste täpsustamistega oluliselt ökonoomsemad – kokku esines 35 VTE-d 80 dialoogis. Põhjusi on mitmeid:

• WOZ dialoogide teema on reisimine, mis tekitab paratamatult võimaluse liiga üldise ja väga suure vastusemahuga küsimuste esitamiseks. Infotelefonile helistades püüavad inimesed formuleerida oma küsimuse nii, et sellele oleks võimalikult vähe vastuseid (tavaliselt on vastuseks mingi asutuse kontaktandmed).

• WOZ dialoogides esitasid katseisikud küsimusi üsna suvaliselt, loomulikes dialoogides aga kindla sooviga saada vajalik vastus.

• WOZ dialoogide "võlur" polnud õppinud klienditeenindamist, infotelefoni töötajatel aga on mõningaid teadmisi ja kogemusi info lihtsamaks leidmiseks ning klientidest arusaamiseks.

Loomulikes dialoogides esitatakse täpsustavaid küsimusi üpris sageli kliendi küsimuse hüpoteetilise tõlgenduse vormis (vrd tabel 2).

Tabel 2. Loomulike dialoogide vastuse tingimuste täpsustamise vormilised märgendid

VTE vormiline märgend	Arv	VTJ vormiline märgend	Arv
KYE: VASTUST PAKKUV	20	KYJ: JAH	21
KYE: ALTERNATIIV	9	KYJ: INFO ANDMINE	7
KYE: SULETUD KAS	8	KYJ: ALTERNATIIV: ÜKS	7
DIE: PAKKUMINE	6	DIJ: NÕUSTUMINE	5
KYE: AVATUD	5	KYJ: JAH + IL: TÄPSUSTAMINE	4
DIE: PAKKUMINE + TVJ: VASTUVÕTMINE	3	KYJ: EI	2
KYE: JUTUSTAV KAS	2	DIJ: MITTENÕUSTUMINE	2
KYE: VASTUST PAKKUV + TVJ: VASTUVÕTMINE	1	KYJ: NÕUSTUV EI	1
KYE: MUU	1	KYJ: MUU	1
		KYJ: ALTERNATIIV: MÕLEMAD + KYE: AVATUD	1
		KYJ: EI + IL: TÄPSUSTAMINE	1

(7)
H: [te:re?] | RIJ: VASTUTERVITUS |
(.) tahtsin=`küsida et sellel=e `Tele kahel on see mingi `infotelefon kus saab e `elistada ja sealt `kõike `arveid ja `kõike õ küsi[da] | DIE: SOOV |
V: [o]pe`raator`teenus [jah.] | KYE: VASTUST PAKKUV | | KYE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |
H: [jah.] | KYJ: JAH |

Loomulike ja WOZ dialoogide paralleelse uurimise käigus selgus veel mõndagi huvitavat:

• WOZ dialoogides esitatakse küsimused täislausetena, loomulikus suhtluses üldjuhul hüpoteesi pakkuva fraasina;

• WOZ dialoogides täpsustab arvuti kliendipoolset küsimust üksikute punktide kaupa (*Kas Teid huvitab mingi konkreetne nädalapäev?*), loomulik vestlus on aga paindlikum ning püüab ökonoomsemalt täita mitut arusaamatut/puudulikku punkti vajaliku info leidmiseks:

(8)
H: ega ˊteie ei oska kogemata öelda Tallinna ˊbusside aegu. | KYE: JUTUSTAV KAS |
V: ja ˊmilliseid busse. | KYE: AVATUD | | VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |
(.)
H: ee (.) noh millal lähevad ˊviimased ˋkiirbussid täna ˊTallinna. | KYJ: INFO ANDMINE | | VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE |
(.)
V: kell=ˊõheksa on viimane ekspress. | KYJ: INFO ANDMINE |

3.3. Parandused

Lisaks vastuse tingimuste täpsustamisele on suhtluse õnnestumise seisukohalt samavõrra olulised parandussekventsid, kus lahendatakse ettetulevaid suhtlusprobleeme. Paranduslõik koosneb ühe kõneleja paranduse algatusest (tüüpiliselt küsimusest, ka direktiivist) ja teise kõneleja paranduse läbiviimisest (vastusest), vastavad tähekombinatsioonid on meie märgendusskeemis PPE ja PPJ. Vabatahtliku reaktsioonina võib lisanduda paranduse hindamine, mis osutab, et parandus on piisav (Hennoste, Rääbis 2004: 56, 139). Võiks oletada, et loomulikus dialoogis on valdkonna ulatuslikkuse ja suulise läbiviimise tõttu rohkem suhtlusprobleeme ning infoandja on sunnitud vastavalt rohkem parandusi algatama. Tegelikkus näitab aga vastupidist. 80 suulises dialoogis leidub 46 infoküsija ja 41 infoandja algatatud parandust. 21 WOZ dialoogis on parandusi 29, neist infoandja algatatud parandusi 28 (vt tabel 3). WOZ testijad-infoküsijad olid mõistagi vähem motiveeritud täpset infot taotlema või andma, sest nende päringuid ei põhjustanud otsene infovajadus. See seletab infoandja algatatud paranduste rohkust ja infoküsija algatatud paranduste vähesust. Infoandja on aga tihtipeale kasutanud parandussekventsi hoopis aja võitmiseks, et infootsinguks vajalik paus liiga pikaks ei veniks.

Tabel 3. Infoandja algatatud parandused loomulikes ja WOZ dialoogides

Paranduse liik	WOZ dialoog	Loomulik dialoog
Mittemõistmine	25	8
Ümbersõnastamine	3	12
Üleküsimine	–	21
Kokku	28	41

3.3.1. Mittemõistmine. Mittemõistmine on selline paranduse algatus, millega antakse teada, et vestluskaaslase eelnev info on jäänud kuulmata või mõistmata (Hennoste, Rääbis 2004: 59). Harilikult vormistatakse mittemõistmist avatud küsimusena (kuidas), harvem direktiivina (*Ropka-Mõisa palun korrake*). WOZ dialoogides on mittemõistmised põhiline ja praktiliselt ainus paranduse algatus (25 juhtu 28st). Nende vormistamiseks on kasutatud kindlaid vormeleid, nagu *Kuidas, palun?* (12), *Täpsustage siht/lähtepeatuse nimi* (8), *Ma ei saanud aru* (2), *esitage oma küsimus palun uuesti* (2), *palun täpsustage* (1). Enamik neist eeldab raskusi põhjustanud vooru ümbersõnastamist arusaadavamal viisil ning on tingitud eelmise vooru puudulikkusest: ebakindla deiktiku kasutamisest (*sel ajal* näites 9), küsimuse elliptilisusest (näide 10).

(9)
Infoklient: Kas sel ajal saab Tartust edasi? | KYE: JUTUSTAV KAS |
Arvuti: Kuidas, palun? | KYE: AVATUD | | PPE: MITTEMÕISTMINE |
Infoklient: Kas peale kella 22.35 on võimalik sõita Tartust Viljandisse? | KYJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE | | KYE: JUTUSTAV KAS |

(10)
Infoklient: Aga nädalavahetusel. | KYE: AVATUD |
Arvuti: kuidas, palun? | KYE: AVATUD | | PPE: MITTEMÕISTMINE |
Infoklient: Kuidas käivad Tartust Tallinnasse bussid nädalavahetusel. | KYJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE | | KYE: AVATUD |

Avatud küsimused ja direktiivid, mis eeldavad vastusena eelmise vooru kordust (nt *Kuidas, palun?*, *Ma ei saanud aru*), on efektiivsed just lühivoorude järel: palvele korrata järgneb täislauseline vastus, mis hõlmab soovitud võtmesõnu (siht-, lähtekoht, ajavahemik). Kui tegemist on täislausega, mida palutakse korrata (näide 11), siis on tihtipeale tulemuseks veelgi segasem lause: inimene saab aru, et eeldatakse ümbersõnastamist, kuid ei tea täpselt, mida ta peaks ümber sõnastama ja üritab muuta probleemvoorus võimalikult palju. Automaatse tuvastuse seisukohalt ei ole see hea: tuvastajal tuleks vaeva

näha kliendi uue täislausega sõltumata sellest, kas mõistmata jäi terve eelmine lause või üksnes kindel fraas.

(11)
Infoklient: Kuidas on võimalik kõige kiiremini jõuda Tartust Kuressaarde? | KYE: AVATUD |
Arvuti: Esitage oma küsimus palun uuesti. | DIE: SOOV | | PPE: MITTEMÕISTMINE |
Infoklient: Milliste liiklusvahenditega on võimalik jõuda Tartust Kuressaarde? | DIJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE | | KYE: AVATUD |

Mõnikord on WOZ dialoogis testija kordamispalve järel ka loobunud küsimuse uuesti esitamisest – kas oskamatusest vooru ümber sõnastada või arvates selle olevat kasutu – ning pakkunud uut teemat (näide 12).

(12)
Infoklient: kas on veel mõned peatumis kohad? | KYE: JUTUSTAV KAS |
Arvuti: Esitage oma küsimus palun uuesti. | DIE: SOOV | | PPE: MITTEMÕISTMINE |
Infoklient: tahaksin teada ka mis kella aegadel läheb helsingisse laev | KYE: AVATUD | | TVE: PAKKUMINE |

Loomulikes dialoogides on infoandja mittemõistmisi märkimisväärselt vähe (kõigest 8 juhtu). Üldiselt on tegu sellega, et eelmises (pikas, keeruliselt sõnastatud või pealerääkimisega) voorus läheb oluline info (aadress, nimi) infoandjale kaduma. Parandusalgatused on vormistatud lünkküsimustena, mis lokaliseerivad mõistmata jäänud osa eelmisest voorust üsnagi täpselt. Paranduse läbiviijal ei ole seega vaja korrata tervet vooru, vaid üksnes täita küsimuse lünk (näited 13–14).

(13)
ma soovin 'Dentese ambaravi Rop (.) Mõ- 'Ropka='Mõisa. | DIE: SOOV |
(0.8)
V: .h e Ropka='Mõisa palun 'korrake. | DIE: SOOV | | PPE: MITTEMÕISTMINE |
H: Ropka=Mõisa A'dentese hambaravi. | DIJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE |

(14)
(.) 'tahtsin küsida 'Tartus=e (.) 'Kalda tie 'kolmkümmend (.) 'pleki 'ukse 'koda (.) kas te 'saate mu anda. | KYE: JUTUSTAV KAS |
(1.2) on sellised (.) {teil või} £ | KYE: SULETUD KAS |
(4.2)
V: ja 'aadress oli 'Kalda 'tee? | KYE: AVATUD | | PPE: MITTEMÕISTMINE |
(.)
H: £ 'kolmkend kui ma 'õieti 'mäletan. £ | KYJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE |
(14.5)

Tundub, et lünkküsimused on kordamispalvetest nii ökonoomsemad (vastus on terviklause asemel fraasipikkune, mõnikord reservatsiooniga, nagu näites 14) kui ka efektiivsemad (vastaja ei ürita kogu lauset ümber sõnastada, mõistatades, mis on raskust põhjustanud). Ka dialoogsüsteem võiks sarnastel juhtudel kasutada lünkküsimusi, reserveerides palved korrata juhtudeks, kus tõesti ei ole tervest lausest aru saadud. WOZ dialoogides kandsid sellist strateegiat fraasid *Palun täpsustage siht/lähtepeatuse nimi*, millele vastaja andis lühikese ja adekvaatse vastuse (näide 15).

(15)
Infoklient: Kuidas saab sõita Tartust Kioki | KYE: AVATUD | | TVE: PAKKUMINE |
Arvuti: Täpsustage sihtpeatuse nimi, palun! | DIE: SOOV | | TVJ: VASTUVÕTMINE | | PPE: MITTEMÕISTMINE |
Infoklient: Koigi | DIJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE |
Arvuti: Peatust Koigi meie andmebaasis pole! | KYJ: INFO PUUDUMINE |

3.3.2. Üleküsimine. Üleküsimine esineb ainult loomulikes dialoogides. Üldiselt on infoandja algatatud üleküsimised eelmise vooru osa kontrollivad kordused, millele vastatakse kinnitusega (Hennoste, Rääbis 2004: 56). Vahel küsitakse üle ka kaugemal eespool vestluses mainitud info. Tõenäoliselt puuduvad üleküsimised WOZ dialoogides seetõttu, et kirjalikus suhtluses saab kirjapandud lauseid uuesti läbi lugeda, suulistes dialoogides aga ei ole infoandjal võimalust eelmist vooru uuesti kuulata. Tavaliselt korratakse olulisimat osa voorust, nagu telefoninumber, aadress, firma nimi. Küsimus on vormistatud fraasikordusena, millele võib lisanduda *jah/või*. Üleküsimine on ökonoomseim paranduse algatuse vahend: sellele järgneb partikliga *jah* või *mhmh* vormistatud paranduse läbiviimine (näide 16), üksikjuhtudel ka *ei*, mida laiendab täpsustav lausung (meie võrdluskorpuses näidet ei leidu, näide 17 pärineb Eesti dialoogikorpusest).

(16)
kas te ´oskate öelda ´Tartus ´Kagu intressite Kagu=vest=ee ´aadressi.
| KYE: JUTUSTAV KAS |
V: Kagu In´vest. (.) | KYE: VASTUST PAKKUV | | PPE: ÜLEKÜSIMINE |
H: jah | KYJ: JAH | | PPJ: LÄBIVIIMINE |

(17)
((reisibüroo dialoog, räägitakse reisija lapse vanusest))
H: kümneaastane. | PPJ: LÄBIVIIMINE | | KYJ: INFO ANDMINE |
(.)
V: e oli ´viie. | PPE: ÜLEKÜSIMINE | | KYE: VASTUST PAKKUV |
(0.5)

H: ei, | PPJ: LÄBIVIIMINE | | KYJ: EI |
'kümneaastane. | PPJ: LÄBIVIIMINE | | KYJ: INFO ANDMINE |

3.3.3. Ümbersõnastamine. Ümbersõnastamisega on tegemist siis, kui paranduse algataja pakub oma tõlgenduse (hüpoteesi, järelduse vms) eelmisele voorule, soovides kinnitust, et ta on vestluskaaslasest õigesti aru saanud (Hennoste, Rääbis 2004: 58). Infoandja algatatud ümbersõnastamist esineb nii simuleeritud (3 juhtu) kui loomulikes dialoogides (12 juhtu, millest kolm kuuluvad samasse keerukasse sekventsi).

Ümbersõnastamisi algatatakse vestluskaaslase kavatsuste teadasaamiseks (näites 18 eelneb ümbersõnastamisele kaudne *kas*-küsimus reisiaegadest *kui palju laevu liigub X marsruudil*), eksimuste parandamiseks (klient pakub ehitusfirma nimeks Ekaruse asemel Ekaras ja Ekaros), liiga üldise ja segase küsimuse kitsendamiseks. WOZ dialoogides vormistatakse ümbersõnastamine täislausena, loomulikes dialoogides aga fraasina, millele võivad lisanduda küsipartiklid *jah/või*.

(18)
Infoklient: kui palju liigub laevu marsruudil Tallinn–Helsingi | KYE: AVATUD | TVE: PAKKUMINE |
Arvuti: Kas Teid huvitab reiside arv? | KYE: JUTUSTAV KAS | | TVJ: VASTUVÕTMINE | | PPE: ÜMBERSÕNASTAMINE |
Infoklient: Ei, | KYJ: EI | | PPJ: LÄBIVIIMINE |
kellaajad? | KYJ: INFO ANDMINE | | PPJ: LÄBIVIIMINE |
Arvuti: Kas Teid huvitab mingi konkreetne ajavahemik? | KYE: JUTUSTAV KAS | | VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE |

(19)
H: a 'ütelge palun: vot=e firma (.) 'Ekoros (.) 'Ekaras või, £ | DIE: SOOV |
V: 'Ekarus. | KYE: VASTUST PAKKUV | | PPE: ÜMBERSÕNASTAMINE |
(.)
H: £ jah. | KYJ: JAH | | PPJ: LÄBIVIIMINE |
see (.) ehitus'firma, (.) ehitus'materjalid. | IL: SELETAMINE |

Ümbersõnastamine on lingvistiliselt markeeritum sedavõrd, kuivõrd hüpotees pole ilmne (vt voor 2 näitest 20). Kui hüpotees ei osutu õigeks, ei muuda infoandja oma strateegiat, vaid esitab uusi hüpoteese (voorud 4 ja 6 samast näitest), mis on järjest mitmesõnalisemad ja eksplitsiitsemad (*te mõtlete, te tahate*). Strateegiat muutmata jõuab infoandja lõpuks soovitud vastuseni, mis kitsendab infoandmisvaldkonda (voor 9).

(20)
1. H: mul on siuke kuradi proble´maatiline küsimus=t ee äkki te oskate mulle öelda mõne {-} ´veoauto: numbrid et mul oleks vaja kuradi ´metalli ära viia {-} ühtegi ´numbrit=tead | DIE: SOOV |
(4.0)
2. V: >täendab< ´veoauto ´renti siis=või. | KYE: VASTUST PAKKUV | | PPE: ÜMBERSÕNASTAMINE |
3. H: ei ma: mina ´rentida ei ´taha | KYJ: EI | | PPJ: LÄBIVIIMINE |
ma tahan et tulek ks ´ise töötaja=noh. | IL: TÄPSUSTAMINE || PPJ: LÄBIVIIMINE |
(0.8) et ma ´autojuhile ´maksad. | IL: TÄPSUSTAMINE | | PPJ: LÄBIVIIMINE |
4. V: ´täendab te ´mõtlete me´talli kokku´ostjaid=või, | KYE: VASTUST PAKKUV | | PPE: ÜMBERSÕNASTAMINE |
5. H: ei ma kok- kok- | KYJ: EI | | PPJ: LÄBIVIIMINE |
ma tahan et kes mul ´ära viiks=tead=noh, | IL: TÄPSUSTAMINE || PPJ: LÄBIVIIMINE |
6. V: no selles mõttes ´ongi=et t te tahate ju ´ära anda sda metalli=ikka=et ned me´talli ´firmad kes kokku ostavad, (0.8) või [kui]das | KYE: VASTUST PAKKUV | | PPE: ÜMBERSÕNASTAMINE |
7. H: {-} ei ´ei | KYJ: EI | | PPJ: LÄBIVIIMINE |
ma tahan: ´veoauto ´veoautot tahan | IL: TÄPSUSTAMINE || PPJ: LÄBIVIIMINE |
8. X: ((kõrvalt öeldakse ette)) ´veo´takso noh | IL: TÄPSUSTAMINE |
9. H: või mingi ´takso ta-takso taoline kes võtaks ´peale need {no} | IL: TÄPSUSTAMINE |

See, et loomulike dialoogide paranduste hulgas on ülekaalus üleküsimine ja ümbersõnastamine, väljendab vestluskaaslaste kooperatiivsust: suhtlushäirete puhul kasutab infoandja kliendilt juba saadud infot, nõudmata talt kogu vooru uuesti esitamist. Hüpoteeside tegemisel toetub ta tõenäoliselt oma töökogemusele, pakkudes välja tõenäolisemaid lahendusi. Statistiliselt tõenäolisemate hüpoteeside tegemine, millele oodatakse kinnitust, osutub palvest korrata ökonoomsemaks ja efektiivsemaks just tänu infoandja kogemuse olemasolule. Tõenäoliselt peaks ka dialoogsüsteem õppima juba peetud infokõnedest, kogudes päringute statistikat.

Paranduse vormilistes märgendites (tabelid 4 ja 5) paistavad WOZ ja loomuliku dialoogi erinevused samuti hästi välja.

Tabel 4. WOZ dialoogide paranduse vormilised märgendid

PPE vormiline märgend	Arv	PPJ vormiline märgend	Arv
KYE: AVATUD	14	DIJ: INFO ANDMINE	7
DIE: SOOV	8	DIJ: INFO ANDMINE + KYE: AVATUD	1
KYE: SULETUD KAS	2	DIJ: INFO ANDMINE + KYJ: JUTUSTAV KAS	3
DIE: SOOV + TVJ: VASTUVÕTMINE	3	KYJ: INFO ANDMINE + KYE: AVATUD	3

PPE vormiline märgend	Arv	PPJ vormiline märgend	Arv
KYE: JUTUSTAV KAS	1	KYJ: INFO ANDMINE + KYE: JUTUSTAV KAS	4
		DIJ: MUU	1
		KYJ: EI	1
		KYJ: JAH	2
		KYJ: INFO ANDMINE	4
		KYJ: INFO ANDMINE + SEE: VÄIDE	1

Prototüüpne parandussekvents WOZ dialoogis on selline:

Infoklient: infoküsimus või infoandmine
1. KYE: AVATUD/DIE: SOOV
2. INFO ANDMINE + infoküsimuse kordus
Arvuti: vastus infoküsimusele või VTE

Tabel 5. Loomulike dialoogide paranduse vormilised märgendid

PPE vormiline märgend	Arv	PPJ vormiline märgend	Arv
KYE: VASTUST PAKKUV	32	KYJ: JAH	27
KYE: AVATUD	7	KYJ: INFO ANDMINE	6
KYE: SULETUD KAS	1	KYJ: EI + IL: TÄPSUSTAMINE	3
DIE: SOOV	1	KYJ: NÕUSTUV EI	2
		KYJ: INFO ANDMINE + KYE: VASTUST PAKKUV	1
		DIJ: INFO ANDMINE	1

Prototüüpne parandussekvents loomulikus dialoogis on lühem:

Helistaja: infoküsimus või infoandmine
1. KYE: VASTUST PAKKUV
2. KYJ: JAH
Vastaja: vastus infoküsimusele või VTE

4. Kokkuvõtteks

Loomulikes ja WOZ dialoogides kasutavad infoandjad samades allsekventsides suhtluseesmärgi saavutamiseks mõnevõrra erinevaid vahendeid. Üllatusena mõjus tõdemus, et loomulike dialoogide allsekventsides ei kasutata täpsustavaid avatud küsimusi, vaid pakutakse välja oletusi, millele oodatakse kinnitust. Oletuste strateegia osutub efektiivseks infoandja pikaaegse kogemuse arvessevõtmise tõttu. Mitmed olulised erinevused dialoogides on tingitud suhtluska-

nalist (kirjalik–suuline). Seega on vaja hakata koguma ka suulisi WOZ dialooge, mis nõuab üsna põhjalikku ettevalmistust, katsetamist ning võluri(te) koolitamist. Tulevane võlur võiks kasutada neid efektiivseid ja ökonoomseid strateegiaid, mida infoandja suulises kõnes järgib.

Kirjandus

Dahlbäck, Nils, Jönsson, Arne, Ahrenberg, Lars 1998. Wizard of Oz studies – Why and how. – http://www.ida.liu.se/~arnjo/papers/KBS.ps

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Valdisoo, Maret, Vutt, Evely 2002. "Võlur Ozi" tehnika ja eesti keeles suhtlev arvuti. – A&A 5, 63–67.

Dialoogsüsteemid – kuupäevade tuvastamine ja vastusemallid[1]

Margus Treumuth
Tartu Ülikool

1. Dialoogsüsteemide areng

Automatiseeritud teenused on saanud elu normiks. Info on arvuti vahendusel kõigile kättesaadav kodust lahkumata, kuid inimeste soovid kasvavad – oleks tore suhelda arvutiga inimkeeles, oleks tore, kui telefonikõnedele oskaks vastata inimkeelt mõistev arvuti. Siis piisaks telefonikõnest, et esitada päring andmebaasile või lasta endale e-posti ette lugeda. Paljudele meeldiks võimalus loobuda tarkvara kasutusjuhendite lugemisest ja pöörduda arvuti poole, sõnastades ülesande vaba tekstiga emakeeles.

Alljärgnevalt antakse ülevaade dialoogsüsteemide arengust, käsitletakse dialoogsüsteemide koostamist teatriinfosüsteemi näitel ning kirjeldatakse üldisi alamülesandeid nagu kuupäevade tuvastamine ja vastusemallide moodustamine.

1.1. Vestlused kohvilaua juures

Dialoogsüsteemi all mõistame arvutiprogrammi, mis suudab inimesega suhelda, kusjuures suhtlus toimub inimkeeles kõne või teksti vahendusel.

Leidub hulgaliselt meelelahutuslikke dialoogsüsteeme (nt soovitan teha Google'i otsing: "chatbot"), mille kasutamine toimub klaviatuuri vahendusel. Need on vestlusprogrammid (juturobotid), mida pahaaimamatu vestluspartner võib mõnda aega isegi reaalseks inimeseks pidada (Nass 2005). Põhiliselt tuleb aga juturobot toime üksnes kuuldu ümbersõnastamisega ja universaalsete käibefraaside kasutamisega. Klaviatuuri vahendusel ennast väljendada ei ole kuigi mugav, liiati kui arvuti vastused on vaid asjatu müra ega aita ini-

[1] Tööd on toetanud Eesti Teadusfond (grant nr 5685) ning HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

mest. Arvuti peaks suutma vesteldes pakkuda uut infot, mitte ainult ootama kasutajapoolset aktiivsust.

On raske uskuda, et keegi sooviks vestelda arvutiga niisama ajaviiteks kohvilaua taga. Kui siiski, peaks arvuti sellisel juhul omama emotsioone, huvitavaid maailmavaatelisi seisukohti jms. Arvutiga peaks olema huvitav lobiseda. Paraku ei pakkunud huvitavat vestlust mitte ükski dialoogsüsteem, mida eespool soovitatud otsinguga leidsin ja uurisin. Võime küll modelleerida inimest, kuid – olgem realistlikud – tulemuseks saab olla ikkagi vaid masin.

1.2. Uuesti asjalikuks

Jätame eelmainitud meelelahutuslikud dialoogsüsteemid ootama sõbralikku publikut ja proovime kavandada masina, mis on ühes kitsas tegevusvaldkonnas piisavalt "intelligentne" ja oskab inimest aidata. Seeläbi leiame ka praktilise õigustuse inimese modelleerimiseks.

Tulemuseks peaksime saama kasutuskõlbliku juturoboti, kes vastab sõbralikult meie telefonikõnedele, ei vihastu, ei solvu, ei väsi (nagu võib juhtuda inimesega). Suuremate keelte puhul (nt inglise jmt) on see saavutatud (Manaris 1998). Infotelefoniteenust osutavaid roboteid leidub näiteks http://www.voicerobots.com/. Firma AT&T pakub võimalust telefoni teel e-posti lugeda (AT&T VoiceTone).

Dialoogsüsteemide atraktiivsus on kasvanud kõnetuvastuse käibeletulekuga, s.t piisab masinaga rääkimisest, klaviatuur jääb puutumata. Kõnetuvastus on ka eesti keele jaoks saamas reaalsuseks (Alumäe 2005) ning seegi annab põhjuse jätkata dialoogsüsteemide arendamist. Eesti keele kõnesünteesi (http://www.phon.ioc.ee/) integreerimist dialoogsüsteemisse ongi juba proovitud (vt Reisiagent http://www.dialoogid.ee/).

1.3. Ideest teostuseni

Millised on need komponendid, millest võiks inimkeeles suhtlev masin koosneda? Millise valdkonna eksperdina võiks selline masin ennast õigustada?

Vestlusprogrammi tegevusvaldkonna kitsendamiseks sobivad lihtsa näitena kõikvõimalikud ajatabelid – teatri-, kino-, telekavad, rongi-, lennu-, bussigraafikud. Sedalaadi info jagamisega inimkeeles peaks masin hakkama saama. Ajatabelites ja graafikutes info otsimi-

ne ning selle inimkeeles väljastamine on üks põhilisi praktilisi valdkondi, kuhu dialoogsüsteemide loojad on keskendunud.

Ajatabelid kajastavad teatud ajahetkedel toimuvaid sündmusi. Neid saab lihtsasti kirjeldada kas ühesainsas tabelis või mõnes omavahel seotud tabelis. Ka sündmust kirjeldavate tunnuste hulka on võimalik hoida üsna väiksena. Tunnusteks võivad olla näiteks sündmuse nimetus, toimumise aeg, kestus, kuuluvus teatud gruppi ja toimumise koht.

Järgnevalt keskendume tarkvaratehnilistele aspektidele taolise masina loomisel, mis suudaks aru saada meie soovist ning leida meid huvitava info masina käsutuses olevast ajatabelist.

2. Dialoogsüsteemi loomine

Vaatame lihtsa dialoogsüsteemi ehitamist, mis on liideseks teatriinfo andmebaasile. Kui helistame infotelefonile, sooviga välja selgitada, missugused etendused on parajasti Eesti teatrites, siis meile võiks vastata ametniku asemel see süsteem.

2.1. Nõuete analüüs

Ülesande lahendamist tuleks alati alustada nõuete analüüsist. Püüdkem kirjeldada loodava toote – dialoogsüsteemi funktsionaalsust.

Lihtsuse huvides püüame hoida nõudmised madalal ning lepime kokku, et dialoogsüsteem võiks osata vastata küsimustele, mis käsitlevad vaid etenduse pealkirja, toimumisaega ja toimumiskohta teatris või linnas. Näiteks:

```
Mida mängitakse Linnateatris 22. mail?
Millal mängitakse etendust "Isad ja pojad"?
Tahaks kuulda Draamateatri mängukava.
Mida mängitakse Tallinnas 15. mail?
```

Teatud kahetsusega jätame loodava masina esialgu mõnest oskusest ilma. Masin ei anna vastuseid näitlejate kohta, ei tea etenduse sisukirjeldust, lavastajat, piletihindu, vabade kohtade olemasolu ega oska teha ka broneeringuid. Kui nimetatud info on kättesaadav, siis võime ka selle dialoogsüsteemile hiljem lisada. Esialgu aga piirdume vaid vähesega.

Lisaks põhioskusele – info jagamine – peaks süsteem olema lihtsasti hallatav ja täiustatav. Dialoogsüsteem võiks pidada vestluste

logi, et juba toimunud vestluste põhjal süsteemi edaspidi testida ja täiustada. Dialoogsüsteemi kasutuses olev teatrikavade andmebaas võiks olla lihtsasti uuendatav, parimal juhul automaatselt uuenev. Pidevalt muutuva info haldamine on lisatöö, mida tuleks võimaluse piirides alati minimiseerida.

Teadmised, mida süsteem omab lisaks teatrikavadele, võiksid paikneda süsteemivälistes andmestruktuurides. Siinjuures on silmas peetud näiteks käibefraaside (tervitused, hüvastijätud jms) kogumit. Süsteemivälised andmestruktuurid pakuvad ka programmeerimisoskuseta inimesele võimaluse süsteemi seadistamiseks. Ka lausemallid, mida vaatleme pisut allpool, peaksid asuma süsteemist väljaspool.

Lähtuvalt püstitatud nõuetest saame määratleda loodava süsteemi andmevajaduse. Antud ülesande lahendamiseks piisab teatrikavadest, kus sisaldub teatri nimi, teatri asukohalinn, etenduse pealkiri, toimumise kuupäev ja algusaeg.

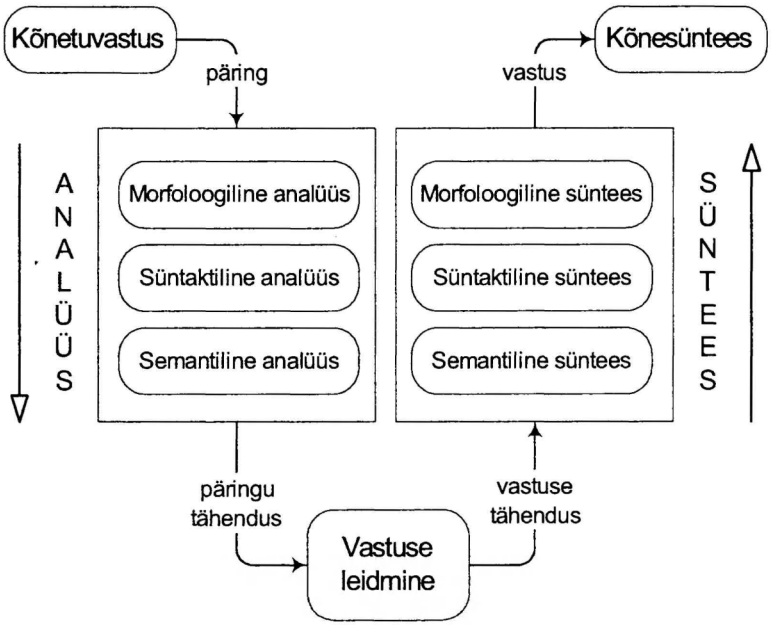
2.2. Realisatsioon

Dialoogsüsteem koosneb mitmest allsüsteemist, mis paiknevad erinevatel tasanditel (vt joonis 1). Kõnetuvastus ja kõnesüntees on kujutatud eraldiseisva tasandina, sest praktikas on need realiseeritud eraldiseisvate komponentidena. Siiski lõimuvad ka sellel tasandil morfoloogiline ja fonoloogiline teadmus (Karlsson 2002). Fonoloogiline tasand on näidatud eraldiseisvana ka seetõttu, et dialoogsüsteem saab funktsioneerida ka ilma selle tasandita. Kõnetuvastuse puudumisel tuleb sisend klaviatuurilt ja kõnesünteesi puudumisel väljastatakse vastus ekraanile.

Morfoloogiline, süntaktiline ja semantiline teadmus on koondatud omaette tasandile, sest praktikas on need realiseeritud fonoloogilisest tasandist eraldiseisvate komponentidena, mis küll on omavahel tihedalt seotud.

Vastuse leidmise allsüsteem ehk dialoogi juhtimiskeskus on tasandiks, kus funktsionaalsus ei ole keelespetsiifiline.

2.2.1. Vastuse leidmine. Kasutaja küsimusele vastuse leidmiseks tuleb küsimusest aru saada. Masinad aga ei saa inimesest aru. Või saavad? Kui peame arusaamiseks sõnale tähenduse leidmist (nt sõnastikust), siis võime tinglikult öelda, et masinad siiski saavad keelest aru.



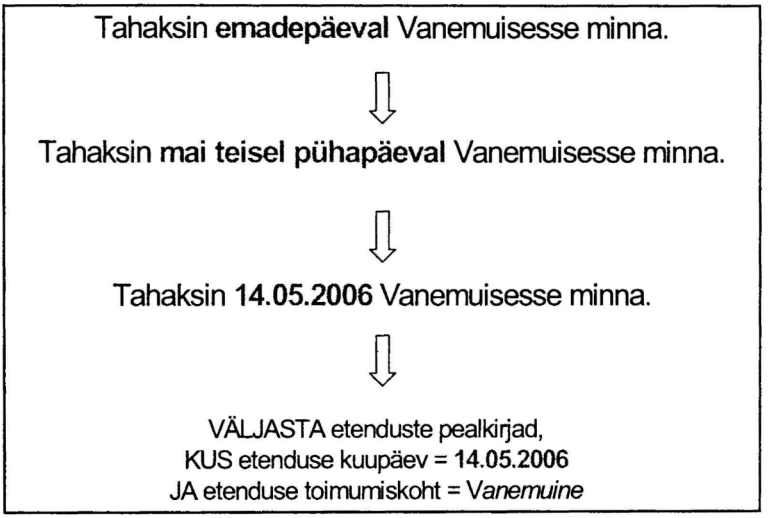
Joonis 1. Dialoogsüsteemi allsüsteemid

Võib aimata, et küsimuse tükeldamine sõnadeks ning nendest sõnadest morfoloogilise analüsaatori abil algvormide genereerimine pakub võimaluse sõnastikust tähenduste otsimiseks. Lisaks võib saadud sõnadest ja tähendustest moodustada päringuid ning esitada need teatriinfo andmebaasile arusaadavas (formaalses) päringukeeles.

Omades sõnastikku ja oskust moodustada andmebaasile arusaadavaid päringuid, suudame konstrueerida kasutaja küsimusest terve hulga päringuid ning saata need andmebaasile. Võib juhtuda, et mõni nendest päringutest saab andmebaasist vastuse. Veel enam, vastus võib osutuda kasutajale sobivaks.

2.2.2. Kuupäevade tuvastamine. Kui räägime graafikutest või ajatabelitest, siis on üks olulisemaid tunnuseid sündmuse toimumise kuupäev. Tihtipeale on kuupäeva järgi vaja otsida graafikust sobiv vastus. Näiteks tuleb otsida reedeseid etendusi.

Siiani ei ole eestikeelsetes korpustes eraldi märgendatud kuupäevi ja/või kellaaegasid. Näiteks lauses "Tahaksin kaks päeva enne



Joonis 2. Kuupäeva tuvastamine

tõnisepäeva koju minna." võiks olla märgendatud "kaks päeva enne tõnisepäeva" => [KUUPÄEV: 15. jaanuar]. See aga on semantilise taseme märgendus, mida ei ole siiani eesti keele korpustes tehtud. Süntaktilisel tasandil märgendatakse mitmesõnalised ajaväljendid määruseks (@ADVL), ei eristata aja-, koha- ja viisimäärusi. Kuupäevade tuvastamist hõlbustaks vastavalt märgendatud korpuste olemasolu. Suurimat kasu võiks saada Eesti dialoogikorpuses ajaliste väljendite märgendusest, sest infodialoogides kasutatavad ajamääratlused sarnanevad dialoogsüsteemides ettetulevatega kõige enam. Ajaliste väljendite soovitatavad märgendusreeglid saab leida internetist ning nende väljatöötamisega eesti keele jaoks eraldi vaeva ei oleks vaja näha (Gerber 2002).

Uurides kuupäevade erinevaid kirjutusvõimalusi eesti kirjakeele korpusest selgub, et kuupäevade sõnastusvõimalusi on suhteliselt palju ning seetõttu on tekstis kuupäeva tuvastamine päris raske ülesanne (vrd Hobbs 2003; Busemann 1997).

Lihtsamal juhul on kuupäev kirjeldatud päeva ja kuuga (31. detsembril), kuid keerulisemad juhud on näiteks kuupäevade vahemikud (jaanuari algul), rahvakalendri tüüpi kuupäevakirjeldused (uusaastaööl), riiklikud pühad ja tähtpäevad (vt joonis 2).

Kuupäevade tuvastajat on võimalik käsitleda kui eraldiseisvat, dialoogi valdkonnast sõltumatut komponenti. Kuupäevade tuvastamine on otsekui omaette keeletehnoloogiline vahend, mida saab integreerida suvalisse sobivasse rakendusse.

Kuupäevade tuvastamist saab kasutada näiteks ka pihuarvutis, kus toimub käekirja tuvastamine. Sellisel juhul saab lisaks käekirjale tuvastada jooksvalt ka kuupäeva ja näiteks lisada selle põhjal kalendrisse sündmuse või märkuse. Kuupäevade tuvastamist saab kasutada ka info visualiseerimisel, leides näiteks tekstist kuupäevad ning järjestades selle põhjal tekstis esitatud info kronoloogiliselt.

Kuupäevade tuvastamisega sarnaseid lahendamist ootavaid alamülesandeid on veelgi. Näiteks pärisnimede tuvastamine ja geograafiliste asukohtade tuvastamine (Volk 2001). Nimede tuvastamised on peamiselt realiseeritud leksikonipõhiselt. Kuupäevade tuvastamine nõuab aga dünaamilisemat lähenemist, sest ei ole mõistlik kirjeldada leksikonis ühe kuupäeva kõikvõimalikke sõnastusvõimalusi. Sestap tuleks kuupäevade tuvastamisel moodustada produktsioonid, mis regulaaravaldisele vastava sõnastusvormi teisendavad kuupäevaks kujul pp.kk.aaaa (päev, kuu, aasta):

```
([0123456789]{1,2})\.\sjaanuar ==> $1.1.%A
([0123456789]{1,2})\.\sveebruar ==> $1.2.%A
```

Taoliselt kirjapandud produktsioonide baas võiks olla süsteemiväline ning lihtsasti laiendatav.

2.2.3. Päringute moodustamine. Andmebaasile saame esitada päringuid. Omaette ülesanne on viia kasutajalt saadud küsimus andmebaasi jaoks arusaadavale kujule (vt joonis 3).

Teisisõnu, kasutajalt saadud küsimus tuleb andmebaasi jaoks ümber sõnastada. Näiteks lause

> "eee tahaks pühapäeval vanemuisesse minna on seal midagi va"

saab ümber sõnastada järgmisel viisil:

```
VÄLJASTA etenduste pealkirjad,
KUS etenduse kuupäev = tuvasta_kuupäev(tagasta_algvorm(pühapäeval))
JA etenduse toimumiskoht = tagasta_algvorm(vanemuisesse).
```

Antud näites on päringu viimases osas rakendatud morfoloogilist analüüsi ning kuupäeva tuvastamise funktsiooni, mille abil saame sõnast "pühapäeval" moodustada vajaliku kuupäeva.

Tekst kasutajalt

eee tahaks pühapäeval vanemuisesse minna
on seal midagi va

Päring andmebaasile

```
SELECT DISTINCT ete.pealkiri
      FROM etendused ete, teatrid tea
    WHERE tea.id = ete.tea_id
      AND ete.toim_kuupaev = str_to_date ('18.09.2005', %d. %m.Y')
      AND tea.nimetus = 'VANEMUINE'
  Order BY ete.toim_kuupaev
```

Joonis 3. Kasutajalt saadud teksti teisendus päringukeelde

Andmebaasi poole on võimalik pöörduda teoreetiliselt lõpmatu hulga päringutega. Saame küsida etendusi teatri nime järgi, etenduse nime järgi, kombineeritult teatri ja etenduse nime järgi jne.

Iga võimalik päring aga ei pruugi olla mõistlik ega vajalik. Näiteks "Palun tagasta etendused, mille pealkiri oleks vähemalt 8 tähe pikkune" on võimalik päring, kuid ei ole mõistlik. Seega näeme, et tuleb leida "mõistlik" hulk päringuid, mida võidakse kasutada andmebaasi poole pöördumisel. Järgnevalt on toodud mõned näited päringumallidest, mis tuleks dialoogsüsteemis realiseerida.

Palun etenduste nimekiri.
Palun etenduste nimekiri kuupäeval K (või perioodil A kuni B).
Palun teatrite nimekiri, kus toimub etendusi.
Palun linnade nimekiri, kus toimub etendusi.
Palun etenduse E toimumisaeg, algusaeg ja koht.
Palun kuupäeval K (või perioodil A kuni B) toimuva etenduse E algusaeg ja koht.
Palun teatris T mängitavad etendused.
Palun teatris T kuupäeval K (või perioodil A kuni B) mängitavad etendused.
Palun linnas L mängitavad etendused.
Palun linnas L kuupäeval K (või perioodil A kuni B) mängitavad etendused.
Palun teatrite nimekiri linnas L.

2.2.4. Vastuse moodustamine. Kasutajale vastamisel tuleb andmebaasist saadud infot kasutades moodustada lause, st lahendada keele genereerimise probleem. Lause tuleb moodustada sõnadest, mis peavad olema korrektses vormis. Lause konstrueerimisel saame kasutada eesti keele morfoloogilise süntesaatori abi, mis suudab genereeri-

da algvormina antud sisendsõnast ja vormitunnustest meie lausesse sobiva sõnavormi.

Analoogiliselt päringu moodustamisel kasutatavatele päringumallidele tuleks vastuse moodustamise juures mõelda läbi võimalikud päringutulemused. Edasi saame iga tulemuse kohta panna kirja vastuses kasutatava lause üldkuju nn lausemalli ehk mustrina.

Näiteks oletame, et kasutaja esitas küsimuse teatud teatri kohta teatud kuupäeval toimuvate etenduste teadasaamiseks. Üheks tulemuseks on, et sel kuupäeval nimetatud teatris etendused puuduvad, ning sellisel juhul võime kasutada järgmisi lausemalle:

> "<kuupäev; ainsuse alalütlev> ei toimu teatris <teater; ainsuse nimetav> ühtegi etendust."
> "Kahjuks ei toimu <kuupäev; ainsuse alalütlev> teatris <teater; ainsuse nimetav> ühtegi etendust."

Antud lausemallides on lüngad, mille täitmiseks tuleb kasutada päringu parameetreid ja teatrikavast saadud infot. Esimesele lünga asemele tuleb kirjutada kuupäev, mida kasutaja oma päringus mainis, ning teise asemele teater, mille kohta kasutaja päringu esitas. Kuupäev võib olla näiteks "1. jaanuar" ja teater võib olla "Vanemuine". Vastuse moodustamisel tuleb kuupäev panna ainsuse alalütlevasse käändesse ehk 'jaanuar' asemel tuleb öelda 'jaanuaril'. Teatrinime saab jätta ainsuse nimetavasse, sest antud lausemall võimaldab seda. Lõplikuks vastuseks saame siis:

> "1. jaanuaril ei toimu teatris Vanemuine ühtegi etendust."

Sedalaadi valdkonnapõhiseid lausemalle tasuks hoida süsteemivälistes andmestruktuurides, mis pakub lihtsa võimaluse vastusemallis muudatuste tegemiseks.

Kokkuvõte

Dialoogsüsteemide loomine sisaldab palju inseneritööd, aga üsna suurel määral tuleb kasutada ka keelelisi teadmisi. Nüüd, kus eesti keele morfoloogiline analüüs ja süntees on täielikult automatiseeritud, saab juba ainuüksi nende komponentide abil lahendada suure hulga keelelisi probleeme.

Käesolevaks ajaks on loodud kaks dialoogsüsteemi, mis on kasutatavad http://www.dialoogid.ee/ kaudu ning mõeldud vastavalt lennuinfo ja teatriinfo andmebaaside päringuliidestena.

Edasises plaanis on luua tarkvaraline keskkond dialoogsüsteemide loomise hõlbustamiseks (võttes eeskujuks nt http://cslu.cse.ogi.edu/toolkit/).

Järgmine samm, kus keeleteadlaste abi tarkvarainsenerile on teretulnud, on dialoogiaktide (nt rituaalid, küsimused, direktiivid) automaatne tuvastamine (vt Fišeli ja Kikase artiklit käesolevas kogumikus). Oodatud on analüüsid, kus püütakse määratleda dialoogiakte reeglipõhiselt. See pakub täiendava võimaluse parandada statistilist tuvastust (Fishel 2005). Lisaks dialoogiaktide märgendamisele korpuses oleks vajalik ka teiste semantiliste üksuste (sh kuupäevade) märgendamine.

Kirjandus

Alumäe, T., Võhandu, L. 2003. Piiratud ulatusega eestikeelne kõnetuvastus. – Toimiv keel. I, Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Tallinn: Eesti Keele Sihtasutus, 50–52.

Busemann, S., Decleck, T., Diagne, A. K., Dini, L., Klein, J., Schmeier, S 1997. Natural Language Dialogue Service for Appointment Scheduling Agents. – Proceedings of the Fifth Conference on Applied Natural Language Processing, 25–32.

Fishel, Mark 2005. Dialogue Act Recognition in Estonian Dialogues using Artificial Neural Networks. – Proceedings of the International Conference The Second Baltic Conference on Human Language Technologies, 231–235.

Gerber, L., Ferro, L, Mani, I, Sundheim, B, Wilson, G., Kozierok, R 2002. Annotating Temporal Information: From Theory to Practice. – Proceedings of the 2002 Conference on Human Language Technology. San Diego, CA, 226–230.

Hobbs, J. R., Pustejovsky, J. 2003. Annotating and Reasoning about Time and Events. – Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning. Stanford, California.

Karlsson, Fred 2002. Üldkeeleteadus. Tõlkinud ja kohandanud R. Pajusalu, J. Valge, I. Tragel. Tallinn: Eesti Keele Sihtasutus.

Manaris, B 1998. Natural Language Processing: A Human-Computer Interaction Perspective. – Advances in Computers 47. Ed M. V. Zelkowitz. New York: Academic Press, 1–66.

Nass, Clifford 2005. Wired for speech: How voice activates and enhances the human-computer relationship. Cambridge, MA: MIT Press.

Volk, Martin, Clematide, Simon 2001. Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition. – Proceedings of 6th International Workshop on Applications of Natural Language for Information Systems. Madrid.

Kuidas võiks masin teha vahet otsesel ja kaudsel *kas*-küsimusel?[1]

Tarmo Truu
Tartu Ülikool

Kaudne kõneakt on keeruline, aga huvitav kõneakt, mida kasutatakse üsna palju. Ta väljendab ebamäärasust ja ebakindlust. Ta jätab vastajale suurema vabaduse vastusevalikul. Kaudne võib olla ükskõik milline kõneakt, nt küsimus, soov, käsk, lubadus jt. Kõigil neil kaasnevad aga eri omadused, millega kaudset tähendust lisatakse. Selleks võib olla puhas keeleline konstruktsioon, aga võib olla ka intonatsioon, miimika või muu väga spetsiifiline meetod.

1. Sissejuhatus

Kõneaktiteooriasse, mille rajaja oli 1960. aastatel inglise filosoof John Austin, tõusis kaudse kõneakti teema teise mõjuka filosoofi, ameeriklase John Searle'i mõtete kaudu. Lisaks filosoofidele on kõneakte uurinud matemaatikud, psühholoogid ja loomulikult ka keeleteadlased. Igaüks oma vaatenurgast. Võib-olla sellepärast pole tekkinud ühtset teooriat.

Kõneakte on läbi aegade üritatud liigitada mitut moodi. Ühtset kõigi poolt aktsepteeritud liigitust ei ole tänaseni koostatud. Eirates spetsiifilisemaid jaotusi, võib esile tuua kolm klassikalisemat lähenemist või liigitust. Austin väitis, et kõneleja, moodustades lausungit, teeb kindlaid tegevusi, mis sisaldavad kolme tüüpi lingvistilisi akte: lokutiivseid (keegi ütleb midagi, mis ei kanna endas kommunikatiivset sisu), illokutiivseid (keegi kihutab kedagi midagi tegema, nt käsib, väidab, palub jms) ja perlokutiivseid (kõneleja saavutab eesmärgi, mille ta illokutiivse aktiga püstitas) (Marmaridou 2000: 173, 174). Seega võivad kõik need aspektid sisalduda ühes kõneaktis, sest Austini järgi tähendab lokutiivse akti esitus ka koheselt illokutiivse akti esitust (Marmaridou 2000:174). Searle jagas kõneaktid otsesteks ja kaudseteks (Searle 1975). Otsesed kõneaktid on aktid,

[1] Tööd on toetanud Eesti Teadusfond (grant nr 5685) ning HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

millega kõneleja mõtleb sõna-sõnalt seda, mida ütleb, ja nii peab ka kuulaja aru saama. Kaudsed kõneaktid on seevastu aktid, millega kõneleja kasutab ühte liiki kõneakti, kuid teist liiki kõneakti vormiga. Itaalia uurimisgrupp on välja pakkunud kõneaktide jaotuseks lihtsad ja keerulised kõneaktid. Lihtsad kõneaktid on aktid, mis on kergesti mõistetavad. Siia kuuluvad otsesed ja konventsionaalsed kõneaktid. Keerulised kõneaktid on aktid, mida on raskem mõista, ja siia kuuluvad kaudsed ja mittekonventsionaalsed kõneaktid. See liigitus on parem kui otseste ja kaudsete liigitus, kuna võimaldab sisse tuua ebastandardseid kõneakte, nagu iroonia ja valelikkus. Tulevikus soovitakse liikuda selle eristusega nii kaugele, et saaks lisada ka dialoogis tegelikult väga tähtsat rolli mängivad miimika, kehakeele jms (Bosco jt 2003).

Austini lähenemine ja kaks nimetatud liigitust seostuvad otseselt kolme erineva kõneaktiteooriaga, mida ei tohi omavahel segamini ajada, sest neil on erinev metoodika ja erinev eesmärk.

Kõneaktiteooria uurib, mismoodi muudavad kõneaktid maailma, kuidas mõjutavad inimesi, kuidas väljendavad emotsioone jne. Kõneaktiteooriasse kuuluvad lokutiivsus, illokutiivsus ja perlokutiivsus.

Kõneaktisemantika uurib, mil määral on lausungitesse (lausung on suulise kõne liigendaja, võrdväärne kirjakeelse lausega) paigutatud lingvistilist koodi, mis aitaksid tuvastada vastavat kõneakti. Üks selline vahend on performatiivverb.

(1)
Ma hoiatan sind, et seal väljakul on üks kuri pull. (Siin ei saa tekkida kahtepidi mõtlemist, performatiivne tegusõna on 'hoiatama'.)

Kõneaktipragmaatika uurib, kuidas kuulaja otsustab, mis kõneakti kõneleja tarvitas. Siia kuuluvad otsesed ja kaudsed kõneaktid.

(2)
Pull on väljakul. (See võib olla nii otsene hoiatus kui kaudne soovitus, et pole mõtet sinna väljakule oma nina pista.)

2. Kaudne kõneakt

Kaudse kõneakti kõige otsesem definitsioon kõlab: öeldakse üht, mõeldakse teist. Näiteks esitan palve, mida tuleb mõista hoopis käsuna, või küsimuse, mida tuleb mõista soovina jts (Searle 1975). Kaudsed kõneaktid tekitavad suhtluses lisatasandi, kus kõneleja suhtleb kuulajaga rohkem kui vaid selle sõnasõnalise infoga. Seoses

sellega tõusevad oluliseks teguriks ühised teadmised (ehk teadmuse ühisosa). Ühisosas mängivad rolli väga paljud asjaolud: kes on suhtlejad, mis on nende sotsiaalne taust, kas nad on tuttavad, kas neil on ühiseid mälestusi, kus see vestlus toimub, kas vestluses osaleb rohkem inimesi, kas teised ei tohi öeldut mõista, mis stiilis suheldakse, mis on nende eesmärgid, kas keel on mõlemale emakeel, mis rahvusest nad on, kus elavad, mis on haridus, mis ametit peavad jne (Clark 1992: 262).

Otsese kõneakti näide:

(3)
H: ee 'kas teil Re'noo juppe ka 'sees on. | KYE: SULETUD KAS | (Küsija soovib konkreetselt teada, kas poes on vajalikke osi või mitte.)

Kaudse kõneakti näide:

(4)
H: e:ga te ei oska öelda: Tartu 'raudteejaama 'numbrit. | KYE: JUTUSTAV KAS | (Küsija võtab viisaka ja tagasihoidliku tooni.)

Samas peab vastaja mõistma, et küsija soovib infot, mitte seda, kas ta oskab või ei oska vastata.[2]

Et mõista kaudset kõneakti, peab kuulaja tõusma mõistmise kõrgemale tasandile, ammutama sealt vajalikku infot. Mõistmise kõrgem tasand tähendabki vestlejate teadmuse ühisosa (ingl *common ground*) (Stalnaker 2002). Kui üks esitab küsimuse kõneakti *Kas sa oskad öelda, kui palju kell on?*, siis peavad mõlemad tõusma hetkeks kõrgemale tasandile. Küsija tahab tegelikult teada, mis kell on, mitte seda, kas vastaja oskab öelda. Vastaja peab mõistma sama. Seega peavad kõneakti õnnestumiseks mõlema eesmärgid ühtima. Kui ei ühti, siis muutub see kaudne kõneakt otseseks. Nurjumisel võivad olla erinevad põhjused, nt võis vastaja nalja heita. Kui minna veel sügavamale, siis võib keegi küsida *Kas sul kella on?*. See küsimus on veelgi kaudsem. Õigem on tegelikult väita, et eelmist näidet on võimalik vormiliselt tuvastada. Teine näide on palju keerulisem. Kui keegi tänaval sellise küsimuse esitab, siis oleneb kõik olukorrast. Kui seda öösel kuskil kahtlasel kõrvaltänaval küsima tullakse, siis võib see küsimus olla otsene ja eesmärgiga see kell lihtsalt ära va-

[2] Dialoogiaktide märgendid vt Lisa 2. Akt | KYE: SULETUD KAS | ootab jah/ei vastust, | KYE: JUTUSTAV KAS | nõuab vastuseks pikemat infohulka (Hennoste, Rääbis 2004: 99).

rastada. Kui aga keskpäeval kesklinnas, siis ilmselt soovib küsija teada, kas vastajal on kell, millelt vaadata kellaaega. Küsimuse vormile otsa vaadates on tegu pigem otsese küsimusega. Kui selline küsimus masinale esitada, siis kindlasti vastab ta automaatselt jaatavalt, sest pole ette anda ühtki märksõna, mis ütleks talle, et tegu on kaudse kõneaktiga. Antud näite põhjal saab kaudsed kõneaktid jagada mingil määral kaheks: küsimusest ja vastusest tuvastatavad. Küsimusest tuvastatav on selline, millel on selge vorm ja kindlad märksõnad. Vastusest tuvastatav on see, mille vorm on pigem otsene kui kaudne ja mille mõte on peidetud ridade vahele.

3. Kaudse kõneakti tuvastamine

Kaudse kõneakti tuvastamiseks on aegade jooksul pakutud välja erinevaid teooriaid. Anas Yasin (2002) pakkus teooria, mis sisaldab nelja tingimust, kuidas olla kindel, et tegu on kaudse kõneaktiga.

1. Lausungid ei saa olla kaudsed kõneaktid, kui nad sisaldavad performatiivseid verbe.

2. Kui ei leidu ühtegi performatiivset verbi, siis tuleb vaadata lausevormi, kas see on kirjeldav, küsiv, käskiv või muu. Muidugi võib ka otsene kõneakt neid vorme võtta, sellepärast peab arvestama kaht järgmist tingimust.

3. Laused, mille sõnalised tähendused rikuvad õnnestumistingimusi, on kaudsed.

(5)
Kas sa sulgeksid ukse? (See rikub õnnestumistingimusi, sest on küsimuse vormis eeldus, vt 2. punkti.)

(6)
Palun sulge uks. (See eeldab 2. punkti.)

4. Kujutage ette konteksti, kus lausung esitatakse, ja püüdke silme ette manada kuulaja reaktsioon. Kui ta vastab nii, nagu peaks vastama, siis võib üsna kindel olla, et see lausung on kaudne kõneakt. Kui kuulajad lihtsalt kinnitavad mingit lausungit, siis oli tegu väitega; kui nad vastavad, siis on tegu ilmselt küsimusega jne.

See pole küll kõneaktiteooriat muutnud lähenemine, kuid on huvitav ja üsna praktiline. Yasin on üles kirjutanud olulised reeglid, mida on võimalik praktikas rakendada.

Hoopis teistmoodi lähenevad tuvastamisele Jill Nickerson ja Jennifer Chu-Carroll, kes võrdlevad otseste ja kaudsete kõneaktide

akustilisi ja prosoodilisi erinevusi. Nende analüüs näitas, et kaudsete kõneaktide puhul hääletoon madaldub oluliselt rohkem kui otseste puhul. Lisaks mängib rolli lauserõhk, nt ebaviisakates lausungites tõuseb rõhk lausungi lõpus (Nickerson, Chu-Carroll 1999).

Mainitud kaks täiesti erinevat teooriat annavad mõista, et kaudsete kõneaktide tuvastamiseks tuleb rakendada erinevaid meetodeid, et saavutada võimalikult häid tulemusi. Kas või näiteks neid kaht teooriat koos. Nagu eelpool mainitud, saab kaudseid kõneakte tuvastada küsimusest ja vastusest lähtuvalt. Selge on see, et kõneakte ei saa täielikult mõista, kui ei arvestata kuulaja ja kõneleja voore koos (Clark 1991). Pidades silmas masinat, mis võtab vastu infotelefonikõnesid, siis vastusest lähtuv lähenemine on kasulik ainult kõneaktide uurijatele. Masina õpetamiseks muutub vastusest lähtuv lähenemine vajalikuks, kui suudetakse masin pärast talle esitatud küsimust "mõtlema" panna. Ta võiks enne vastamist läbida mitu protsessi, mille tulemusena oleks tal valmis mitu vastusevarianti. Käesoleva artikli praktilise analüüsi osas kasutatakse küsimusest lähtuvat lähenemist ja püütakse leida esialgsed tulemused kaudse *kas*-küsimuse tuvastamiseks.

4. Otsene ja kaudne *kas*-küsimus

Tiit Hennoste juhtimisel loodud dialoogiaktide tüpoloogia sisaldab üle 100 akti. Rohkus on põhjendatud eesmärgiga uurida dialoogis esinevaid nähtusi. Dialoogsüsteemile pole niipalju spetsiifilisi akte küll vaja, aga pikemas perspektiivis võib süvauuringust palju abi olla. Selle kohta on tuua praktiline näide. Tüpoloogias märgib *kas*-küsimus kaht akti: suletud ja jutustav ehk otsene ja kaudne. Masinale oleks palju lihtsam, kui nendel vahet ei tehtaks. See võimaldaks masinal alati juhuslikult valida, kas vastata küsimusele otseselt või kaudselt. Kui kaht vormi uurida, siis erinevus seisneb selles, mis vastust ootab küsija. Suletud ehk otsese puhul tahab küsija saada kinnitust, reeglina lühikest, esitatud küsimusele. Jutustav ootab pikemat informatiivset vastust (Hennoste, Rääbis, 2004). Selline eristamine kergendab kaudsete kõneaktide uurimist. Mitteametlikult on nende kahe akti ingliskeelsed vasted *open yes/no question* ja *closed yes/no question* (Hennoste, Rääbis 2004). Võib-olla peaks nende aktide nimed tulevikus selguse huvides nimetama nt otseseks ja kaudseks *kas*-küsimuseks.

5. Analüüsi materjal

Oletame, et kõik *kas*-küsimused sisaldavad küsisõna *kas*, mille järgi tuvastab masin *kas*-küsimuse. Mis võiksid olla *kas*-küsimuses sisalduvad märksõnad, mille abil suudaks ta teha vahet otsese või kaudse *kas*-küsimuse vahel? Teooria kinnitamiseks või ümberlükkamiseks võtsin ette 312 dialoogi (edaspidi minikorpus). Praktilistel kaalutlustel valisin välja institutsionaalsed telefonikõned, mis kuuluvad dialoogikorpusesse ning on märgendatud ja ühtlustatud 2004. aastal. Spetsiifilisemalt jagunevad dialoogid infotelefoni- (228), registratuuri- (14) ja muu infopäringuga dialoogitüüpideks (70). Kõigis nendes dialoogides on helistaja eesmärk koguda informatsiooni. Dialoogides märgib H helistajat ja V vastajat. Seda on oluline meeles pidada, sest praegu uuritakse dialooge, kus mõlemad osapooled on inimesed, aga tulevikus peaks V rolli täitma juba masin.

Analüüsimiseks ja statistika koostamiseks kasutasin Margus Treumuthi programmeeritud Dialoogikorpuse tööpinki (Treumuth 2005). Lisaks sellele kasutasin ka operatsioonisüsteemi Unix, sest väga konkreetsetele küsimustele ei saa tööpingi abiga vastuseid, näiteks kui on vaja leida, kui palju esineb *kas*-küsimustes küsisõna *kas*. Tööpink küll leiab need üles, aga sinna hulka arvab ta ka need lausungid, mille ühe sõna sees on tähtede järjend *-kas-*.

Helistaja H voore, mille aktiks jutustav *kas*-küsimus (edaspidi JKK), oli 163, vastaja V omi aga 19. Suletud *kas*-küsimust (edaspidi SKK) oli H esitanud 76 korda ja V 67 korda. Märkimisväärse erinevuse põhjustas see, et kuna helistaja eesmärk on infot koguda, siis loomulikult loodab ta oma *kas*-küsimusele saada pikema sisuga infot ja on selle *kas*-küsimuse moodustanud JKK kujul. SKK puhul on H-i ja V esinemiskordade võrdsus tingitud eelkõige sellest, et vastaja esitab tihti lühikesi järelküsimusi, mis aitavad vastajal täpsemat infot kokku panna ja väljastada. Kui kaudne küsimus koosneb kahest osast, siis otsene küsimus on üheosaline. Kaudses küsimuses on peidus nii otsene kui ka kaudne tähendus (Clark 1991). Antud juhul meenutab SKK otsest küsimust, mille taga ei peitu kaudset soovi.

Olgu lisatud, et edaspidi uuritakse ainult 163 JKK ja 76 SKK vooru, mille autoriks on H. See on põhjendatud eesmärgiga välja selgitada, kuidas masin M suudaks tuvastada kaudset kõneakti.

6. Põhilisemad märksõnad

Analüüs on jagatud kahte ossa: esimeses on kommenteeritud ja näidetega varustatud olulisemaid märksõnu, teises koos küsisõnaga *kas*. Märksõnu nimetatakse sõnedeks. Peatüki lõpus on toodud võrdlev tabel.

6.1 Esimene grupp

6.1.1. Kas. 163 JKK koosseisu kuulus küsisõna *kas* või selle variant *ks* 92 korda ehk 56%, 76 SKK sisaldas küsisõna *kas* 53 korda ehk 70%. Minikorpuses leidus lisaks 75 muud sõnet *kas* sisaldavat akti. Seega ei saa kinnitada, et *kas* esineb ainult *kas*-küsimuses. Samas jääb masin, kui ainult küsisõna *kas* uurida, kaudse ja otsese kas-küsimuse tuvastamisega hätta.

(7)
H: kas teil `üliõpilastele mingit `hinnasoodustust ka on. | KYE: JUTUSTAV KAS |

(8)
H: ma ei tea kas `kaart jääb siia kaardi järele ei pea tulema | KYE: SULETUD KAS |

(9)
H: mm ma=ei kujuta ette et kump oleks `odavam kas `bussiga või > `lennukiga=lennukiga oleks muidugi `see=et < saaks `kiiremini `kohale. | KYJ: INFO ANDMINE |

6.1.2. Mingi. Mingi (11), mingid (10), mingit (9), mingisugune (3), mingisuguseid (2), mingisugused (2), mingil (1); mingisugust (1).

Sõne *mingi* ja tema erinevad variandid eristavad selgelt JKK ja SKK. JKK-s leidus selliseid sõnesid 38, SKK-s aga ainult 1 (*mingisugust*). Seega saab nende andmete põhjal öelda, et kui tegu on *kas*-küsimusega, siis võib üsna kindla veendumusega märgendada selle JKK-ks. Sõne *mingi* on justnagu lünk, mille peab infoga täitma vastaja. *Mingi* eri vormid esinesid kokku 34 erinevas JKK-s (mõnes oli kaks korda) ja kuna minikorpuses esines *mingi* veel 52 korral erinevates kõneaktides, siis õige automaatse märgendamise tõenäosus on 41%, mis on madal, aga vähemalt ei saa eksida *kas*-küsimuse liigi määramisega.

(10)
H: e kas mul on veel üks küsimus kas neil on mingid erinevad modifikatsioonid 'ka selles=suhtes=et 'pikkuste erinevused või [või 'käigukasti] | KYE: JUTUSTAV KAS |

(11)
H: ee 'sellele te mingisugust=ee | KYE: SULETUD KAS |

6.1.3. Mõni. Sõne *mõni* funktsioon sarnaneb *mingi* omaga. Sõnesid *mõni* leidus JKK-des kokku 5 ja alati koos küsisõnaga *kas*. SKK-des polnud ühtegi.

(12)
H: .hh kas: Võrust 'Käärikule ka mõni 'buss läheb. | KYE: JUTUSTAV KAS |

6.1.4. Midagi. Ka selle sõne põhimõte sarnaneb eelnimetatutega. Märkimisväärne oli *midagi* esinemissagedus. Seda leidus JKK-s 21 korral, aga mitte kordagi SKK-s. Lisaks 21 lausungile sisaldas minikorpus veel 24 lausungit, mille üks sõne oli *midagi*, aga kõneakt muu. Seega oleks avastamise tõenäosus 46%.

(13)
H: kas midagi ette vaja veel 'seitset või .hh või või mida | KYE: JUTUSTAV KAS |

6.1.5. Saama. Sõne *saama* andis üllatava tulemuse. Minikorpuses esines see 210 vormis, millest 31 JKK-s ja 10 SKK-s. Üllatus selgub, kui uurida tabelist, mis on tuvastusprotsent, kui ühes kõneaktis on koos küsisõna *kas* ja *saama*.

(14)
H: .hh kas teie käest saaks informatsiooni kui palju võiks maksta sõit 'Inglismaale. | KYE: JUTUSTAV KAS |
H: 'kas te mulle 'saate firma 'nime vaadata 'numbri järgi,= | KYE: SULETUD KAS |

Ülejäänud märksõnad olid *vä, ütlema, võimalik, näiteks* ja *ega*. Nende andmed on nähtavad tabelist. Tärn tabelis tähistab seda, et märksõnadel on ühine tüvi ja nad on kokku võetud.

Tabel 1. Otseste ja kaudsete *kas*-küsimuste hulk ja võimalikud tuvastusprotsendid

Märksõna	JKK	SKK	Korpus	JKK %	SKK %	Korpuse %
Kas	92	53	75	41,8	24,1	34,1
*Mingi**	38	1	52	41,8	1,1	57,1
Vä	3	6	0	33,3	66,7	0,0
Mõni	5	0	2	71,4	0,0	28,6
Midagi	21	0	24	46,7	0,0	53,3
Ütlema	27	1	43	38,0	1,4	60,6
Võimalik	9	4	24	24,3	10,8	64,9

Märksõna	JKK	SKK	Korpus	JKK %	SKK %	Korpuse %
Näiteks	10	3	48	16,4	4,9	78,7
Saama	31	10	210	12,4	4,0	83,7
Ega	8	1	9	44,4	5,6	50,0
Kas + muu küsisõna	7	0		100,0	0,0	0,0
*Kas + mingi**	17	0	6	73,9	0,0	26,1
Kas + mõni	3	0	0	100,0	0,0	0,0
Kas + midagi	9	0	1	90,0	0,0	10,0
Kas + võimalik	5	2	0	71,4	28,6	0,0
Kas + ütlema	13	0	11	54,2	0,0	45,8
Kas + näiteks	4	3	1	50,0	37,5	12,5
Kas + saama	21	9	10	52,5	22,5	25,0
Tahtsin + küsida	6	0	1	85,7	0,0	14,3
Kokku	237	40	442	33,0	5,6	61,5

6.2. Teine grupp

Teise grupi moodustavad paarid, kus peamiseks on küsisõna *kas* ja kõrvaliseks üks märksõna. Nagu võis eeldada, kasvab märksõnapaaride korral tuvastusprotsent, aga see, et kasv on niivõrd suur, oli üllatus. Statistiliselt võivad tulemused kõikuda siia ja sinna, aga võib loota, et suhe jääb samaks. JKK puhul on kasv 30,1% kuni 65,9%. SKK puhul jääb praktiliselt samaks. Seega langeb minikorpuse muude kõneaktide osakaal JKK kasuks.

Kõige drastilisem näide on sõnega *saama*. Kui üksi oleks JKK tuvastus *saama* abil heal juhul küündinud kõigest 12% ja SKK tuvastus 4%, siis koos küsisõnaga *kas* kerkis JKK puhul tuvastus 52,5% ja SKK puhul 22,5%.

7. Kaudne küsimus ja otsene vastus ning otsene küsimus ja kaudne vastus – lahendamatu probleem?

Dialoogi märgendamisel tekib olukordi, kus ka inimene ei tee vahet, kas tegu on otsese või kaudse küsimusega. Suletud ja jutustavad kasküsimused võivad tihti minna sõlme neile järgnevate vastustega. Märgendusjuhend ütleb, et märgendamisel tuleb lähtuda küsija tõlgendusest. Kui konteksti ja küsimust vaadates selgub, et tegu on

kindlalt suletud *kas*-küsimusega, aga kuna vastaja seda ei mõistnud, vaid hakkas küsijale soovimatut infot jagama, siis antud olukorras tuleb see akt märgendada jutustavaks ehk kaudseks. Sama kehtib ka teistpidi.

Järgnevas näites on tegu kaudse küsimusega. Sellele viitab tema vorm (Kas te oskate öelda kellaaega?). Vastajal puudub aga info ja ta vastab eitavalt. Seega tegelikult peaks teooria järgi kaudne küsimus muutuma tagantjärgi hoopis otseseks, sest see ei õnnestunud. Vastaja edasised väited ei puutu enam asjasse, sest need põhjendavad ainult seda, miks ta ei tea ja ei saa vastata.

(15)
H:[ee] kas te ei oska öelda `tunnihinda ka neil. | KYE:JUTUSTAV KAS |
V: `ei, ei `tea, | KYJ: INFO PUUDUMINE |
V: seepärst `ma: olen nagu `Rahva=auto aga see on `Aspar, aga meil on paraleel`telefon, | IL: PÕHJENDAMINE |
V: (.) [nii=et=ma=i:] `nende `tööga ma=i=ole `ültse `kursis, ei=e ei [`tund] | KYJ: INFO PUUDUMINE |

Selles näites on olukord vastupidine, kus küsija esitab tegelikult kaudse küsimuse, aga on märgendatud otseseks, sest lähtutakse vastusest. Tegelikult annab infoandja vastuse küll kaudsele küsimusele, aga ainult selle esimese osa kohta. Kui korraks mõelda masinale, siis antud näite puhul polegi ju tegelikult oluline, mis aktiga on tegu. Esitatud küsimusele hakkab masin andmebaasist otsima, milline on mandlioperatsiooni järjekord septembris, ja kui ta leiab, et vaba aega pole, siis vastab eitavalt. Suhtlusprobleemi antud näite põhjal ei teki ka inimese ja masina vahel.

(16)
H: aga oskate öelda kuidas seal nüüd see mandli operatsiooni järjekord on, et kas: kas on võimalik septembris ära teha. | KYE: SULETUD KAS |
V: vot seda ma küll ei oska öelda, | DIJ: INFO PUUDUMINE |
V: see (.) see on jah kõrva kurguarsti (.) kõrva kurguarst oskab sellele vastata. | KYJ: MUU |

Kuidas sellised olukorrad lahendada, sõltub eelkõige sellest, mida arvestada – kas arvestada küsimuse vormi, mis oleks masinat silmas pidades kasulikum, või vaadelda küsimust ja vastust koos, mis annab uurimisel ja märgendamisel eelise. Selline keeruline olukord võibki jääda lahendamatuks, aga tihti pakub lahenduse naaberpaarile järgnev dialoogi areng. Kui küsija esitab jutustava *kas*-küsimuse, millele vastatakse jaatavalt, siis võib tekkida pikem paus, kus küsija loodab, et vastaja tõlgendab küsimust siiski õigesti ja annab sisukamat infot.

Kui pausist ei piisa, siis võib küsija sama küsimuse uuesti küsida, aga seda juba rõhutatumalt. Kui küsija esitab suletud *kas*-küsimuse, aga nõustumise asemel jagab vastaja infot, mida tegelikult ei küsitud, siis võib küsija infojagamisele vahele segada ja infotulva katkestada. Nii tekib küsimus, kas alati saab lähtuda küsija tõlgendusest – ka vastajal on üsna suur roll. Nagu eespool juba mainitud, ei saa kõneakte täielikult mõista, kui ei arvestata kuulaja ja kõneleja voore koos (Clark 1991). Kui masin eksib, vastates ilmselgele kaudsele küsimusele otse, siis päästab teda see, et olukorra lahendamiseks kasutab küsija oma vahendeid, et soovitud info teisiti kätte saada. Ta kas küsib uuesti või laseb pausil kesta, võimalusi on teisigi.

8. Kokkuvõte

Kaudsel ja otsesel *kas*-küsimusel on võimalik vahet teha. Seda kinnitab läbiviidud analüüs. Mõlemal vormil on küll sarnaseid külgi, kuid nagu selgub, leidub ka palju selliseid pisikesi detaile, mis neid kaht kindlalt eristavad. Masina jaoks on esmane ülesanne tuvastada *kas*-küsimus. Alles pärast seda on võimalik jätkata otsese või kaudse küsimuse väljaselgitamist. Tulemustest selgub, et kõige olulisem roll langeb küsisõna *kas* esinevusele. Kui see on olemas, siis kasvab tõenäosus akt tuvastada. Eriti kasvab kaudse *kas*-küsimuse tuvastusprotsent.

Kaudse kõneakti sees on vähemalt kaks tähendust, vastama peab aga kõige tähtsamale osale. Ka selles analüüsis selgus, et kasutatakse kõige tavalisemat kaudse küsimuse vormi *Kas sa tead, mis kell on?*. Siin esineb kaks nähtavat tähendust: kas sa tead ja kas sa ütleksid kellaaja. Lisaks on vähemalt üheks nähtamatuks tähenduseks soov, et vastaja ütleks kellaaja. Neid tähendusi võib veel mitmeid olla. See artikkel ei pööra tähelepanu sellele, mis põhjustel kaudseid kõneakte vormistatakse ja kuidas mõistetakse, vaid piirdub keeleliste vahendite uurimisega.

Oletame, et masin suudab tuvastada ca 60% kaudsetest kõneaktidest. Kas see on hea või halb? Mis probleeme võib see inimese ja masina suhtluses tekitada? Nendele küsimustele vastused veel puuduvad, aga lõppude lõpuks sünnib dialoog koostöö tulemusena. Ka inimeste vahel tekivad suhtlusprobleemid, ka inimesed ei mõista tihti, kas tegu on otsese või kaudse küsimusega. Miks peaks masin eksimatu olema?

Kirjandus

Austin, John 1962. How to do Things with Words. Clarendon Press.

Bosco, Francesca, Sacco, Katiuscia, Colle, Livia, Angeleri, Romina, Enrici, Ivan, Bo, Gianluca, Bara, Bruno 2003. Simple and Complex Extralinguistic Communicative Acts. 26th Annual Meeting of the Cognitive Science Society.

Clark, H. Herbert 1991. Responding to Indirect Speech Acts. – Pragmatics. Ed by S. Davis. Oxford: Oxford University Press.

Clark, H. Herbert 1992. Arenas of Language Use. The University of Chicago Press.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Marmaridou, Sophia 2000. Pragmatic Meaning and Cognition. Amsterdam/Philadelphia: John Benjamins.

Nickerson, Jill, Chu-Carroll, Jennifer 1999. Acoustic-Prosodic Disambiguation of Direct and Indirect Speech Acts. – Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco, California.

Searle, John 1975. Indirect speech acts. – Syntax and semantics 3: Speech acts. Ed by P. Cole, J. L. Morgan. New York: Academic Press.

Stalnaker, Robert 2002. Common ground. Linguistics and Philosophy, 25 (5–6), 701–721.

Treumuth, Margus 2005. A software tool for the Estonian Dialogue Corpus. – Proceedings of Second Baltic Conference on Human Language Technologies, Tallinn, 341–346.

Yasin, Anas 2002. Pragmatics (lectures). [WWW document]. URL http://www.geocities.com/anas_yasin/.

Dialoogiaktide automaatne tuvastamine[1]

Mark Fišel, Taavet Kikas
Tartu Ülikool

1. Sissejuhatus

Dialoogiaktide tuvastamine on dialoogide analüüsi oluline alamülesanne. Üldine eesmärk on tuvastada dialoogi iga lausungi klass, nn dialoogiakt; näiteks lausung võib olla küsimus, rituaal (nagu tervitus või tänamine), direktiiv jne.

Dialoogiakti automaatne tuvastamine on vajalik inimesega suhtlevas dialoogsüsteemis, et määrata kõneleja kavatsus ja reageerida sellele vastavalt. Kui lausung oli küsimus, siis peab dialoogsüsteem leidma vajaliku info ja esitama selle kasutajale, tervituse korral peab süsteem hoopis vastu tervitama jne.

Taust. Seosed kõne helisignaali ja lausungitüüpide vahel on liiga kaudsed ja keerulised selleks, et iseõppiv automatiseeritud süsteem või isegi inimene saaks neid leida. Seepärast tuleks esitada dialoogis esinevad lausungid mingi tunnuste hulgaga. Üks tunnuste valimise viis on tekstipõhine: nt võib määrata võtmesõnad ja defineerida binaarsed tunnused, mis näitavad, kas võtmesõna esineb antud lausungis või mitte. Võtmesõnade esinemisest ja mitteesinemisest võib teha järeldusi selle kohta, millist dialoogiakti lausung väljendab. Teine võimalus on kasutada keelemudelit ja näiteks kodeerida tunnusteks lausungi morfoloogiline kirjeldus. Lisainformatsioonina kasutatakse ka kõne prosoodia parameetreid (hääle toon, kõne kiirus, pauside pikkus jms). Kasuks tuleb ka arvestamine eelnevate, juba määratud dialoogiaktidega – reeglina aitab see samuti potentsiaalsete dialoogiaktide hulka väiksemaks taandada. Näiteks suure tõenäosusega järgneb küsimusele kas info andmine või küsimuse täpsustamine. Mudelite kõrgema täpsuse saavutamiseks on kasulik kombineerida tunnuseid, kuna siis on tunnuste hulk täielikum ja kirjeldab lausungeid täpsemini.

[1] Uurimistööd on toetanud Eesti Teadusfond projekti nr 5685 raames. Autorid tahaksid väljendada erilist tänu prof Mare Koidule, kes on olnud mõlema autori juhendaja.

Tunnuste töötlemiseks (ja dialoogide analüüsimiseks üldiselt) on kaks põhilist meetodite klassi: reeglipõhised ning statistilised meetodid. Reeglipõhiste meetodite puhul määrab inimene ise, millisel viisil ja milliseid tunnuseid peab tuvastamissüsteem interpreteerima (klassikaline näide on juturobot Eliza[2]). Viimasel ajal on populaarsemad statistilised meetodid, mille puhul süsteem ise valib tähtsamaid tunnuseid ja loob nende interpreteerimismudeli, tuginedes statistilisele teadmusele dialoogide ja dialoogiaktide kohta. Enamus rakendatavaid meetodeid on masinõppe meetodid, kus süsteem "õpib", kasutades dialooge, milles lausungite dialoogiaktid on eelnevalt märgendatud; rakendatud on Bayesi klassifitseerijat (Keizer jt 2002), otsustuspuid (Stolcke jt 2000), neurovõrke (tehisnärvivõrke), Markovi peitmudeleid (Ries 1999) jmt.

Ei reeglipõhised ega statistilised meetodid ole siiani andnud tulemuseks dialoogiaktide tuvastamist 100% täpsusega. Inimene ei jõua füüsiliselt analüüsida kõikide tunnuste mõju tulemusele. Statistilised meetodid on tihti edukad mudelite üldistamises ja mitte nii edukad rohkete erandjuhtumite puhul – aga viimaseid on dialoogide analüüsis piisavalt palju. Praegusel tasemel sõltub tuvastamise täpsus tugevasti dialoogiaktide tüpoloogia detailsusest. Kui dialoogiaktid on üldised ja erinevaid akte on vähe, on saavutatud suhteliselt kõrge täpsus (Ries 1999).

Käesolevas artiklis on kirjeldatud otsustuspuude ja neurovõrkude rakendamist aktide tuvastamiseks. Andmete esitamiseks on kasutatud nii tekstipõhiseid kui ka keelemudeli tunnuseid.

2. Dialoogiaktid

Selles artiklis kirjeldatud uurimistöös kasutatakse eestikeelsete dialoogiaktide tüpoloogiat, lühendatult EDiT (Hennoste, Rääbis 2004). Nimetatud tüpoloogia põhineb paljuski Eesti dialoogikorpusel, mis kujutab endast eestikeelsete suuliste dialoogide kogumit. Korpusesse on dialooge kogutud alates 1997. aastast ning 2005. aasta oktoobri seisuga sisaldas see 781 litereeritud teksti, millest suurema osa moodustasid telefoni teel peetud infodialoogid.[3]

[2] Vt nt http://www-ai.ijs.si/eliza/

[3] Aktide tüpoloogia vt Lisa 2.

EDiT-i puhul on lähtutud sellest, et erinevat tüüpi aktid oleksid selgelt eristatavad ning et aktitüüpide loetelu oleks ammendav. Kokku eristatakse 126 erinevat dialoogiakti. Siinjuures on oluline märkida, et eristamine toimub ennekõike aktide funktsiooni, mitte keelelise vormi põhjal. Seetõttu leidub dialoogikorpuses ka akte, mida ei saa üheselt määrata. Aktimäärangute usaldusväärsuse hindamiseks kasutatakse niinimetatud κ-koefitsienti, mis näitab, kui suur osa kahe inimese tehtud aktimäärangutest kokku langeb. Üldiselt loetakse heaks tulemuseks juba κ väärtusi, mis on suuremad kui 0,8. Eesti dialoogikorpuse puhul on saavutatud κ väärtus 0,74. Samas sisaldab EDiT rohkem akte kui enamik teisi tüpoloogiaid.

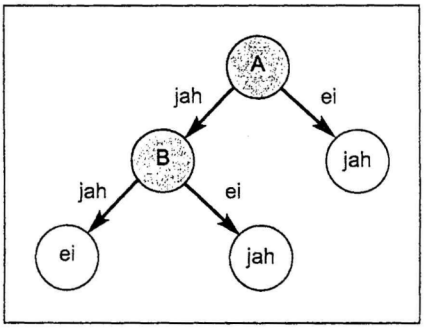
Nimetatud tüpoloogias on igale aktile antud kahest osast koosnev nimi, kus esimene osa tähistab akti üldist rühma ja teine tema täpset funktsiooni. Näiteks tähistab nimi KYE: AVATUD küsimuse eesliikmeks olevat avatud küsimust ning RIJ: VASTUTERVITUS rituaalide rühma kuuluvat vastutervitust. Lõik transkribeeritud ja märgendatud dialoogist võib välja näha umbes nii:

H: [tere= | RIJ: VASTUTERVITUS |
palun]=`öelge kuskohas Pärnus on ilusalong Di`one. | KYE: AVATUD |

3. Otsustuspuud

Otsustuspuuks (ingl k *decision tree*) nimetatakse hierarhiliste otsuste puukujulist esitust. Otsustuspuu moodustub kahesugustest tippudest – sisemistest tippudest ja lehtedest (joonis 1). Puu sisemised tipud kujutavad endast otsuseid, mida on vaja teha. Sisemisest tipust väljuvad servad tähistavad aga valikuid, mida me antud otsuse langetamisel teha saame. Liikudes puu juurest mööda servi järk-järgult edasi, jõuame lõpuks otsustuspuu leheni. Erinevalt sisemistest tippudest ei ole siin enam tegu otsuse langetamise, vaid lõpliku vastusega. Kuna puu läbimisel tehtud otsused lähtuvad puu sisendatribuutidest, on sisendatribuutidega määratud ka vastus.

Otsustuspuude tõlgendamine on seega intuitiivselt küllaltki lihtne. Nende konstrueerimine on aga mõneti keerulisem. Esiteks võib mingit probleemi kirjeldavaid atribuute olla nii palju, et nende seast oluliste väljavalimine on ilma süstemaatilise lähenemiseta peaaegu võimatu. Teiseks ei ole paljud otsustused vähemalt pealtnäha olulisuse järgi hierarhiliselt järjestatavad. Mõlema probleemiga puutume



Joonis 1. Kahe atribuudiga binaarne otsustuspuu

kokku, kui proovime dialoogiakte tuvastavaid otsustuspuid "käsitsi" konstrueerida. Seetõttu vajame algoritmi, mis teeks selle töö meie eest automaatselt ära.

Algoritme otsustuspuude automaatseks konstrueerimise on välja pakutud isegi mitmeid. Dialoogiaktide tuvastamiseks oleme neist kasutanud Ross Quinlani ID3 algoritm (Russell, Norvig 2003). Nimetatud algoritm on masinõppe algoritm. Nagu seda tüüpi algoritmidel tavaks, toimub õppimine näidete ning statistiliste meetodite abil. Algoritmi eesmärgiks on koostada otsustuspuu, mis oleks ühest küljest kooskõlas näidetega ning teisest küljest üldistaks neid, st oleks võimeline toime tulema ka sarnastes, kuid mõneti erinevates situatsioonides.

Algoritmi toimimiseks on vaja piisavalt suurt näidete hulka ehk treeninghulka, millel on fikseeritud näiteid kirjeldav atribuutide hulk. Atribuudid on kõikide näidete jaoks samad, kuid atribuutide väärtused võivad näiteti erineda. Atribuudid paigutatakse otsustuspuusse sedamööda, kuidas atribuudi väärtuste järgi treeninghulka alamhulkadeks jagades kahaneb alamhulkade entroopia. Entroopia on suurus, mis iseloomustab mingi hulga homogeensust – mida sarnasemad on hulga elemendid, seda homogeensem on see hulk ja seda väiksem on entroopia. Otsustuspuude puhul huvitab meid ennekõike treeninghulga entroopia väljundväärtuse suhtes.

Eelpool kirjeldatud algoritm koostab küll treeninghulka täpselt määratleva otsustuspuu, kuid see puu ei oma üldistusvõimet. Seda seetõttu, et otsustuspuusse satuvad kõik atribuudid, sh ka need, mille korral alamhulkade entroopia vähenemine on tegelikult tühine. Üldiselt on see probleem tuntud andmete ületäpsustamise nime all (ingl

k *data overfitting*) (Russell, Norvig 2003). Selle vältimiseks hinnatakse entroopia vähenemise olulisust veel hii-ruut testi abil ning heidetakse ebaolulised atribuudid kõrvale. Tulemuseks on nn lihtsustatud otsustuspuu, mis toimib praktilistes olukordades enamjaolt paremini kui lihtsustamata puu.

Dialoogiaktide tuvastamiseks oleme kasutanud binaarse hargnemisteguriga otsustuspuid. Ühest küljest on neid kõige lihtsam konstrueerida ning teisest küljest on suur osa akte iseloomustavatest atribuutidest just jah–ei-väärtustega. Samuti tuleb märkida, et ka suurema hargnemisteguriga otsustuspuud on taandatavad binaarseteks puudeks. Lisaks oleme lähtunud sellest, et iga dialoogiakti tuvastamiseks kasutatakse erinevat otsustuspuud. Esiteks teeb see võimalikuks binaarsete puude kasutamise. Teiseks päästab see meid ülisuurtest puudest ja võimaldab üksikuid puid paindlikumalt konstrueerida.

Siiski ei piisa otsustuspuude konstrueerimiseks üksnes ID3 algoritmist – lahendamata jääb probleem, milliste atribuutidega akte üldse kirjeldada. ID3 algoritm paigutab küll akti kirjeldavad atribuudid loogilisse struktuuri, kuid ei ütle, millised need atribuudid olema peaksid, v.a nii palju, et ebaolulised atribuudid jäetakse otsustuspuust välja.

Kuna dialoogiaktide tüpoloogias viitab akt ennekõike tähendusele, mitte keelelisele vormile, siis on tegu küllaltki keerulise probleemiga, mille täielik lahendamine on aktide tähendusest aru saamata tõenäoliselt võimatu. Siiski oleme püüdnud aktide kirjeldamise probleemile läheneda järgmiste meetoditega.

Neist kõige lihtsamas proovitakse aktitüüpe eristada teatud märksõnade ja intonatsioonimärkide põhjal. Kuna võimalikke märksõnu on palju, siis otsib arvuti dialoogikorpusest üles kõige sagedamini esinevad märksõnade vormid ja intonatsioonimärgid ning konstrueerib nende põhjal otsustuspuu.

Järgmine loogiline samm aktide tuvastamisel on dialoogide morfoloogiline analüüs. Seni on selleks otsustuspuude juures kasutatud Eesti Keele Instituudi morfoloogilist analüsaatorit. Morfoloogiline analüüs võimaldab teha aktide kohta üldistusi. Näiteks ei ole enam probleemi, et ühe sõna erinevaid vorme nähakse erinevate märksõnadena. Tänu sellele väheneb ka otsustuspuu konstrueerimiseks vajalik näidete hulk.

Kõrvale ei saa jätta ka naaberaktide olulisust – EDiT-i põhjal moodustavad mitmed aktid koos esinevaid naaberpaare. Üks võimalus on lihtsalt vaadelda akti naaberaktides sagedasti esinevaid atribuute (Kikas 2005). Teine lähenemisviis, mis annab paremaid tulemusi, seisneb selles, et esiteks määratakse vaadeldava akti naaberlausungite täpsed aktitüübid ja seejärel nende põhjal akti enda tüüp. Tagasilöögina suureneb küll valede aktimäärangute arv, sest ühes aktimäärangus tehtud viga kandub teistesse määrangutesse edasi.

Eksperimendid. Otsustuspuude meetodi testimiseks viidi läbi rida eksperimente, mis keskendusid rituaalide, direktiivide ja küsimuste tuvastamisele (Kikas 2005). Seejuures lähtuti sellest, et otsustuspuude valede aktimäärangute protsent ei oleks treeninghulgal suurem kui 1%. Puude lihtsustamisel võeti aluseks hii-ruudu väärtus 3,84. Välistamaks ebaolulisi atribuute, kasutati otsustuspuudes üksnes atribuute, mida esines treeninghulgas vähemalt kahel korral. Lisaks piirati ühte aktitüüpi kirjeldavate atribuutide arvu, nii et see ei oleks suurem kui 128.

Tulemused. Eksperimentides tuli ilmekalt esile erinevaid atribuute kasutavate otsustuspuude erinevus (tabel 1). Keskmise tuvastusprotsendi järgi osutusid selgelt kõige edukamaks morfoloogilisi atribuute kasutavad kombineeritud puud. Samas tegid kombineeritud otsustuspuud ka kõige rohkem valesid aktimääranguid. Seega ei pruugi need praktikas ilmtingimata eelistatuks osutuda. Kuigi morfoloogiliste atribuutidega otsustuspuud osutusid keskeltläbi mõnevõrra edukamateks kui märksõnu ja intonatsioonimärke kasutavad puud, ei saa ühte teisele selgelt eelistada. Leidus ka akte, mida tuvastati märksõnade ja intonatsioonimärkide abil tunduvalt edukamalt. Üheks selliseks aktiks oli näiteks RIE: TERVITUS.

Tabel 1. Dialoogiaktide tuvastusprotsent, kasutades märksõnu ja intonatsioonimärke (MRK) ning morfoloogilisi atribuute (MRF)
Tabelist on välja jäetud aktid, mille esinemissagedus 200 dialoogis oli alla 10

	MRK tuvastusprotsent		MRF tuvastusprotsent	
Akt	Tavaline	Naaber-aktidega	Tavaline	Naaber-aktidega
RIE: HÜVASTIJÄTT	41	40	97,9	90,9
RIE: KUTSUNG	93,5	100	97,8	100
RIE: LÕPUSIGNAAL	43,2	0	30,1	25

Akt	MRK tuvastusprotsent		MRF tuvastusprotsent	
	Tavaline	Naaber-aktidega	Tavaline	Naaber-aktidega
RIE: TERVITUS	80,9	79,3	0	87,9
RIE: TÄNAN	69,8	90,3	96,6	97
RIJ: KUTSUNGI VASTUVÕTMINE	58,8	100	29,3	100
RIJ: LÕPETAMISE VASTUVÕTMINE	18,1	0	0	25
RIJ: PALUN	88,9	96	98,6	100
RIJ: VASTUHÜVASTIJÄTT	27,5	0	80,3	100
RIJ: VASTUTERVITUS	78,6	72,7	12,4	81,5
RY: TUTVUSTUS	73,9	98,2	29,3	98
DIE: ETTEPANEK	11,1	0	45,1	15,4
DIE: PAKKUMINE	15,6	14,3	37,4	0
DIE: SOOV	25,4	44	36,4	30,2
DIJ: EDASILÜKKAMINE	39,1	16,7	82	77,8
DIJ: INFO ANDMINE	45,2	48,8	33,5	37,9
DIJ: INFO PUUDUMINE	66,7	0	13,3	28,6
DIJ: NÕUSTUMINE	6,5	37,5	12,9	0
DIJ: PIIRATUD NÕUSTUMINE	0	0	0	0
KYE: ALTERNATIIV	14,4	100	43,3	14,3
KYE: AVATUD	28,7	33,3	40,9	67,9
KYE: JUTUSTAV KAS	12,9	52,9	28,4	35,1
KYE: MUU	8,3	0	11,1	0
KYE: SULETUD KAS	24,5	45,5	8,3	37,5
KYE: TÄPSUSTAV	8,3	0	8,3	0
KYE: VASTUST PAKKUV	16,4	22,6	14,7	17,1
KYJ: ALTERNATIIV: ÜKS	6,7	0	0	20
KYJ: EDASILÜKKAMINE	12,5	0	48,3	20
KYJ: EI	23,3	75	69,6	30
KYJ: INFO ANDMINE	15,4	38,6	19,9	31,2
KYJ: INFO PUUDUMINE	8,3		33,3	37,5
KYJ: JAH	29,3	31,4	6,5	38,6
KYJ: MUU	4,2	0	0	14,3
KYJ: NÕUSTUV EI	8,3	0	54,2	62,5
Keskmine tuvastusprotsent	32,5	36,4	35,9	44,7
Keskmine veaprotsent	0,8	1	0,8	1,3

Samuti ilmnes, et akte, mille esinemissagedus oli 200 akti seas väiksem kui 10, ei õnnestunud ühegi meetodiga edukalt tuvastada. Siinjuures tuleb märkida, et morfoloogilisi atribuute kasutavad otsustuspuud olid väikese treeningnäidete arvu korral keskeltläbi edukamad kui märksõnu ja intonatsioonimärke kasutavad otsustuspuud. Seda tänu morfoloogiliste atribuutide suuremale üldistusvõimele.

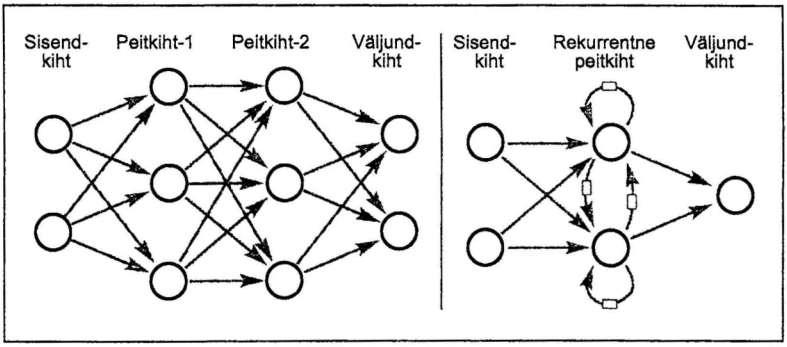
Samuti tuleb märkida, et otsustuspuud osutusid treeninghulga suuruse, hii-ruudu väärtuse ja lubatud veaprotsendi osas küllaltki tundlikuks. Seega õnnestub täpse häälestusega kõikide otsustuspuude tulemusi tõenäoliselt mõnevõrra veel parandada.

4. Neurovõrgud

Neurovõrgud on üks masinõppe liik. Nende eripära seisneb selles et arvutamise/õppimise süsteem on koostatud väiksematest lihtsatest arvutusühikutest, mida nimetatakse (tehis-)neuroniteks. Neurovõrgud võtavad malli inimese aju ehitusest – nii aju kui neurovõrgu neuronid on omavahel ühendatud suunatud kaalutud seostega, mida nimetatakse sünapsseosteks. Võrkude õppimine ehk treenimine seisneb sünapsseoste kaalude iteratiivses muutmises, eesmärgiga saavutada väljundit, mis oleks võimalikult lähedane soovitud väljundile.

Selles töös kasutatakse mitmekihilist pertseptroni (tajurit) ja rekurrentset kihilist neurovõrku. Mõlema neuronid on organiseeritud kihtidesse; on paika pandud sisend- ja väljundkiht. Pertseptroni puhul on lubatud sünapsseosed ainult eelnevast kihist järgnevasse kihti. Sisendsignaal läbib pertseptroni algusest lõpuni ning genereerib väljundsignaali. Seega sõltub ühe iteratsiooni väljund ainult selle iteratsiooni sisendist ning võrgu seoste kaaludest. Rekurrentsel võrgul on lubatud ka sünapsseosed samasse või eelnevasse kihti. Need seosed on viivitatud, mis tähendab et sihtneuroni sisendiks saab lähteneuroni eelmise iteratsiooni väljund. See võimaldab töödelda fikseerimata pikkusega sisendite jadasid. Joonisel 2 on toodud mitmekihilise pertseptroni ja rekurrentse võrgu näited.

Eeltöötluse meetodid. Neurovõrgud on suhteliselt "õrn" tehnika. Treenimise tulemus sõltub tugevalt treenimise parameetritest, võrgu topoloogiast ning eriti sisend- ja väljundandmete esitusmeetodist. Võrgu sisendiks ja väljundiks on reaalarvude vektorid ning tähtis on kasutada eeltöötlusmeetodit, mis ühest küljest kirjeldaks sisenddi-



Joonis 2. Mitmekihiline pertseptron kahe peitkihiga (vasakul) **ja lihtne rekurrentne võrk** (paremal, kastidega sünapsseosed on viivitatud)

alooge võimalikult täpselt ning kaotaks nende kohta võimalikult vähe informatsiooni ja teisest küljest oleks "arusaadav" neurovõrgu jaoks, st võrk oskaks hästi teha vahet kahel erineval kodeeritud lausungil.

Väljund kodeeritakse kahendvektoriks, kus kõik vektori komponendid on üksüheses vastavuses olemasolevate dialoogiaktidega. Iga komponendi väärtus on 1, kui lausungi akt on sellele komponendile vastav dialoogiakt, ning 0 vastasel korral. Selliseid vektoreid nimetatakse tunnusvektoriteks.

Üks kasutatud meetoditest on võetud projektist WEBSOM (Honkela jt 1997). Iga sõna baasvormile seatakse vastavusse juhuslikult genereeritud n-mõõtmeline vektor, mis on seda pikem, mida unikaalsem on antud sõna vastavalt lausungi dialoogiaktile. Lausungile vastab sinna kuuluvate sõnade vektorite summa. Seega, iga aktispetsiifiline sõna, mille vektor on teistest pikem, määrab n-mõõtmelise ala, kuhu tõenäoliselt kuuluvad seda akti väljendavate lausungite vektorid. See eeltöötlusmeetod on ainus, mida saab kasutada nii pertseptroni kui ka rekurrentsete võrkudega. Pertseptroni puhul arvestab see meetod ainult vaatluse all oleva lausungiga ega arvesta eelnevate lausungite märgenditega (aktidega). Rekurrentsele võrgule esitatakse kõik lausungite vektorid sama iteratsiooni jooksul. Seega on akti määramisel kasutatud ka infot eelnevate lausungite aktidest.

Ülejäänud meetodid on mõeldud puhtalt rekurrentsete võrkude jaoks ja kodeerimine toimub sõnade, mitte lausungite kaupa. Ühe iteratsiooni jooksul saab võrgule esitada nii ühe lausungi (sel juhul loetakse

võrgu väljundit alles pärast viimase sõna esitamist) kui ka terve dialoogi (ühe iteratsiooni jooksul esitatakse järjest kõik dialoogi sõnad ning väljundit loetakse iga lausungi viimase sõna esitamisel). Kaks sellist meetodit oli ajendatud projektist WEBSOM, selle erinevusega, et vektoreid ei summeerita, vaid esitatakse võrgule ükshaaval. Esimene variant oli esitada neidsamu juhuslikke vektoreid. Teises variandis asendati juhuslike reaalarvude vektoreid juhuslike kahendvektoritega, selle motivatsiooniga, et neurovõrgud töötavad paremini diskreetsete andmetega, juhul kui väljund on diskreetne (nagu antud ülesandes).

Alternatiivne meetod, mida on kasutatud arvutilingvistikas rekurrentsete võrkudega (Wermter 2000), on tähtsusvektori kodeerimine. Tähtsusvektori iga element vastab ühele dialoogiaktile ning võrdub antud sõna esinemissagedusega antud aktis, jagatud sõna üldise esinemissagedusega terves korpuses. Seega, kui sõna esineb üks kord RIE lausungis ja kaks korda KYE lausungis, siis RIE aktile vastav komponent võrdub 1/3 ning KYE komponent võrdub 2/3; ülejäänud aktidele vastavad komponendid võrduvad 0.

Kõik seni kirjeldatud meetodid koostavad ja kasutavad sõnastikku – sest igas meetodis kasutatakse kodeerimiseks mingit infot märgistatud väljundi kohta. Sõnastik koostatakse treeninghulga põhjal ning testimisel/rakendamisel otsitakse sõnad sõnastikust üles. Ühest küljest on see hea, kuna eeltöötlus on arvutuslikult kallis ja pärast treenimist pole enam vaja arvutusi sooritada. Teisest küljest aga ei saa pärast treenimist enam lisada uusi sõnu, sest puudub info väljundi kohta ja tundmatuid sõnu lihtsalt ignoreeritakse.

Viimane kasutatud eeltöötlusmeetod asendab lausungi selle morfoloogilise kirjeldusega – sõnaliik, nimisõna kääne, arv, tegusõna tegumood, aeg jne. Iga parameetrit kodeeritakse tunnusvektoriga. See meetod ei ole sõnastikupõhine, seega tundmatute sõnade hulk oluliselt väheneb, tundmatud on ainult keeles ebakorrektsed sõnad.

Eksperimendid. Mitmekihilise pertseptroni topoloogiatest kasutati ühe ja kahe peitkihiga võrke. Ühe kihiga võrk on teoreetiliselt suuteline lähendama suvalist pidevat funktsiooni, kuid samas ei ole garanteeritud, et siin kasutatud ja enimlevinud treeningalgoritm (vea tagasilevi, *error backpropagation*) leiab ideaalse lahenduse. Tihti annab sama algoritmiga treenitud kahe peitkihiga pertseptron paremaid tulemusi.

Rakendatud rekurrentsete võrkude topoloogiatest on enimlevinud lihtne rekurrentne võrk (ingl k *simple recurrent network*) ning selle laiendatud variant. Lihtsas rekurrentses võrgus on peidetud kiht täiesti sidus viivitatud seostega. Seega info eelmisest sammust satub peidetud kihti n-ö kodeeritud kujul. Siin kasutatud laiendatud variandis olid lisatud ühendused väljundkihist peidetud kihti. Seega, lisaks kodeeritud infole oli kättesaadav ka "puhas" info eelmisel sammul määratud dialoogiaktist.

Eksperimendid sooritati kahe erineva dialoogiaktide hulgaga. Üks hulk koosnes üldistest dialoogiaktidest (nt RIE, KYE, YA) ning teine täpsetest aktidest (nt KYE: AVATUD, DIJ: INFO ANDMINE).

Tulemused. Mitmekihilise pertseptroni eksperimentide tulemused olid suhteliselt head ning neid on kirjeldatud (Fishel 2005). Seda tehnikat hiljem rakendades ei suudetud saavutada niisama kõrget täpsust, seega varem publitseeritud tulemused ei paista olevat usaldusväärsed.

Täpsete dialoogiaktide klassifitseerimise täpsus jäi alla 15%, seega detailne klassifitseerimine praeguste meetoditega ei õnnestunud.

Tabel 2. Laiendatud lihtsa rekurrentse võrgu tuvastatud aktide sagedustabel üldiste dialoogiaktide süsteemis

Tegelik dialoogiakt	Võrgu väljund (protsent dialoogiakti suhtes)									
	DIE	DIJ	IL	KYE	KYJ	RIE	RIJ	RY	VR	YA
DIE	27,4	3,1	7,9	8,3	5,3	0,1	3,1	0,4	0,4	0
DIJ	8,5	48,1	8,8	6,2	10,4	1	0,2	0	3,2	0
IL	12,9	1,8	21,9	9,6	11,4	1	0,1	0	1,2	0
KKE	0	0	0	0,1	0	0,1	0,1	0	0,7	0
KYE	19,7	9	21,9	41,4	14,8	1,1	1,0	0	4,7	0
KYJ	17,3	21	17,5	21,7	33,4	1	1,4	0,4	20,4	33
RIE	0,7	0,2	0	0,3	1,5	86,1	8,9	0	1	0
RIJ	0,2	0,2	0	0,7	1,1	8,2	84,1	2	1,1	0
RY	0,7	0	0	0,2	0,5	0,1	0,4	97,2	0,3	0
SEE	1,5	0	2,6	0,8	1,5	0	0	0	0	0
SEJ	1,1	0,1	0,9	0,7	0,9	0	0	0	0,7	0
VR	3,5	14,5	6,1	4,8	8,6	0,1	0,1	0	64,8	66,7
YA	6,6	2	9,6	5,3	10,5	0,4	0,4	0	1,5	0
Aktide osakaalud (%)	**6,7**	**17,6**	**1,7**	**14,4**	**11,5**	**10,3**	**13,2**	**3,7**	**20,9**	**0,0**

Treenitud võrkude täpsused üldise dialoogiaktide hulgaga olid 40% ja 57% vahel. Parim tulemus (57,58%) saavutati modifitseeritud lihtsa rekurrentse võrguga, kasutades tähtsusvektori eeltöötlust. Selle poolt klassifitseeritud aktide sagedustabelist (tabel 2) on näha et neurovõrk õppis suhteliselt hästi tuvastama mõningaid sagedasemaid (DIJ, RIE, RIJ, VR) ja ühte harvaesinevat (RY) dialoogiakti, mille leksikon on suhteliselt piiratud. See tähendab arvatavasti, et võrk eraldas selle akti võtmesõnad. Kõikide eksperimentide tulemused on madalad ja tuvastamistäpsused on sarnased, mis arvatavasti tähendab, et rekurrentsed võrgud ei suutnud ära õppida andmetesiseseid sõltuvusi ning vaatamata eeltöötlusmeetodile on sisendvektorid neurovõrgu jaoks lihtsalt erinevad juhuslikud vektorid.

5. Võrdlev analüüs ja edasiarendamise võimalused

Neurovõrgud ja otsustuspuud on oluliselt erinevad masinõppetehnikad ning kummalgi on omad tugevad küljed. Otsustuspuud on valge kasti tüüpi meetod, see tähendab, et otsustuspuu struktuur on inimese poolt hästi tõlgendatav. Seevastu neurovõrgud on musta kasti tüüpi meetod, sest treenitud võrgu parameetrid ei ole inimese jaoks hästi analüüsitavad. Teine erinevus seisneb selles, et võrgud näitasid dialoogiaktide tuvastamise puhul head üldistusvõimet. Samas oskasid otsustuspuud arvestada hästi väikeste erinevustega sisendis. Autoritel on plaanis mõlemad meetodid kombineerida. Ühe võimalusena võib neurovõrkude abil esiteks tuvastada akti üldise klassi ja seejärel puude abil täpse dialoogiakti.

Otsustuspuude puhul tuleb ühe edasiarendusena kõne alla ka keerukamate keeleliste struktuuride kasutamine – seni oleme piirdunud üksikute märksõnade, intonatsioonimärkide ja morfoloogiliste atribuutidega või siis nende paaride ja kolmikutega (Kikas 2005). Samuti võib katsetada alternatiivseid puude konstrueerimise algoritme ning pöörata suuremat rõhku puude optimeerimisele ning veaprotsendi vähendamisele. Lisaks tasub katsetada märksõnapõhiste ja morfoloogiapõhiste otsustuspuude ühendamist.

Neurovõrkude osas on tulevikus plaanis rakendada pikka lühiajalist mälu (LSTM, Hochreiter, Schmidhuber 1997). See tehnika näitas mitmel korral oluliselt paremaid tulemusi, võrreldes tavaliste rekurrentsete võrkudega. Nt on ta suuteline leidma pikaajalisi sõltuvusi,

ajalisi mustreid jms ning on võimalik, et LSTM-iga õnnestub see, mis ei õnnestunud rekurrentsete võrkudega – ära õppida andmetesisesed sõltuvused. Teine idee, mida on plaanis proovida, on kombineerida sisendis sõnastikupõhiseid meetodeid (mis oleksid leksikoni rollis) morfoloogilise töötlusega (mis esineks grammatilise kirjeldusena).

Kirjandus

Fishel, Mark 2005. Dialogue Act Recognition in Estonian Dialogues using Artificial Neural Networks. – Second Baltic Conference on Human Language Technologies, Tallinn, 231–236.

Hennoste, Tiit, Rääbis, Andriela 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Hochreiter, S., Schmidhuber, J. 1997. Long Short-Term Memory. – Neural Computation 9(8), 1735–1780.

Honkela Timo, Kaski Samuel, Lagus Krista, and Kohonen Teuvo 1997. WEBSOM – Self-Organizing Maps of Document Collections. Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Helsinki University of Technology, Finland, 310–315.

Keizer, S., Op den Akker, R., Nijholt, A. 2002. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. – Proceedings 3rd SIGdial Workshop on Discourse and Dialogue. Ed by K. Jokinen, S. McRoy. Philadelphia, Pennsylvania, 88–94.

Kikas, Taavet 2005. Dialoogiaktide tuvastamine eestikeelsetest dialoogides otsustuspuude abil. Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut.

Ries, K. 1999. HMM and Neural Network Based Speech Act Detection. – International Conference on Acoustics and Signal Processing. http://www.is.cs.cmu.edu/papers/speech/ICASSP99/ICASSP99-klaus.pdf

Russell, S., Norvig, P. 2003. Artificial Intelligence: A Modern Approach. Prentice Hall.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M. 2000. Dialogue Act Modeling for automatic tagging and recognition of conversational speech. – Computational Linguistics 26(3), 339–373.

Wermter, S. 2000. Neural Network Agents for Learning Semantic Text Classification. – Information Retrieval 3(2), 87–103.

Sõnastike haldussüsteem Eesti Keele Instituudis[1]

Andres Loopmann, Kati Sein, Ülle Viks
Eesti Keele Instituut, Tallinn

1. Sissejuhatus

Kolm ja pool aastat tagasi leidis Margit Langemets Eesti Keele Instituudi (EKI) keelekogudest ülevaadet tehes, et aastate jooksul on kogutud ja korrastatud päris kaunis hulk elektroonilist keelevara ning sõnastikke, millest mõned erinevates keeletehnoloogilistes rakendustes juba ka laia kasutajaskonda teenivad (Langemets 2002). Samas tõdes autor, et meil puudub paar olulist asja: (1) ühtne struktuurimärgistuse standard, mis ühtlustaks kõigi meie sõnastike andmete esitust; (2) standardit toetav sõnastike haldamise tarkvara. Praegu võime rõõmustada, et viimaste aastate töö on olukorda tublisti parandanud – liigume jõudsasti ühtse märgistuse poole, millesse viime üle olemasolevaid sõnastikke, ning uusi luua saame juba oma majas sündinud tarkvara – sõnastike haldussüsteemi – abil. Sõnastikutekstide märgistamises on kõikjal toimunud analoogiline areng: tüpograafiliselt (kirjastiile näitavalt) ja deskriptiivselt (lineaarselt) märgistamiselt on jõutud üldistava, s.o teksti struktuuri kirjeldava ning algoritmilist töötlemist võimaldava märgistuseni (Langemets 2000).

Pisut ajalugu. Alates 1978. aastast on EKI-s loodud paarkümmend elektroonilist sõnastikku, mis oma teostuselt on üsna erinevad: tehnilised võimalused on ligi 30 aastaga palju muutunud.

Esimene elektrooniline sõnastik EKI-s oli "Õigekeelsussõnaraamat" (ÕS 1976), mis tipiti arvutisse tabelina, kus sõnaartikli iga struktuurielemendi jaoks oli ette nähtud kindel positsioon reas, nt positsioonides 1–35 oli märksõna, positsioonides 36–37 muuttüübi number jne. Analoogiline süsteem pisut teisel kujul oli kasutusel paaris väiksemas sõnastikus, kus sõnaartikli elemente eraldas tühik ja elemendi tähendus oli määratud tema järjekorraga reas.

---

[1] Aitäh Margit Langemetsale ja Peeter Pällile kasulike märkuste eest. Tööd on toetanud: EKRK47, ETF5969, SF0052488s03.

Mitmed suuremad sõnaraamatud (nt VMS, VES jt) sisestati arvutisse polügraafilise märgistusega, selleks et neid fotolao vahendusel trükki anda: eraldi koodidega tähistati kirjastiilid, taanded, suurtähed jne.

1980. aastate teisel poolel võeti kasutusele sisuline deskriptiivne märgistus, kus iga struktuurielement oli varustatud prefiksilaadse tähtkoodiga, nt märksõna ees oli kood *m+*, tähenduse ees *t+*, grammatilise info ees *g+* jne. Suurem osa EKI sõnastikke ongi märgistatud nii, sõltuvalt sõnastikust võis koodide valik olla erinev. Selline märgistus võimaldab edukalt kirjeldada sõnaartikli struktuuri – juhul kui elemendid paiknevad lineaarselt: uue elemendi algus on ühtlasi eelmise lõpp. Hierarhilist paigutust, kus üks element paikneb teise sees, niiviisi edasi anda ei saa. Deskriptiivse märgistuse suurim puudus on aga struktuurikontrolli puudumine. Sõnaartiklite struktuuri saab analüüsida alles tagantjärele, võrreldes artiklite koodijadasid omavahel ja otsides nende hulgast ebatüüpilisi (ehk vigaseid). Kuid just korrektne struktuur on see, mis tagab sõnastiku hilisema arvutitöötluse korrektsuse, nii keeletehnoloogilistes rakendustes kui ka sõnastiku enda edasiarendamisel. Ülevaade EKI elektrooniliste sõnastike varasemast (kuni 1990) seisust on artiklis (Viks 1990).

Esimene katse siduda sõnastiku sisestamisega struktuurikontroll tehti EKI-s 1990. aastate alguses, kui oli valminud "Eesti–vene sõnaraamatu" (EVS) 1. köite käsikiri. Koostöös TPI-ga loodi tarkvarasüsteem, mis sõnaartikli sisestamise igal sammul andis ette just selle valiku struktuurielemente, mis on antud kontekstis võimalikud. Nt pärast märksõna sisseviimist sai valida kas tähendusnumbri või seletuse või vaste, aga vaste järel võis tulla ainult vene grammatika jne. Süsteemi väljundiks oli tekstifail, kus iga struktuurielemendi ees oli numbriline kood (sisuliselt sama mis tähtkood plussiga). Artiklite struktuur oli küll kontrollitud ja vastas etteantud kirjeldusele, kuid see struktuur oli ikkagi lineaarne: element, mille sees juhtus olema mõni teine element, lõhuti tükkideks. Nt kui näite tõlke keskel sattus olema rektsiooniküsimus, siis tekkis kahe sisulise elemendi asemel kolme elemendi jada: "tõlge", "rektsioon", "tõlke jätk":

<näide>mängib tunnetega <tõlge>игр"ает <rektsioon>чьими <tõlke jätk>ч"увствами

EVS jäi selle süsteemi ainsaks rakenduseks – tema kohandamine teiste sõnastikega oleks nõudnud ulatuslikku programmeerimistööd.

Vana süsteemiga on sisestatud EVS-i 3 köidet, alates 4. köitest läksime üle uuele sõnastike haldussüsteemile, mida tutvustame lähemalt 3. alaosas. Aga enne räägime märgistuskeelest XML, mis on uue süsteemi aluseks.

2. XML

Sõnastikuandmete esitamise vorminguks oleme valinud XML-i, mida ka mujal maailmas sõnaraamatute jm keeleressursside kirjeldamisel laialdaselt kasutatakse. XML (*Extended Markup Language*) kujutab endast standardiseeritud kirjelduskeele SGML (*Standard Generalized Markup Language*, ISO standard 8879) edasiarendust (XML 1.0). Märgistusega kodeeritakse andmete paigutus dokumendis ja dokumendi loogiline struktuur. Selleks et veenduda, kui üldine see standard tänapäeval on, piisab, kui lehitseda paari viimast Euralexi leksikograafiakonverentsi kogumikku: näiteks 2002. a kirjeldati tšehhi kirjakeele sõnaraamatu teisendamist leksikaalseks XML-andmebaasiks (Smrz 2002), ungari sõnastiku töötlust XML-vahenditega (Prószéky, Kis 2002), inglise slängisõnastiku hoidmist XML-baasina (Dalzell, Victor, Williams 2002) jne. Ka põhjalik kakskeelsete arvutisõnastike ülevaade (Calzolari jt 2001) toob sõnastike ühe keeletehnoloogilise põhivahendina esile XML-i.

2.1. XML-vormingu eelised

XML-vormingul on mitmeid omadusi, miks ta just sõnastike esitamiseks hästi sobib. Olulisim neist on info struktureeritus sisu järgi, mis võimaldab sooritada päringuid ükskõik millise olemasoleva tunnuse järgi. XML-andmed paiknevad "kindlas kohas", teatud tähiste ehk siltide (ingl *tag*) vahel, tänu millele saab teha täpsemaid päringuid: otsides nt tüübinumbrit 8, me leiamegi ainult tüübinumbrid, samas muud kaheksad jäävad kõrvale. Struktureerimata (nt ainult küljendatud) tekstis ei saa vahet teha, milline leitud kaheksatest on tüübinumber. Oluline on ka see, et XML-i kasutamiseks ei pea tingimata soetama mingit spetsiaalset tarkvara (andmebaasisüsteeme vms), sest andmeid ja struktuuri kirjeldust hoitakse lihttekstina, mida saab lugeda ning parandada ka kõige lihtsama tekstiredaktoriga.

XML-vormingu põhiomadused võib kokku võtta järgmiselt (vt nt MSDN Library – October 2002):

1) XML võimaldab igaühel ise defineerida oma märgistatud struktuure informatsiooni salvestamiseks;
2) XML-vorming on avalik, hästi defineeritud ja laialdaselt kasutatav, mis võimaldab lihtsamalt vahetada informatsiooni erinevate platvormide (Windows, Linux jm) ning andmetöötlusprogrammide vahel;
3) XML põhineb Unicode standardil, mis võimaldab lihtsamalt luua rahvusvahelisi rakendusi;
4) XML-skeemi (ingl *schema*) kasutamise korral saavad rakendused andmete struktuuri ning andmetüübi kontrollil toetuda standardsetele XML parseritele;
5) XML-vorming on tekstipõhine, mis hõlbustab dokumenteerimist ja teeb andmed paremini loetavaks;
6) fakultatiivsed omadused, nt XML-andmete küljendusinfo defineeritakse eraldi failis, mitte koos andmetega andmefailis;
7) XML-andmete töötlemise tarkvara on kättesaadav erinevatel platvormidel;
8) XML-dokumentide töötlemisel on võimalik laialdaselt kasutada HTML standardi jaoks loodud infrastruktuuri, k.a HTTP protokolli ning brausereid;
9) XML-dokumente on lihtne luua ning XML-projektid on kiiresti realiseeritavad.

Kui universaalsus, avatus ja tarkvara kättesaadavus on suureks eeliseks igasuguse standardi puhul, siis XML pakub sõnastike töötluseks hõlpsasti rakendatavaid lisavõimalusi. Et XML-vormingus sõnastikku võib käsitleda andmebaasina, siis saab temaga ette võtta ka samasuguseid toiminguid. Kesksel kohal on struktureeritud päringud: otsing kindlast elemendist (nt järjend *viilu* elemendist "homonüüm") või sõnaartikli teatud piirkonnast (nt stiilimärgend *argi* päise tsoonist – mitte näidetest), otsing elemendi olemasolu või puudumise järgi (nt EVS-i artiklid, kus puudub märksõna vaste), otsing elemendi korduste arvu järgi (nt märksõnad, millel on vähemalt 7 vastet) jne. Keeleuurimise seisukohast on oluline võimalus sõnavara ümber sortida ja mitme kandi pealt sõeluda (nt sõnastikuartiklites oleva erialainfo järgi või tähenduste hulga järgi jne). Häid võimalusi pakub sõnastikuteksti ümberstruktureerimine koos järgneva sortimisega (nt lihtsa kakskeelse sõnastiku saab ümber pöörata, kui vahetada tõlkevaste ja märksõna positsioonid). XML-märgistus võimaldab sõnas-

tiku infot statistiliselt analüüsida, mis tõhustab keeleuurimist ja leksikograafilist tööd. Rakenduste jaoks on kasulik võimalus teha elementide alamosa väljavõtteid (nt lühendatud sõnastikuartikleid tasku- või mobiilisõnastiku jaoks, temaatilisi sõnastikke jm). Korrektselt ja sisule vastavalt struktureeritud sõnastikku on üsna hõlbus kasutada erinevates keeletehnoloogilistes rakendustes, nt keeleõppeprogrammides ja keeletestides, "seletavas veebilehitsejas", kus sõnale klõpsates näeb tema kohta käivat sõnastikuartiklit. Mõeldavad on ka SMS-teenused, nt sõnaartikli või käänamisjuhise näitamine, päeva sõna (ingl *word of the day*) pakkumine mobiilile jne jne.

2.2. Andmete esitamine

Igal XML-dokumendil on oma loogiline ja füüsiline struktuur. Dokumendil saab olla üks juurelement ehk esimene element, milles kõik ülejäänud elemendid sisalduvad. XML-dokumendi tekst koosneb andmetest (elemendi sisust) ja siltidest. Iga dokument sisaldab teatava hulga elemente, mis on piiritletud algus- ja lõpusildiga. Igal elemendil on (ainukordne) nimi, atribuudi abil võib täpsemalt määratleda elemendi tüübi (atribuudil on nimi ja väärtus). Näiteid:

```
(1)
<marksona>hommik</marksona>
<marksõna msliik="yhendverb">alla vanduma</marksona>
<marksona homnr="1">aas</marksona>
```

XML-vormingus andmed peavad rahuldama mitmeid eri tingimusi, näiteks:

1) XML-elemendid ei tohi üksteisega kattuda – lubamatu on olukord: <marksona>hommik<tyybinr></marksona>8</tyybinr>
2) atribuutide väärtused peavad olema jutumärkide vahel;
3) andmete sees ei saa kasutada mõningaid märke (< > &), nende asemel tuleb kasutada muutujaid (< > &) jne.

Toome näite EVS sõnaartiklist *rukkilill*, kus seletus <def> paikneb sisuploki <sisu> tähendusgrupi <txhgrp> alltähenduse grupi <dvgrp> definitsioonirühma <defgrp> sees. Sildid "sisu", "txhgrp" jne on elementide nimed, "txhnr" on atribuudi nimi. (Näidete tekstiosad on parema loetavuse huvides antud poolpaksus kirjas.)

```
(2)
<sisu>
  <txhgrp txhnr="1">
```

```
      <dvgrp>
        <defgrp>
          <def>umbrohuna kasvav korvõieline rohttaim</def>
        </defgrp>
      </dvgrp>
    </txhgrp>
  </sisu>
```

Tihti tekib vajadus kasutada kusagil mujal defineeritud elementi või atribuuti. Näiteks tuleb artiklisse lisada taime ladinakeelne nimetus, mis võetakse nt mingist andmebaasist. Siis tuleb eristada allikaid, kus elemendid/atribuudid on defineeritud. Selleks lisatakse elemendi nimele prefiks (ingl *prefix*) ning seostatakse omavahel prefiks ja allikas. Sisu plokk võiks nüüd välja näha selline:

```
(3)
<x:sisu xmlns:x="http://www.eki.ee/dict/evs">
  <x:txhgrp x:txhnr="1">
    <x:dvgrp>
      <x:defgrp>
        <x:def>umbrohuna kasvav korvõieline rohttaim</x:def>
        <x:def xml:lang="la">Centaurea ceanus</x:def>
      </x:defgrp>
    </x:dvgrp>
  </x:txhgrp>
</x:sisu>
```

Prefiksid "x" ja "xml" eristavad erinevates allikates defineeritud elemente. "Allikate" identifikaatoriteks on siinkohal "http://www.eki.ee/dict/evs" (prefiksi "x" jaoks) ning "http://www.w3.org/XML/1998/namespace" (prefiksi "xml" jaoks). Kõik sama identifikaatoriga defineeritud elemendid moodustavad nn nimeruumi (ingl *namespace*).

2.3. Skeem

Skeemi (*schema*) kasutades on võimalik andmeid XML-failis allutada kindlale struktuurile. Skeem määrab dokumendi füüsilise ja loogilise struktuuri:

1) millised elemendid milliste elementide sees sisalduda võivad, st milline on elementide ja nende tütarelementide vahekord;
2) mis tüüpi peavad olema mingid elemendid või atribuudid;
3) mis sisu võib olla ühel või teisel elemendil jne.

Skeemi järgi saame märgistatud teksti valideerida (ingl *validate*), st kontrollida XML-faili vastavust skeemi reeglitele.

EVS-i *rukkilille*-artikli andmetega XML-faili skeemi fragment võiks välja näha järgmine:

(4)

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://www.eki.ee/dict/evs"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:x="http://www.eki.ee/dict/evs">
<xs:import namespace="http://www.w3.org/XML/1998/namespace"
schemaLocation="xml.xsd" />
...
<xs:attribute name="txhnr" type="xs:positiveInteger"/>
<xs:element name="def" type="xs:string"/>
...
<xs:element name="txhgrp">
  <xs:complexType>
  <xs:sequence>
    <xs:element ref="x:dvgrp" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute ref="x:txhnr" use="required"/>
  </xs:complexType>
</xs:element>
...
</xs:schema>
```

2.4. Sõnastikud XML-vormingus

Toome näiteks ühe sõnaartikli ÕS 1999-st:

(5)

aabe <31: .aape, aabet> *kirjatäht*. Algus‿ , suur‿ , väike‿ aabe

ja sama artikkel XML-vormingus:

(6)

```
<q:art q:nr="5">
 <q:hdr>
  <q:m_grp>
   <q:m>aabe</q:m>
   <q:grm>
    <q:tyyp_grp>
     <q:tyyp>31</q:tyyp>
     <q:vorm>.aape, aabet</q:vorm>
    </q:tyyp_grp>
   </q:grm>
  </q:m_grp>
 </q:hdr>
 <q:tahendus>
```

```
<q:tah_grp>
 <q:t_grp>
  <q:t>kirjatäht</q:t>
 </q:t_grp>
</q:tah_grp>
</q:tahendus>
<q:naited>
 <q:nl_grp>
  <q:kvm>.</q:kvm>
  <q:nl>
   <q:n>Algus/, suur/, väike/aabe</q:n>
  </q:nl>
 </q:nl_grp>
</q:naited>
</q:art>
```

Sõnastiku XML-märgistus võib tekkida põhiliselt kolmel moel:
1) "käsitsi" suvalises tekstiredaktoris;
2) poolautomaatse teisenduse teel mõnest teisest märgistusest;
3) täisautomaatselt sõnastike haldussüsteemi abil.

Kõik uued sõnastikud EKI-s saavad edaspidi oma märgistuse sõnastike haldussüsteemi vahendusel, varem sisestatud sõnastikud oma spetsiifiliste struktuuride ja märgistusviisidega tuleb aga enne haldussüsteemi viimist üle viia standardsele XML-kujule. See nõuab eraldi tööd iga sõnastikuga, kohati isegi sõnaartiklite ümbertegemist, et tulemuseks oleks võimalikult lihtne skeem ja sellele vastav XML-struktuur (mida keerulisem struktuur, seda raskem on süsteemiga töötada). XML-failide erinevad keerukusastmed sõltuvad eelkõige sõnastiku tüübist (nt taskusõnastiku struktuur on oluliselt lihtsam kui suurel kakskeelsel tõlkesõnastikul), aga osalt ka struktuurikirjelduse enda loogikast. Lähtemärgistuse automaatse teisendamise võimalused sõltuvad samuti sõnastiku ülesehitusest, aga veel suuremal määral lähtemärgistuse liigist ja korrektsusest. Üleminek korrektselt deskriptiivselt märgistuselt XML-ile on suhteliselt hästi automatiseeritav, palju tülikam on hakkama saada polügraafilise märgistusega (sellest lähemalt Viks 1985).

Standardiseerimise vaev tasub ennast ära sõnastike puhul, mida on kavas ka edaspidi kasutada: kas sama sõnastiku uuendamiseks või uute sõnastike lähtematerjalina või ka muudes keeletehnoloogia rakendustes. Seni on korrektse XML-märgistuse saanud EKI-s kolm sõnastikku: EVS-i 3 esimest köidet ja ÕS 1999 poolautomaatse teisenduse teel ning eesti–X-sõnastik, mille eeltööd tehti "käsitsi"

Wordi makrodega. Töö vanade elektrooniliste sõnastike teisendamisega jätkub vastavalt vajadusele ja võimalustele.

3. Sõnastike haldussüsteem

Sõnastike haldussüsteem võimaldab sõnastikumaterjali "pidamist" XML-vormingus. See hõlmab leksikograafi töö põhisisu: sõnaartiklite koostamist ja toimetamist, kõikvõimalikke otsinguid kogu sõnastiku ulatuses, sõnaartiklite järjestamist ja väljastamist. Haldussüsteem teeb valmis ka küljenduse ning hoiab silma peal kõigel, mis toimub: kontrollib artikli struktuuri õigsust ja registreerib tehtud muudatused.

3.1. Sõnastike haldamise põhifunktsioonid

Haldussüsteemi põhifunktsioonid on:

1) uue artikli sisestamine (koostamine);
2) artikli sisu valideerimine;
3) olemasolevate artiklite kustutamine ja muutmine, elementide/atribuutide lisamine/kustutamine artiklis (toimetamine);
4) artiklite otsimine eri tunnuste järgi: elementide/atribuutide olemasolu ja väärtuste järgi; lisavõimalustena metamärkide kasutamine, tõstutundlikkuse arvestamine, erisümbolite (transkriptsioonimärgid, kirjavahemärgid jms) arvestamine/mittearvestamine;
5) artiklite vaatamine ja väljastamine: üksikartikli vaade töö käigus, otsingu tulemuste väljastamine (faili või väljatrükki), faili küljendus;
6) artiklite järjestamine;
7) töökäigu registreerimine: koostamise/toimetamise aeg ja isik.

3.2. Sõnastike haldamise põhinõuded

Haldussüsteemi ülesehitamisel on silmas peetud järgmisi põhinõudeid.

1. Sõnastikku peab saama kollektiivselt koostada ja toimetada: mitu tegijat korraga, kusjuures eri taseme kasutajatel on erinevad õigused.
2. Töökoha tarkvara peab põhinema üldkasutatavale standardile ning komponendid peavad olema lihtsalt installeeritavad.

3. Artiklit peab saama lihtsalt otsida, võimalikult paljude tunnuste järgi.
4. Artikli toimetamise protsessis ei pea toimetaja olema teadlik skeemi tehnilistest nõudmistest; süsteem võimaldab ainult neid parandusi, mis on antud kontekstis lubatud.
5. Toimetamise käigus peab olema võimalik kontrollida artikli vastavust skeemile, sellele mittevastavat artiklit ei lasta salvestada.
6. Toimetaja parandused on sisulised; artikli vaade, väljatrükk ja küljendus genereeritakse automaatselt.
7. XML-element ja talle vastav vaate tekstifragment on seotud: klõpsamine ühel toob nähtavale ka teise.
8. Andmed artikliga töötamise käigu kohta (isikud, aeg) fikseeritakse automaatselt.

3.3. Sõnastike ettevalmistamine

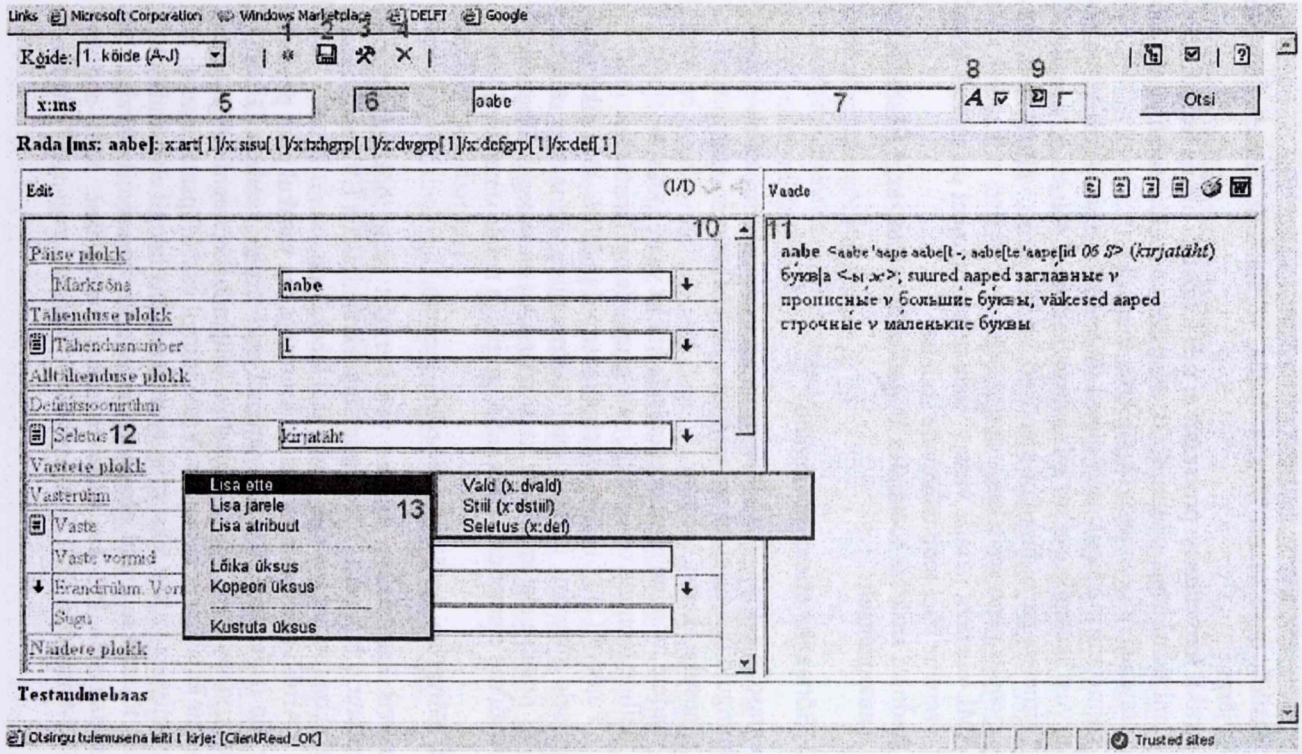
Olemasoleva sõnastiku sisseviimine haldussüsteemi vajab järgmisi toiminguid (eeldusel, et andmed juba on XML-vormingus):
1) sõnastiku skeemi loomine;
2) vaadete loomine nii toimetamisala kui ka vaate jaoks;
3) mallide loomine uue artikli, uute gruppide jne lisamiseks.

Punktid (1) ja (2) on automatiseeritud ja XML-sõnastiku sisseviimisele süsteemi järgneb vaid häälestamine. Päris uue sõnastiku puhul luuakse skeem käsitsi koos sõnastiku autoritega: fikseeritakse võimalikud elemendid, atribuudid, grupid ja nende omadused.

3.4. Sõnastiku haldamise tehniline teostus

Serveris kasutatakse Apache veebiserverit ning töökohal Internet Explorer 6 brauserit. Näide toimetaja ekraanipildist EVS artikli *aabe* korral on toodud joonisel (järgmisel leheküljel).

Artikli parandamise ajal on toimetaja ees vasakul pool parandatav ala (toimetamisala 10), paremal pool artikli vaade (11). Artikli vaade vastab igal hetkel sõnastiku küljendusversioonile: iga tehtud muutus kajastub kohe vaates. Vaate mingil osal klõpsates avaneb toimetamisalas vastav tekst ning toimetamisalas mingil elemendil klõpsates tõstetakse antud element vaates esile teise taustavärviga. Nii toimetamisala kui ka vaate loomisel kasutatakse XML-andmete XSLT-teisendusi (XSLT: *Extensible Stylesheet Language Transfor-*


Links
Microsoft Corporation
Windows Marketplace
DELFT
Google
Kõide: 1. kõide (A-J)
1 2 3 4
x:ms 5
6
aabe 7
8 9
Otsi
Rada [ms: aabe]: x:art[1]/x:sisu[1]/x:tzhgrp[1]/x:dvgrp[1]/x:defgrp[1]/x:def[1]
Edit
(1/1)
Vaade
10
11
Päise plokk
Märksõna
aabe
Tähenduse plokk
Tähendusnumber
1
Alltähenduse plokk
Definitsioonirühm
Seletus 12
kirjatäht
Vastete plokk
Vasterühm
Vaste
Vaste vormid
Sugu
Näidete plokk
aabe <aabe'aape aabe[t-, aabe[te'aape[id 06 S> (kirjatäht)
буква <-ы ж>; suured aaped заглавные ν
прописные ν большие буквы, väikesed aaped
строчные ν маленькие буквы
Lisa ette
Lisa järele
Lisa atribuut
13
Lõika üksus
Kopeeri üksus
Kustuta üksus
Vald (x:dvald)
Stiil (x:dstil)
Seletus (x:def)
Testandmebaas
Otsingu tulemusena leiti 1 kirje: [ClientRead_OK]
Trusted sites

mations), mis esitavad artikli sisu HTML kujul. XSLT-teisenduste abil määratakse, milline element esitatakse poolpaksult, milliste elementide ümber pannakse sulud jne.

Mõlemast alast ülevalpool paikneb menüüde, nuppude ning info ala. Kõige ülemises reas paiknevad artikli lisamise (1), salvestamise (2), ümbernimetamise (3), kustutamise (4) nupud. Teise rea menüüst (5) on võimalik valida otsitava elemendi nimi. Lahtrisse (6) on võimalik sisestada otsitava atribuudi väärtus, lahtrisse (7) otsitava elemendi tekst. Linnuke (8) määrab otsingu tõstutundlikkuse ja erisümbolite (9) arvestamise.

Mingi elemendi nimel (12) hiire parema nupuga klõpsates avaneb toimetajale kontekstmenüü (13), mille abil saab elemente ja atribuute lisada, kustutada, kopeerida või kleepida. Vastavalt kontekstile saab valida, milliseid elemente võib antud elemendi ette, taha või sisse lisada. Vaate (11) kohal on väljatrüki ning küljenduse nupud. Küljendamiseks valitud artiklid esitatakse MS Wordis.

4. Lõpetuseks

Töö sõnastike haldussüsteemiga on jõudnud rakendusjärku. EVS-i 4. köite koostamine ja toimetamine on lõppemas; käsil on Eesti–X-keele sõnastiku koostamine; ÕS 1999 on süsteemi sisse viidud ja käimas on uue väljaande (2006) jaoks vajalike muudatuste tegemine.

Sõnastike haldussüsteemist on olemas nii lokaalne kui ka veebiversioon ning ta sobib põhimõtteliselt igasuguse struktureeritud info haldamiseks, mitte ainult sõnastike jaoks. Lähema info saamiseks võib ühendust võtta e-posti aadressil andres.loopmann@eki.ee.

Kirjandus

Calzolari, Nicoletta, Grishman, Ralph, Palmer, Martha (responsible authors) 2001. Survey of Major Approaches Towards Bilingual/ Multilingual Lexicons. ISLE Computational Lexicons Working Group. Deliverable D2.1-D3.1. February 2001.

Dalzell, Tom, Victor, Terry, Williams, John 2002. 'A labour so ungrateful': Report of a project to update Eric Partridge's *Dictionary of slang and unconventional English*. – Proceedings of the Tenth Euralex International Congress, Euralex 2002. Vol I–II. Ed by A. Braasch, C. Povlsen. Center for Sprogteknologi, CST, 331–340.

EVS = Eesti–vene sõnaraamat 1–3 (5). 1997–2003–. Tallinn: Eesti Keele Sihtasutus.

Langemets, Margit 2000. Leksikaalse info kodeerimine. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim T. Hennoste. Tartu: Tartu Ülikooli kirjastus, 101–126.

Langemets, Margit 2002. Eesti Keele Instituudi elektrooniline keelevara. – A&A 5, 39–46.

MSDN Library – October 2002 (MSDN: Microsoft Developer Network).

Prószéky, Gabor, Kis, Balázs 2002. Development of a Context-Sensitive Electronic Dictionary. – Proceedings of the Tenth Euralex International Congress, Euralex 2002. Vol I–II. Ed by A. Braasch, C. Povlsen. Center for Sprogteknologi, CST, 281–290.

Smrz, Pavel 2002. Lexical Databases in XML: A Case Study of Up-Translation of the Dictionary of Literary Czech Language. – Proceedings of the Tenth Euralex International Congress, Euralex 2002. Vol I–II. Ed by A. Braasch, C. Povlsen. Center for Sprogteknologi, CST, 729–735.

VES = Vene–eesti sõnaraamat I–IV 1984–1994. Tallinn: Valgus.

Viks, Ülle 1985. Arvuti ja kakskeelsed sõnastikud. – Keel ja Kirjandus 6, 356–361.

Viks, Ülle 1990. Sõnastike andmebaas: milleks, mis ja kuidas. – Arvutuslingvistika sektori aastaraamat 1988. Toim J. Ross. Tallinn: Keele ja Kirjanduse Instituut, 167–175.

VMS = Väike murdesõnastik I–II 1982–1989. Toim. V. Pall. Tallinn: Valgus.

XML 1.0 = Extensible Markup Language (XML) 1.0. W3C Recommendation 10-February-1998. Referaat. Tõlkinud Uuno Vallner. Eesti Informaatikakeskus, 2000. Internetis: http://www.riik.ee/xml/trans/REC-xml-19980210-ee.html#sec-intro.

ÕS 1976 = Õigekeelsussõnaraamat. Toim R. Kull, E. Raiet. Tallinn: Valgus.

ÕS 1999 = Eesti keele sõnaraamat ÕS 1999. Toim T. Erelt. Tallinn: Eesti Keele Sihtasutus.

Eesti keel internetis[1]

Anni Oja
Tartu Ülikool

1. Sissejuhatus

Artikli eesmärk on esitada ülevaade suhtlusest Eesti internetiruumis, tutvustades erinevaid suhtluskeskkondi ning spetsiifilist keelekasutust mõjutavaid tegureid. Arvestades tehnoloogia kiiret arengut, on see tehnilise poole pealt ennekõike 2005. a olukorra ülevaade, kuid keele kohta käivad üldistused peaks paika pidama ka üldisemalt.

Andmed põhinevad osalt autori uurimustel, osalt uurimusteks veel vormistamata kogemustel, osalt (internetis) peetud vestlustel ja mingil määral loetud töödel. Aspektide valik ja suhtluskeskkondade jaotus on pikaajalise kogemuse tulemus. Seni leitud materjalidest on vähe abi olnud, sest tihtipeale on uurijad ilmselt kas midagi valesti tõlgendanud või kellegi teise valetõlgendust usaldanud. Nii pannakse mõnedes töödes foorumid ühte patta IRC-ga, kuid tegelikkuses on need kaks täiesti ise asja: IRC on üks jututoa vorme ja seega kõnekeelse sünkroonse suhtluse esindaja, foorumis aga suheldakse pigem postiloendi või uudisgrupi stiilis. Kaalusin võimalust koos jaotusega esile tuua ka teiste autorite jaotused ja nende nõrgad küljed, kuid see muudaks üldpilti vist veelgi segasemaks.

2. Internetisuhtluse olulisemad aspektid

Arvuti vahendusel toimuvat suhtlust saab liigitada mitmeti (vt Crystal 2001). Keeleuurija seisukohast on olulisemad omadused aeg ja vestluspartnerite arv ning suhe. Ajafaktorit arvestades jaguneb internetisuhtlus kaheks: **sünkroonne** ja **asünkroonne** infovahetus. Sünkroonne suhtlus tekib juhul, kui vestluspartnerid viibivad samaaegselt samas keskkonnas ja näevad üksteise teksti koheselt nagu tavapärases suulises vestluses (vrd telefonikõne, kohvikuvestlus). Sünkroonset suhtlust esindavad internetis näiteks jututoad, rollimän-

[1] Tööd on toetanud HTM (riiklik programm "Eesti keel ja rahvuslik mälu").

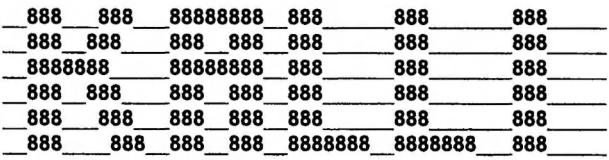
gud ja paarivestlusprogrammid. Asünkroonse suhtluse puhul ei pea kõik suhtlejad samal ajal internetis viibima, üks osapool teeb oma info teistele kättesaadavaks, teadmata, millal seda vaadatakse (vrd tavakiri, artikkel, mobiilisõnum). Konkreetsetest rakendustest võiks nimetada e-posti, foorumeid, kommentaariume. Sünkroonne ja asünkroonne suhtlus on keeleliselt hästi eristuvad: esimese puhul on oluline info edastamise kiirus, mistõttu pannakse rõhku lühidusele ja optimaalsusele. Teisel puhul on teksti loojal rohkem aega oma mõtet sõnastada ning ülesehitus on läbimõeldum. Üldiselt võivad sünkroonsus ja asünkroonsus ka ühe vestluse raames vahelduda: näiteks e-posti puhul võib kirja saaja samal ajal arvuti taga olla ning kohe vastata.

Optimaalsus väljendub internetikeeles mitmel moel: lühendatakse levinumaid sõnu ja käibefraase, suurtähti ning kirjavahemärke (osa)lausepiiri märkimiseks ei pruugita, välja jäetakse kõik, mida välja jätta saab. Kui kirjutades tekib näpuvigu, ei hakata neid enamasti parandama. Leidub ka internetikasutajaid, kes jälgivad rangelt kirjakeele tavasid, kuid valdavalt annab tooni ikka nooremate internetikodanike vaba keelekasutus.

Kui tavalises suulises suhtluses kasutavad inimesed mitmesuguseid mitteverbaalseid vahendeid, et öeldut rõhutada või pehmendada, siis internetis on selle jaoks **emotikonid** (ingl k *emoticon, smiley*). Emotikonid moodustatakse kirjamärkidest ning nende funktsiooniks on edasi anda mingit emotsiooni või tundevarjundit. Kõige tuntumad on naeratus :) ning kurbus :(, kuid loomingulised internetikasutajad moodustavad ka keerukamaid pilte, et vestlust värvikamaks muuta:

(_8-(|)	Homer Simpson	//
%-)	vaatasin liiga palju televiisorit	
***<:-)>>**	päkapikk	

Peale emotikonide kasutatakse mõnel pool ka suuremaid sümbolpilte (ingl k *ascii graphics*):



Emotiivsuse osakaalu mõjutab suhtluses osalejate arv ning omavaheline suhe. Internetisuhtluses on esindatud nii **monoloog**, **dialoog** kui **multiloog**. Monoloogi puhul koostab teksti üks inimene, tagasiside saamine on pigem ebatüüpiline ja enamasti on publik anonüümne: siia alla käivad veebilehed ja ajaveebid. Dialoogilises vestluses on osapooli kaks ja suhtlusvõimalusi laialdaselt, alates e-postist ja lõpetades paarisuhtlusprogrammidega. Multiloogis osaleb rohkem inimesi, tüüpilised rakendused on näiteks jututoad, foorumid ja kommentaariumid. Mono- ja multiloogilise suhtluse vahele jääb hall ala, kus üks vorm võib üle minna teiseks: nii on paljudes ajaveebides olemas kommenteerimise võimalus ja algselt monoloogilise ajaveebi sissekande alla võib tekkida kommentaaridest multiloog. Ajaveebide puhul leidub ka keerukamat multiloogi, kus ühes ajaveebis kommenteeritakse mõnes teises ajaveebis leiduvat teksti.

Sellise suhtluse analüüsiks tuleb tutvust teha sotsioloogiast pärineva **sotsiaalse võrgustiku** mõistega. Sotsiaalne võrgustik kirjeldab inimestevaheliste suhete struktuuri, mis internetis on üsna lihtsasti jälgitav: ajaveebide (vt Asükroonsed suhtluskeskkonnad) puhul viitavad autorid teistele kirjutajatele, kelle tekste nad loevad jne. Sotsiaalne võrgustik tuleb internetis eriti teravalt esile mingi konflikti ajal, olgu siis tegu laiema ühiskondliku probleemi või kitsa ringi sisese vastasseisuga. Keeleuurijale on siin laialdased võimalused diskursuse analüüsimiseks.

Tuntavalt mõjutab suhtlust **anonüümsuse määr**. Eriti selgesti on seda näha internetikommentaarides: kui kommenteerimiskeskkond võimaldab anonüümseks jäämist, kipuvad arvamused olema teravamad ja solvavamad kui identifitseerimist nõudvates kohtades. Samas võivad ka anonüümses keskkonnas tekkida eristatavad kasutajad, kes suudavad oma **identiteeti** piisavalt esile tuua. Inimese "nägu" internetis kujunebki tihtipeale vaid selle põhjal, mida ja kuidas ta räägib. Tänapäeval on tavaline, et aktiivse internetikasutaja sõpruskonda kuulub hulga inimesi, keda ta pole iialgi näinud, ainuke kujutelm põhineb vaid virtuaalse sõbra jutust saadud teabel. Koge-

mustega võrgusuhtleja suudab oma vestluspartneri keelekasutuse iseärasuste põhjal teha mõningaid järeldusi, mida saab kontrollida ka tekstikorpuste analüüsimisel. Autori töö internetikorpustega näitab, et enim saab keelekasutuse varieerumise alusel eristada sugu ja vanuserühmi.

Uurija seisukohast on väga oluline eristada **avalikku** ja **privaatset suhtlust**. Võimaluse korral tuleks uuritavaid uurimistööst teavitada ja hoolega jälgida, et esitatavasse materjali ei satuks infot, mis võib kedagi kahjustada.[2] Eriti ettevaatlik tuleb olla privaatse suhtluse ja info töötlemisel. 1999. aastal Washingtonis korraldatud internetiuurimuste eetika töötoas leiti, et internetis suhtlejate virtuaalsed identiteedid väärivad samasugust kaitset kui füüsilised isikud (Frankel jt 1999). See kehtib ka kasutajanimede kohta: tekstinäidetes oleks hea kasutajanimed asendada või varjata, sest asjasse pühendatud isikute jaoks on kasutajanimi sama kõnekas kui tavaelus täisnimi.

3. Sünkroonsed suhtluskeskkonnad

Jututoad kujutavad endast keskkonda, kus on paljud inimesed saavad korraga reaalajas suhelda, nii avalikult kui privaatselt. Tehniliselt saab jututubasid üsna laialt liigitada (tekstipõhised telneti- ja IRC vestlusruumid, graafilise kasutajaliidesega JAVA jutukad, interaktiivsed kolmemõõtmelised virtuaalmaailmad jne). Keelelisest seisukohast on need siiski kõik ühesugused, seega ei näe autor põhjust keeleuurimustes tehnilist jaotust sisse tuua. Jututubade levinumate funktsioonide hulka kuuluvad avalik ja privaatne reaalajas vestlus ning mitteverbaalsete tegevuste kirjeldamine (nt *Tom noogutab innukalt, JS võtab nüüd Potteri ja poeb põhku*). Sõltuvalt konkreetsest tarkvarast on veel võimalik endale kirjeldusi luua, kasutajanime vahetada ja palju muudki. Tavaliselt on jututubades administraatorid,

[2] Alati ei ole võimalik uuritavatega kontakteeruda, näiteks veebilehtedelt materjali kogudes on vahel pea võimatu lehe autorit või tema kontakte leida – või osaleb suhtluses sadu tuhandeid inimesi, kellelt ükshaaval loa küsimine suuremaks statistiliseks uurimuseks oleks võib-olla suurem töö kui uurimus ise. Sellisel juhul tasuks ringi vaadata, võib-olla on keskkonna jaoks koostatud kasutustingimused, mis privaatsust piiritlevad ja materjali kasutamist reglementeerivad.

kes vestlusel silma peal hoiavad ja korrarikkujaid eemaldavad. Jututoa keelekorpuse kogumiseks saab kasutada programme, mille üldnimetajaks on *bot* (tuletatud sõnast *robot*). Neid programme saab panna ööpäevaringselt jututoas viibima ja kogu avalikku vestlust talletama. Enne taolise programmi paigaldamist tuleb luba küsida jututoa administraatorilt ning kaasvestlejatelt, kuna vastasel juhul võidakse *bot* lihtsalt jututoast välja heita.[3] Eestis on jututubade uurimisega tegelenud näiteks Sigrid Salla (Salla 2001) ja Pille Vengerfeldt-Pruulmann (Vengerfeldt 2001).

Rollimängud ühendavad endas jututubade suhtlusvõimalusi ja fantaasiamaailma seiklusi. Populaarsemad internetipõhised rollimängud on innustust saanud J. R. R. Tolkieni raamatutest, mängijatel on võimalik müütilises maailmas ringi seigelda, ülesandeid täita ja oma tegelaskuju arendada. Tihti kandub mängulisus üle ka mängijatevahelistesse suhetesse, püütakse valida oma tegelaskujule sobilik kõnepruuk ja suhtlusmall. Selline virtuaalne ühiskond on olnud eriti tänuväärseks materjaliks sotsioloogidele, keda huvitas kogukonnasiseste reeglite kujunemine ja rakendamine (Wallace 2002). Virtuaalsed maailmad on põnevad ka kognitiivteadustega tegelejatele, eriti tekstipõhised mängud, kus ümbrust ja tegevusi kirjeldatakse vaid keelevahenditega. Ühest küljest on huvitav see, kuidas mängu loojad ja kujundajad keskkonda teksti abiga visualiseerivad, teisalt aga see, kuidas mängijad maailma kujutlevad. Eestis on seda uurinud Kalver Keskküla (Keskküla 2002), kirjeldades virtuaalmaailmas ruumi kujutlemist.

Paarisuhtlusprogrammid on kujundatud põhiliselt privaatse dialoogilise vestluse tarbeks. Tüüpiliselt peab kasutaja programmi oma arvutisse paigaldama ning suhtluspartnerid tuleb lisada vastavasse nimekirja, nii et kasutaja saab teha eelvaliku, kellega kõnelda. Lisaks tavalisele tekstipõhisele jutule on võimalik vahetada faile, kasutada veebikaamerat, kutsuda vestlusse rohkemaid osalejaid.

Põgusalt võiks mainida ka **internetitelefoni** ja **videokonverentse**. Internetitelefon täidab tavatelefoni funktsioone, kasutades lihtsalt tavapärase telefoniaparaadi ja -kaabli asemel vastavat tark-

[3] Illegaalse boti kasutamine ei jõuagi ebaeetiliseks saada, sest tavaliselt avastavad jututoa administraatorid salakõrva nii ruttu, et see ei suuda veel suurt midagi talletada.

vara, internetti ja kõrvaklappe/mikrofoni. Veebikaamera olemasolu korral saab internetitelefoniga pidada ka videokõnesid.

Videokonverentsid on näide interaktiivsest multiloogist, tavaliselt edastatakse interneti teel videopilti ja heli. Videokonverentsi eesmärgiks on ühendada erinevates kohtades viibivaid osalejaid mingi ettevõtmise tarbeks üheks auditooriumiks.[4]

4. Asünkroonsed suhtluskeskkonnad

Portaalid on suuremad veebikeskkonnad, mis pakuvad oma kasutajatele tavaliselt korraga mitmeid erinevaid suhtlusvõimalusi. Osa informatsioonist ja suhtlusest võib olla privaatne, osa avalik. Ka funktsioone on portaalidel tavaliselt üsna palju, alates uudiste jagamisest ja lõpetades virtuaalse ostukeskkonnaga. Portaali kasutajaskonda võib ühendada nii suhtlemissoov kui ühine huvi (nt arvutimängud). Suhtlemisviisideks on muuhulgas erasõnumite saatmine, virtuaalsete kingituste tegemine, portaalisisesed ajaveebid, jututuba ja portaalisisesed kasutajat tutvustavad lehed.

Foorumid pakuvad internetikasutajatele võimalust mingitel kindlatel teemadel arutleda. Foorumid asetsevad veebis ja tavapäraselt on neil kindel üldteema, mis jaotub alateemadeks. Kasutajaskonda ühendab huvi just selle valdkonna vastu. On nii piiratud ligipääsuga foorumeid kui päris avalikke, vestlusel hoiavad silma peal moderaatorid ja vähemalt postituse lisamiseks tuleb end enamasti registreerida. Tekst on üldjuhul pisut kirjakeelsem, suhtlus pigem asjalik. Samas tekib ka foorumites tihti sõnasõdasid ja erinevate tõekspidamiste konflikte. Aktiivse kasutajaskonnaga foorumil on tavaliselt ka oma autoriteedid, keda hinnatakse vastava valdkonna eksperdina ja kellele ühiselu kirjutamata reeglites on lubatud rohkem kui teistele.

Postiloendid ja **uudisgrupid** on oma olemuselt foorumitega sarnased, kuid suhtlus toimub veebi asemel pigem e-posti teel. Postiloend koondab teemast huvitatud inimeste e-posti aadresse, kuhu

[4] Interaktiivse kübervestluse keele erilisuse või mitte-erilisuse kohta on hetkel raske midagi öelda, sest vastavaid keeleuurimusi ei ole autorile teadaolevalt veel tehtud. Ilmselt seal on erijooni, sest internetitelefon pakub ohtralt erinevaid lisavõimalusi ning videokonverentsidegi puhul erineb suhtlus pisut tavapärasest konverentsist. Tõin need siin ära üldpildi täielikkuse huvides.

edastatakse kõik postiloendisse saadetud kirjad. Uudisgrupid on pärit varasematest internetiaegadest, mil võrgus käidi telefonimodemi abil ja korraga pigem lühikest aega – siis oli vaja head asünkroonset keskkonda, kuhu sõnumeid jätta. Nüüd on arvutivõrgud kordades kiiremaks muutunud, ent uudisgrupid ei ole oma populaarsust kaotanud. Foorumiga sarnaselt on postiloenditel ja uudisgruppidel oma administraatorid, anonüümsust aga pisut vähem, sest inimesed kirjutavad oma e-posti aadressidelt.

Kommentaariumid toimivad tavaliselt osana mingist teemaveebist, lugejatel on võimalik teksti juurde lisada oma arvamus. Enamasti on tegu anonüümse suhtlusega, mistõttu inimesed avaldavad oma mõtteid tihti vabamal kujul kui nad seda identifitseeritult teeksid. Sõnasõjad on kommentaaride puhul tavalised ning lahknevate vaadete kaitsjad panevad suurt rõhku oma arvamuse tõestamisele. Anonüümsusest hoolimata tekivad ka kommenteerijate seas autoriteedid, kelle "käekirja" on võimalik tuvastada ning kelle väljaütlemised rohkelt toetajaid leiavad.

Vikipeediad (ingl k *wikipedia*) on kollektiivselt koostatavad vabad võrguentsüklopeediad. Informatsiooni lisada ja parandada saab igaüks ning nii koonduvad entsüklopeediasse suure kogukonna teadmised, mida pealegi operatiivselt muudetakse ja täiendatakse. Vikipeedia võib olla nii universaalne kui kitsast valdkonda hõlmav, keelekasutus on asjalik ja informatiivne.

E-post on elektrooniline asendus tavaelulisele kirjasaatmisele, olles kasutatavuse poolest oma eeskujust ammu mööda läinud. E-kirjad esindavad asünkroonse suhtluse dialoogilist vormi, levinud on ka masspostitus, ent see on enamasti pigem ühepoolne info levitamine.

Ajaveebid (ingl k *blog, weblog*) kujutavad endast internetis enam või vähem avalikult peetavaid päevikuid. Ajaveebi vormiks on veebilehekülg, sisu valmib ühe või ka mitme autori tööna, tihti on lugejatel võimalik teksti kommenteerida. Ajaveeb võib olla nii isiklikku elu ja mõtete kajastajaks kui informatsiooni edastamise kohaks. Tavapärase teksti asemel võib ajaveeb koosneda ka fotodest või videoklippidest.

Isiklikud veebilehed on kõigist seni kirjeldatud suhtlusvahenditest kõige monoloogilisemad ja ka kõige staatilisema sisuga, st sisu muudetakse pigem harva. Isiklikel veebilehtedel tutvustatakse en-

nast, oma mõtteid ja huvisid. Tekst on läbimõeldum ja tagasisidet sellele pigem ei oodata.

5. Kokkuvõte

Eelnevast on näha, et eestikeelse internetisuhtluse maailm on üsna lai ja kirev, koondades üpris erinevaid tekstiliike ja inimesi. Keeleuurija jaoks on see kõik lihtsasti kättesaadav, vaja on vaid pisut süvenemist ja harjumist, et õppida virtuaalses kogukonnas orienteeruma. Siinne ülevaade püüab olla väikeseks reisijuhiks sel põneval teekonnal:)

Viiteid ja märksõnu
Eesti virtuaalsete suhtluskeskkondadega tutvumiseks

Jututoad
Cafe jututuba: telnet cafe.ee 5555, http://www.cafe.ee
Puhhi jututuba: telnet vw1.chem.ut.ee 6666
OK jututuba: www.ok.ee

Rollimängud
Stonia (tekstipõhine): telnet stonia.ttu.ee 4000, http://stonia.ttu.ee
Runescape (rahvusvaheline, graafiline): http://www.runescape.com

Paarisuhtlusprogrammid
MSN, Jabber, ICQ, Skype (internetitelefon)

Suhtlusportaalid
Rate: http://www.rate.ee
Hei: http://www.hei.ee
Flirt: http://www.flirt.ee

Foorumid
Perekool: http://www.perekool.ee
Dragon: http://www.dragon.ee
Biker: http://www.biker.ee

Postiloendid, uudisgrupid
http://lists.ut.ee, http://news.ut.ee

Kommentaariumid
Delfi: http://www.delfi.ee

Vikipeediad
http://et.wikipedia.org/wiki/Esileht
http://wiki.linux.ee

Ajaveebid
http://www.siimteller.com
http://blog.moment.ee
http://blog.tr.ee

Kirjandus

Crystal, David 2001. Language and the Internet. Cambridge: Cambridge University Press.

Frankel, Mark S., Siang, Sanyin 1999. Ethical and legal aspects of human subjects research on the internet. A report of a workshop. Washington. http://www.aaas.org/spp/sfrl/projects/intres/report.pdf

Keskküla, Kalver 2002. Küberruum kui reaalsus: sotsiaalsus ja ruumitaju Aardwolfi 'muda' näitel. Bakalaureusetöö. Tartu Ülikool. http://www.ut.ee/~kalver/baka/

Salla, Sigrid 2001. Interneti meelelahutusliku jututoa kui võrgusuhtlusvormi erijooni Virtual City jututoa näitel. Magistritöö. Tallinna Ülikool.

Vengerfeldt, Pille 2001. Power, Praise and Punishment in Cybercommunities. Bakalaureusetöö. Tartu Ülikool. http://www.ut.ee/~pille_v/akadeemia/power.html

Wallace, Patricia 2002. Internetipsühholoogia. Tallinn: Valgus.

Piiratud inglise keel ACE ja sellega seotud tarkvara

Kaarel Kaljurand
Tartu Ülikool, Zürichi Ülikool

1. Sissejuhatus

Piiratud loomulik keel (inglise keeles kasutatakse mõisteid *controlled natural language, processable natural language, pseudo natural language*) on loomuliku keele alamhulk, st mõned loomuliku keele konstruktsioonid ning nende interpreteerimisvõimalused on piiratud loomulikust keelest eemaldatud. Piiratud keelel on formaliseeritav süntaks ja semantika, seega saab alati öelda (tõestada), kas mingi tekst kuulub keelde või mitte, ning juhul kui kuulub, siis saab rääkida selle teksti tähendusest (kirjapanduna mingis formalismis). Saab ka nõuda, et teksti tähendus oleks ühene.

Seega on piiratud loomulik keel ühest küljest formaalne keel nagu nt Java, UML või HTML, kuid teisest küljest on keele disainimisel pisut erinevad eesmärgid. Oluline on, et keele kasutamine oleks inimesele loomulik ega ei nõuaks teadmisi, mida formaalsed keeled üldjuhul eeldavad.

Käesolev artikkel käsitleb piiratud loomulike keelte mõistet, disaini ja rakendusi piiratud inglise keele ACE näitel. ACE (Attempto Controlled English) on Zürichi Ülikooli informaatika instituudi projektis Attempto (vt http://www.ifi.unizh.ch/attempto) arendatav keel. Pikemalt kirjeldame keele ACE parserit APE (Attempto Parsing Engine) ning erinevaid abivahendeid, mis ACE-keelsete tekstide loomist lihtsustavad.

2. ACE ja selle piirid

Keele ACE süntaks ja semantika põhinevad inglise keele omadele. Keele disainimisel on püütud jälgida, et kõik laused oleksid aktsepteeritavad ning mõistetavad nagu inglise keele laused ning et erinevused inglise keelest oleksid vähese vaevaga õpitavad.

2.1. Süntaks

Keele ACE lihtlaused sisaldavad üht verbi, mis võib olla nii intransitiivne, transitiivne kui ditransitiivne ning mida saab prepositsioonifraaside ja adverbidega täiendada, ning verbi komplemente: subjekti ja objekti, milleks on substantiivifraasid. Substantiivifraasi moodustab substantiiv, millele peab tingimata eelnema artikkel või kvantor ja millele võib valikuliselt eelneda adjektiivifraas ning valikuliselt järgneda apositsioon, *of*-konstruktsioon ning relatiivlause. Sama tüüpi fraase saab reeglina omavahel koordineerida. Järgnevas toome keele ACE põhikonstruktsioonid koos näidetega.

NP = Det (AdjP) Noun (Apposition) (of-PP) (RelCl)
(1) *Every big and red button "Go!" of the upper panel that is illuminated*

VP = TransitiveVerb NP (Adverbs) (PPs)
(2) *controls the aircraft on the runway*

Sentence = NP VP .
(3) *Every big and red button "Go!" of the upper panel that is illuminated controls the aircraft on the runway.*

Kõiki keele ACE lauseid saab omavahel koordineerida (*and, or*), eitada (*it is not the case that*) ning *if–then* konstruktsiooniga siduda:

(4) *If a dog barks then it is not the case that a man is happy.*

Lisaks deklaratiivlausetele (mis lõpevad punktiga), lubab ACE küsilauseid (mis lõpevad küsimärgiga):

(5) *What does the button control?*
(6) *Where does the button control the aircraft?*

Paljud inglise keeles lubatud konstruktsioonid puuduvad keelest ACE. Nt prepositsioonifraas peab tingimata järgnema verbile (ainus erand on *of*-konstruktsioon, mis võib järgneda ainult substantiivifraasile), verbid ei saa üksteist täiendada, sest keelatud on infinitiivtäiendid, jms:

(7) * *A man with a telescope waits.*
(8) * *A man intends to wait.*

Reeglina peab substantiivifraasi alustama artikkel (*a, the, some*) või kvantor (*every, no, ...*), erand on ainult substantiivifraasid, mis koosnevad pärisnimest.

2.2. Semantika

Keele ACE semantika näitab, millised laused on samatähenduslikud (need saavad täpselt ühesuguse semantilise esituse) ning millised mitte. Semantilisele esitusele tuginedes saab (automaatselt) leida tekstist võimalikke vastuolusid või liiasust. Erinevate võimalike semantiliste esituste arv näitab keele väljendusrikkust.

Loomulike keelte tekstide semantika puhul võib rääkida polüseemiast (ühele tekstile vastab mitu erinevat tähendust) ja sünonüümiast (mitmele erinevale tekstile vastab sama tähendus). Formaalsete keelte disainimisel püütakse kindlasti vältida polüseemiat. Isegi kui mingi süntaktiline konstruktsioon on mitmene, selgub selle tähendus lokaalse konteksti põhjal ning kogu teksti (nt programmikoodi) tähendus on üksainus. Tihti püütakse vältida ka sünonüümiat, sest erinev vorm võib kasutajale tekitada illusiooni erinevast tähendusest.

ACE on formaalne keel, mille tekstid on ühese tähendusega, kuid sünonüümsed konstruktsioonid on lubatud ning isegi teadlikult lisatud. Keele ACE disainimine peab arvestama olulise kitsendusega, millega teised formaalsed keeled eriti kokku ei puutu. Nimelt, keele ACE kasutaja on inglise keele oskaja, kel seetõttu on keele ACE toimimisest (lausete kuulumisest keelde ning nende tähendusest) teatav eelarvamus.

Inglise keelele omast mitmesust on keeles ACE välditud kolmel viisil. Esiteks, inglise keele mõttes väga mitmesed süntaktilised konstruktsioonid keelde ACE ei kuulu. Kasutaja peab omakorda selgeks õppima keele ACE süntaksireeglid, et selliste lausete kasutamist vältida.

Teiseks, mõned laused, mis on inglise keeles selgelt mitmetähenduslikud (nt *Every man loves a woman, A man who runs quickly eats*, jne) kuuluvad küll keelde ACE, kuid neil on ainult üks tähendus. Reeglina pakub ACE sellisel juhul välja alternatiivse konstruktsiooni, mis teist võimalikku tähendust esitab. Kasutaja peab omakorda selgeks õppima paarkümmend interpretatsioonireeglit, mis keele ACE poolt omistatud tähendust üldisel viisil selgitavad. Nt tuntud kvantoriskoobi näite

(9) *Every man loves a woman.*

puhul võib olemasolukvantoril olla inglise keeles nii lai kui ka kitsas mõjuala: st kas "leidub üks ja seesama naine, keda iga mees armas-

tab" või alternatiivselt "iga mees armastab mingit naist, kes pole tingimata seesama, keda teised mehed armastavad". Keeles ACE on see lause mõistetav ainult olemasolukvantori kitsas mõjualas ning alternatiivse tähenduse esitamiseks on hoopis konstruktsioon

(10) *There is a woman who every man loves.*

Vaatleme veel adjunkt- vs komplement-tüüpi mitmesust. ACE käsitleb prepositsioonifraasi alati adjunktina. Komplementditõlgenduse "pealesurumiseks" tuleb prepositsioon verbiga sidekriipsu abil ühendada:

(11) *A steward waits on the table.*

(12) *A steward waits-on the table.*

Seega väljub ACE vahel pisut inglise keele piiridest.

Põhimõtteliselt piisab kasutajal keele ACE süntaksi- ja interpretatsioonireeglite teadmisest, et ta suudaks iga teksti kohta öelda, kas see kuulub keelde ACE, ning kui kuulub, siis mis tähendus sellel tekstil on. Siiski, keelt ei õpita kunagi korraga selgeks, see toimub järkjärgult. Seega peab iga keele ACE implementatsioon (nt parser) andma kasutajale tagasisidet (nt parafraseerides ACE-keelset sisendit), tagades nõnda, et kasutaja arusaam loodud teksti tähendusest oleks keele ACE semantikaga kooskõlas.

Selleks et tagada keele loomulikkus ja kasutusmugavus, lubab ACE reeglina paljusid erinevaid vorme ühe ja sama tähenduse esitamiseks. Näiteks süntaktiliselt võib kasutada eitusemarkerit nii substantiivifraasi, verbifraasi kui ka terve lause ees, semantiliselt annab see aga ühe ja sama tulemuse; *every*-kvantori kasutamine on semantiliselt ekvivalentne *if-then* konstruktsiooni kasutamisega; adverb võib lauses esineda mitmes positsioonis lause tähendust muutmata, jne. Näiteks laused (13) ja (14) on keele ACE mõttes samatähendus-likud:

(13) *It is not the case that John's dog quickly walks.*

(14) *No dog of John walks quickly.*

Oluline tehnika, mis teeb loomuliku keele lihtsalt kasutatavaks, on anafoorid. ACE toetab asesõnade, definiitsete substantiivifraaside, pärisnimede ja muutujate kasutamist, viitamaks tekstis eelnevalt esinenud substantiivifraasidele:

(15) *There is a man$_i$. He$_i$ walks.*

(16) *There is a big man$_i$. There is a small man. The big man$_i$ walks.*

(17) *There is a man$_i$ who walks. His$_i$ dog sees a man who talks. The man$_i$ who walks eats.*

(18) *John$_i$ walks. John$_i$ talks.*

Substantiivifraaside tähistamine muutujatega ning nende hilisem kasutamine viitamiseks on kõige võimsam ja võibolla ka lihtsam viis anafooriliste viidete loomiseks. Samas sarnaneb see võte pigem programmeerimiskeeltega kui loomulike keeltega, mistõttu on see keeles ACE ainult üheks alternatiiviks:

(19) *There is a big man$_i$ A who eats. A$_i$ walks.*

Viitamine peab olema kooskõlas reeglitega, mis püüavad modelleerida anafooride kasutamist inglise keeles. Nt kvantori mõjualas olevale substantiivifraasile saab viidata ainult selle mõjuala seest:

(20) * Every man walks. *He talks.*

(21) * *No man has a dog. He likes it.*

(22) *Every man$_i$ who$_i$ walks talks.*

(23) *There is no man$_i$* who$_i$ has a dog$_j$ and who$_i$ likes it$_j$.

Keele ACE süntaksi ja semantika täielik ja detailne kirjeldus on esitatud ACE dokumentatsioonis (http://www.ifi.unizh.ch/attempto/tools).

3. Keele ACE parser APE

APE on keeles Prolog kirjutatud programm, mis teisendab ACE-keelse teksti kujule DRS (Discourse Representation Structure). APE on avalik REST-veebiteenus (vt http://www.ifi.unizh.ch/attempto/tools/documentation/ape_webservice.html).

3.1. Sõnavara

Lisaks vähesele hulgale eeldefineeritud funktsioonisõnadele (*if, then, for each, him,* jne), lubab ACE kasutada kõiki inglise keele sisusõnu (substantiive, verbe, adverbe ja adjektiive). Selle nõude täitmiseks kasutab APE 100 000 sõnavormist koosnevat inglise keele sõnastikku, mis põhineb sõnastikul COMLEX (Wolff jt 1998). Sõnastikuga

COMLEX võrreldes on APE sõnastik säilitanud vaid vähese informatsiooni, kirjeldades sõnade erinevad kirjutusvariandid, ainsuse ja mitmuse vormid, lihtsa semantilise tüübi (elus/elutu) ning verbide puhul informatsiooni transitiivsuse kohta.

Erialatekstide kirjutamiseks sellest sõnastikust siiski ei piisa, kuid parseri APE kasutaja võib uusi sisusõnu kirjeldada nn kasutajasõnastikus. Lisaks on võimalik uusi sõnu kirjeldada nn *online*-meetodil, märgendades sõnad tekstis sõnaliigi markeriga:

(24) *A man works in a n:pizza-delivery-service.*

Antud juhul tähistab marker *n*, et sõna *pizza-delivery-service* näol on tegu substantiiviga. Selline märgendamine ei spetsifitseeri sõna tüüpi täielikult, sest puudu on nt informatsioon sõna arvu kohta. Puuduv informatsioon järeldub aga sõna kontekstist, sest ainsuse artikkel tagab, et substantiiv on ainsuses.

3.2. Süntaksi ja semantika analüüs

Keele ACE süntaks on kirjeldatud Prologi grammatika Definite Clause Grammar reeglitega. Lisaks kasutatakse Prologi laiendust ProFIT (Erbach 1995), et siduda reeglitega tunnusstruktuurid (*feature structure*). Keele ACE grammatikas on kokku ligi 100 reeglit.

Igale süntaksireeglile vastab semantikareegel, nõnda et lause analüüsimisel luuakse süntaksipuu ning semantiline esitus samaaegselt. Parsimise lõpuks rakendub eraldi moodulina anafooride lahendaja.

APE on deterministlik parser, seega on garanteeritud, et igale sisendile vastab ainult üks analüüs. Kui sisend ei kuulu keelde, siis väljastab APE veateate. Seega erineb APE traditsioonilistest loomulikku keele parseritest, mis väljastavad sisendteksti põhjal mitu võimalikku analüüsi, üritades need kuidagi järjestada (nt korpusest kogutud statistika põhjal), ning mis püüavad igasuguse sisendi korral mingi tulemuse anda.

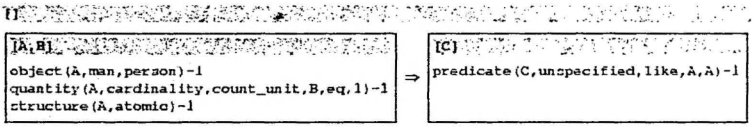
3.3. Semantika esitusformaat DRS

Parseri APE väljund on DRS-struktuur (Fuchs jt 2005). Attempto projektis kasutav DRS-keel on valdavas osas sama, mida kirjeldab Discourse Representation Theory (DRT) (Kamp jt 1993; Blackburn jt 1999).

DRT on 1980. aastatel alguse saanud formaalse semantika teooria, mis lisaks lausetele keskendub diskursusele (st mitmest lausest moodustatud üksusele), analüüsides põhjalikult lauseid omavahel siduvaid anafoore.

DRS (Discourse Representation Structure) kustutab esialgsest ACE-keelsest tekstist funktsioonisõnad, asendades need spetsiaalsete operaatoritega, toob ilmutatult välja sisusõnade tüübi ning omavahelised suhted (nt verbi ja selle komplementide vahelised suhted, vt joonis 1):

(25) *Every man likes himself.*



Joonis 1. Lause *Every man likes himself* struktuur

DRS keel on ekvivalentne esimest järku loogika keelega. Seega lihtsustub DRS struktuuridega (ja selle läbi ka keele ACE tekstidega) ümberkäimine: nende teisendamine erinevatesse formaalsetesse keeltesse, nendest informatsiooni ammutamine jne. Keele ACE kasutaja ei pea tegelikult DRS struktuuridest midagi teadma. Keele ACE roll ongi ju kasutajaid formaalsetest esitustest säästa. Suhtlus saab toimuda täielikult keeles ACE ning DRS on kasutusel ainult taustal, arvuti tasemel. Selles mõttes on DRS nagu masinkood (nt Java baitkood), mida programmeerija ei pea lugeda oskama, sest tema suhtleb arvutiga ainult programmeerimiskeeles (nt Java).

4. Abivahendid

Nagu iga keele puhul, peab ka piiratud loomuliku keele puhul kuidagi tagama, et tekstide kirjutajal oleks võimalik mingil viisil veenduda, et tekstide lugejad kirjutatust õigesti aru saavad. Piiratud keele puhul on "õigesti" formaalselt defineeritud ning aktsepteeritavaid tõlgendusvõimalusi on ainult üks. Kuna tekstide põhitarbijaks on arvutid, mis saavad asjast kindlasti "õigesti" aru, jääb üle ainult tagada, et teksti looja (ilmselt inimene) on ka ise tekstist just nõnda aru saanud. Seega on oluline luua lisaks parserile ka erinevaid vahendeid, mis inimest ACE-keelse teksti loomisel abistavad.

Tagasiside võib olla nii süntaktilist kui ka semantilist laadi. Süntaktiline tagasiside seisneb süntaktiliselt mittekorrektsete lausete raporteerimises ning vigade detailses kirjeldamises. Alternatiivne võimalus on süntaksivigade vältimine süntaksiteadliku tekstitoimeti kasutamise läbi (Schwitter jt 2003). Selline toimeti teavitab kasutajat igal sammul (st iga järgmist sõna sisestama asudes), millised sõnaliigid või isegi konkreetsed sõnad on antud positsioonis võimalikud, ega luba seega sisestada keelde mitte kuuluvaid vorme.

Süntaksivigade leidmine, nendest teavitamine ja neile (kasutajapoolne) reageerimine on võrdlemisi lihtne, kuid nagu programmeerimiskeelte puhulgi, algavad tõelised probleemid alles siis, kui tekst (programmi kood) on süntaktiliselt korrektne, aga millegipärast "ei tööta".

Programmeerimiskeelte puhul saab kasutada silumisvahendeid *(debugger)*, mis täidavad programmi käskhaaval ja raporteerivad vahetulemusi. Samuti saab programme testida komponenthaaval (eeldades, et korrektsete komponentide kogusumma on korrektne programm). Lisaks on võimalik tõestada programmide mõningaid omadusi (*for*-tsükkel ei rakendu mitte ühelgi juhul, muutuja on kasutamata jms). Need on erinevad viisid kasutaja tähelepanu juhtimiseks potentsiaalsetele vigadele. Põhimõtteliselt on kõik needsamad võimalused kasutada ka piiratud loomulike keelte puhul. Senini on Attempto projekti raames fookuses olnud teoreemitõestamine (Fuchs jt 2003) ja parafraseerimine (Fuchs jt 2005).

Teoreemitõestaja viib ACE-keelse teksti esimest järku loogika kujule, millele saab rakendada matemaatikas hästi läbiuuritud meetodeid vastuolude või liiasuste leidmiseks. Seega saab näidata, et tekst

(26) *Every man is happy and walks. If somebody walks then he is happy.*

on liiane ja tekst

(27) *John walks. No man walks.*

on vasturääkiv. Võib eeldada, et kasutaja ei tee liiasuse ja vasturääkivuse vigu meelega ning sellistest vigadest teavitamine on talle kasulik. Teoreemitõestaja võib otsida ka vastust teksti kohta käivale küsimusele:

(28) *Who walks?*

Parafraseerimise puhul viiakse sisendtekst semantiliselt ekvivalentsele, kuid süntaktiliselt teistsugusele kujule. Parafraseerides ACE-keelse teksti samuti keeles ACE, kasutatakse ainult selle väikest (nn Core ACE) alamhulka. Parafraasikeel allub seega samadele reeglitele mis tervik-ACE ning parafraasi mõistmiseks ei pea kasutaja uusi keelereegleid selgeks õppima. Samas on parafraasikeelest eemaldatud ainult konstruktsioonid, mis keele väljendusrikkust ei vähenda, st semantiliselt on parafraasikeel niisama rikas kui tervik-ACE, üksnes sünonüümiavõimalusi on kärbitud miinimumini. Teksti teisendamine süntaktiliselt erinevale kujule ja saadud kuju analüüsimine toob lähteteksti tähenduse paremini esile. Näiteks lause (29) ja selle parafraas (30):

(29) E*very man who loves a woman who loves him is happy.*

(30) *If a* woman loves a man and the man loves the woman then the man is happy.

Parafraasis on kadunud relatiivlause, üldsuskvantor *(every)* ning kasutatakse ainult üht tüüpi anafoorilisi viitu (definiitset substantiivifraasi). Seega võib väita, et parafraasikeelest on kadunud kompaktsus ja loomulikkus, mis keele ACE mugavaks tegid. Samas võib jälle vastu väita (vähemalt antud näite puhul), et parafraasi on lihtsam mõista, sest kõik tekstisisesed suhted on selgelt välja toodud. Lisaks erineb parafraas lähtelausest tunduvalt, rõhutades erinevate süntaktiliste konstruktsioonide samasust keeles ACE.

5. Rakendusvõimalused

Piiratud loomulik keel, mis on ühest küljest formaalne keel, kuid teisest küljest loomuliku süntaksiga ning kergesti õpitav, on kompromiss loomuliku keele ja "arvutikeele" vahel. Seega on piiratud loomulike keelte rakendamine mõeldav kõikjal, kus on tegu inimese ja masina vahelise suhtlusega. Tänu piiratud loomulikule keelele saab inimene anda masinale ülesandeid (nt andmebaasi päring, käsk robotile) loomulikul viisil ning samamoodi tagastab masin inimesele vastuse.

Konkureeriva meetodina võib siin nimetada graafilisi kasutajaliideseid, mis kasutavad suhtlusvahenditena ikoone, menüüsid, aknaid, hiirt jms, kuid erinevalt visuaalsest suhtlemisest arvutiga on teksti põhjal võimalik sünteesida kõnet. See võimaldab rikka suhtlu-

se arvutitega viia ka sellistesse keskkondadesse, kus visuaalne suhtlemine on raskendatud (nt autojuhtimine).

Keele ACE loomise algfaasis nähti keele peaeesmärgina tarkvaraspetsifikatsioonide kirjutamise toetamist. Nt tarkvarafirma klient (kes üldjuhul pole tarkvaratehnika formaalsete meetoditega tuttav) spetsifitseerib oma soovi ja kitsendused piiratud loomulikus keeles. See spetsifikatsioon teisendatakse automaatselt nt UML-skeemiks, mille põhjal alustab firma omakorda soovitud toote loomist.

Viimasel ajal arendatakse keelt ACE projekti REWERSE (http://rewerse.net) raames ning fookuses on keele ACE võimalik roll semantilises veebis. Semantilise veebi idee (Lee jt 2001) näeb ette, et kõik veebilehed ja veebiteenused on varustatud annotatsioonidega (nn metaandmetega), mis kirjeldavad vastava ressursi "tähendust" masinmõistetavas vormis. Selle tulemusena lihtsustuksid (automatiseeruksid) paljud tegevused veebis. Klassikaline kasutusnäide kirjeldab reisiplaneerimist, kus arvuti ostab iseseisvalt inimesele lennupiletid, reserveerib hotelli ja valib restorani, kus õhtust süüa.

Semantilise veebi aluskivi on RDF (Resource Description Framework), mis kirjeldab veebi *(subjekt,predikaat,objekt)*-tüüpi andmete hulgana. Raamistikule RDF ja sellel põhinevatele väljendusrikkamatele keeltele on välja pakutud peamiselt XML-kujulist süntaksit ning keskendutud enam keelte semantika defineerimisele ja väljendusjõu uurimisele. Keelte inimkasutatavus on seni jäänud tagaplaanile, aga seda lünka võimaldaksid täita just piiratud loomulikud keeled (Schwitter 2005; Marchiori 2004).

6. Kokkuvõte

Formaalsetel keeltel on selge süntaks ja semantika, st keelde kuuluvad konstruktsioonid on selgelt piiritletud keelde mitte kuuluvaist ning konstruktsioonide tähendus on nende vormi põhjal selgelt ja üheselt tuletatav. Loomulike keelte puhul see nii pole. Paljude konstruktsioonide korrektsus pole kõigile ühtmoodi selge, konstruktsioonid on lokaalselt väga mitmesed ning nende täpset tähendust on raske kindlaks teha. Piiratud inglise keel ACE on kompromiss arvutirakendustes seni kasutatatud keelte ja inimkeelte vahel, olles ühest küljest formaalne keel, kuid teisest küljest loomuliku süntaksiga ja inimesele kerge vaevaga mõistetav.

Kirjandus

Blackburn, Patrick, Bos, Johan 1999. Working with Discourse Representation Theory. An advanced Course in Computational Semantics.

Berners-Lee, Tim, Hendler, James, Lassila, Ora 2001. The Semantic Web. – Scientific American 284(5), 34-43.

Erbach, Gregor 1995. ProFIT: Prolog with Features, Inheritance and Templates. Proceedings of EACL'95. Dublin. http://arxiv.org/abs/cmp-lg/9502003

Fuchs, Norbert E., Hoefler, Stefan, Kaljurand, Kaarel, Schneider, Gerold, Schwertel, Uta 2005. Extended Discourse Representation Structures in Attempto Controlled English. IFI-2005.08. Department of Informatics, University of Zurich.

Fuchs, Norbert E., Kaljurand, Kaarel, Schneider, Gerold 2005. REWERSE Deliverable I2-D5. Verbalising Formal Languages in Attempto Controlled English I. http://rewerse.net/deliverables.html

Fuchs, Norbert E., Schwertel, Uta 2003. Reasoning in Attempto Controlled English. – Workshop on Principles and Practice of Semantic Web Reasoning (PPSWR 2003), Lecture Notes in Computer Science. Hannover: Springer.

Kamp, Hans, Reyle, Uwe 1993. From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic Publishers.

Marchiori, Massimo 2004. Towards a People's Web: Metalog. http://www.w3.org/RDF/Metalog/

Schwitter, Rolf 2005. Controlled Natural Language as Interface Language to the Semantic Web. – 2nd Indian International Conference on Artificial Intelligence (IICAI-05). Pune, India.

Schwitter, Rolf, Ljungberg, Anna, Hood, David 2003. ECOLE – A Look-ahead Editor for a Controlled Language. – Controlled Translation, Proceedings of EAMT-CLAW03, Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Application Workshop. Dublin City University, Ireland, 141–150.

Wolff, Susanne Rohen, Macleod, Catherine, Meyers, Adam 1998. COMLEX Word Classes Manual. http://cs.nyu.edu/cs/projects/proteus/comlex/manual-98.ps

LISA 1. Transkriptsioon

Tartu Ülikooli Eesti suulise keele korpuse transkriptsioonimärgid

Sõnad

- kirjakeeles olemas olevad sõnad, mille hääldus vastab kirjakeele tüüphäälduselе, kirjutatakse vastavalt kirjakeele ortograafiale;
- sõnad, mille hääldus ei vasta kirjakeelele või millel ortograafia puudub, kirjutatakse vastavalt häälduselе: *sis, vä, onju, nimodi;*
- pealkirjad ja nimed märgitakse samade reeglite järgi nagu muud sõnad. Nende alguses on suurtäht. Muid nimeortograafia võtteid ei kasutata.

Üneemid

- üneemid märgitakse tavaliselt kahekordse tähega: *aa, ee, õõ, mm*;
- tagasiside jaatussõna märgitakse kahe aa-ga: *jaa*;
- jaatusüneemid: *mhmh*; *õhõh, ähäh*;
- eitusüneemid: *mqm*; *õqõ, äqä.*

Liigendusüksused

Kõnevool jagatakse intonatsioonilisteks üksusteks, mitte lauseteks:

- lausungi lõpus on selgelt langev intonatsioon – seda märgib punkt;
- lausungid võivad jaguneda osadeks, mille lõpus on poollangev intonatsioon – seda märgib koma;
- tõusva intonatsiooniga lõppeva üksuse lõpus on küsimärk;
- lausungi algust eraldi ei märgita;
- üksinda lausungit moodustavate tagasisideüksuste puhul kasutatakse tasase häälduse märkimiseks ilma punktita sõna või üneemi (*mhmh*).

Pausid

(.) – mikropaus (0,2 sekundit või lühem);
(...) – mikropausist pikem paus;
(1.2) – pausi pikkus sekundites.

Rõhk ja intonatsioonitõus

`võimalik – terve sõna rõhutus või intonatsiooni tõus;
ho`tellis – ebatavaline või esimesest silbist kaugemal olev rõhuline silp.

Kõnetempo ja hääle tugevus

>.....< (sissepoole osutavad nooled) – kiirendatud lõik;
<.....> (väljapoole osutavad nooled) – aeglustatud lõik;
..... – muust kõnest vaiksem lõik;
AHA (suurtähed) – hääle kõvendamine.

Venitused
Häälikute venitused märgitakse kooloniga venitatud hääliku järel: *te:re.*

Naer ja köhatused
hehe – lahtise suuga naer;
s(h)õna (sulgudes olev *h* sõna sees) – sõna on lausutud naerdes;
$.....$ – naerva häälega öeldud sõna või pikem lõik, kuid mitte naer.

Jäljendamine ja aktsent
£....£ – muulaste aktsent.

Sisse- ja väljahingamine
.hh – häälekas sissehingamine (üks *h*-täht vastab 0,1 sekundile);
=h – sõna lõpul olev väljahingamine ja õhkamine *tule=h.*

Pooleli jäänud ja kokkuhäälдatud sõnad
- – sõna poolelijäämine osutatakse sidekriipsuga sõnaosa järel: *si-*;
= – kokkuhäälдatud eraldi sõnad seotakse võrdusmärgiga: *tulin=ja.*

Pealerääkimised ja haakumised
Pealerääkimised ühendavad erinevate kõnelejate voore:
[– pealerääkimise algus;
] – pealerääkimise lõpp;
= – kaks iseseisvat üksust on häälдatud kokku.

Ebaselgused
{*või*} – halvasti kuuldud tekstilõik või kõneleja nimi;
{--} – ebaselgeks jäänud sõna või kõneleja;
{---} – pikem ebaselgeks jäänud lõik.

Kommentaarid
((....)) – kommentaarid: ((tuleb laua juurde)).

LISA 2. Dialoogiaktide loend

Aktide nimetused koosnevad kahest osast:

1) kahe- või kolmetäheline lühend, kus kahetäheline märgib üksikakte (nt VR = vabatahtlik reaktsioon) ning kolmetäheline naaberpaariakte, kus kaks esimest tähistavad akti tüüpi ning kolmas märgib, kas tegemist on esi- või järelliikmega (nt KYE = küsimuse esiliige, KYJ = küsimuse järelliige);

2) akti nimi, mis annab akti semantilise/funktsionaalse sisu, nt DIJ: INFO ANDMINE.

Pikemalt vt Tiit Hennoste, Andriela Rääbis, *Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs.* Tartu 2004.

I. Naaberpaare moodustavad aktid	II. Üksikaktid
Naaberpaare moodustavad rituaalid RIE: TERVITUS RIJ: VASTUTERVITUS RIE: HÜVASTIJÄTT RIJ: VASTUHÜVASTIJÄTT RIE: SOOVIMINE RIJ: TÄNAMINE RIJ: VASTUSOOVIMINE RIE: TÄNAN RIJ: PALUN RIE: PALUN RIJ: TÄNAN RIE: VABANDUS RIJ: VABANDUSE VASTUVÕTMINE RIE: KUTSUNG RIJ: KUTSUNGI VASTUVÕTMINE RIE: LÕPUSIGNAAL RIJ: LÕPETAMISE VASTUVÕTMINE RIJ: LÕPETAMISE TAGASILÜKKAMINE RIE: MUU RIJ: MUU	**Üksikrituaalid** RY: TUTVUSTUS RY: ÄRATUNDMINE RY: KONTAKTEERUMINE RY: ÜLEANDMINE RY: MUU
Teemavahetus TVE: PAKKUMINE TVE: MUU TVJ: VASTUVÕTMINE TVJ: TAGASILÜKKAMINE TVJ: MUU	
Partneri algatatud parandused PPE: ÜMBERSÕNASTAMINE PPE: ÜLEKÜSIMINE PPE: MITTEMÕISTMINE PPE: MUU PPJ: LÄBIVIIMINE PPJ: MUU	**Üksi tehtavad parandused** PA: ENESEPARANDUS PA: MUU

I. Naaberpaare moodustavad aktid	II. Üksikaktid
Vastuse tingimuste täpsustamine VTE: VASTUSE TINGIMUSTE TÄPSUSTAMINE VTE: MUU VTJ: VASTUSE TINGIMUSTE TÄPSUSTAMINE VTJ: MUU	
Kontakti kontroll KKE: ALGATUS KKE: MUU KKJ: KINNITAMINE KKJ: MUU	
Direktiivid DIE: SOOV DIE: ETTEPANEK DIE: PAKKUMINE DIE: PALVE OODATA DIE: MUU DIJ: INFO ANDMINE DIJ: INFO PUUDUMINE DIJ: KEELDUMINE DIJ: NÕUSTUMINE DIJ: MITTENÕUSTUMINE DIJ: PIIRATUD NÕUSTUMINE DIJ: TEGEVUS DIJ: EDASILÜKKAMINE DIJ: MUU	**Üksik-infoaktid** YA: JUTUSTAMINE YA: LUBADUS YA: INFO ANDMINE YA: RETOORILINE KÜSIMUS YA: RETOORILINE VASTUS YA: REFERAAT YA: EELTEADE YA: JUTU PIIRIDE OSUTAMINE YA: MUU YA: PRAAK
Küsimused KYE: AVATUD KYE: JUTUSTAV KAS KYE: SULETUD KAS KYE: VASTUST PAKKUV KYE: ALTERNATIIV KYE: TÄPSUSTAV KYE: MUU KYJ: INFO ANDMINE KYJ: JAH KYJ: EI KYJ: NÕUSTUV EI KYJ: MITTENÕUSTUV JAH KYJ: ALTERNATIIV: ÜKS KYJ: ALTERNATIIV: MÕLEMAD KYJ: ALTERNATIIV: KOLMAS VALIK KYJ: ALTERNATIIV: EITAV KYJ: INFO PUUDUMINE KYJ: KEELDUMINE KYJ: EDASILÜKKAMINE KYJ: VASTUS ALTERNATIIVINA KYJ: TEGEVUS KYJ: MUU	**Vabatahtlikud reaktsioonid** VR: NEUTRAALNE INFO OSUTAMINE UUEKS VR: HINNANGULINE INFO OSUTAMINE UUEKS VR: NEUTRAALNE JÄTKAJA VR: HINNANGULINE JÄTKAJA VR: NEUTRAALNE PIIRITLEJA VR: HINNANGULINE PIIRITLEJA VR: NEUTRAALNE VASTUVÕTUTEADE VR: HINNANGULINE VASTUVÕTUTEADE VR: PARANDUSE HINDAMINE VR: MUU

I. Naaberpaare moodustavad aktid	II. Üksikaktid
Seisukohavõtud	**Infolisad**
SEE: VÄIDE	IL: SELETAMINE
SEE: ARVAMUS	IL: PÕHJENDAMINE
SEE: MUU	IL: JÄRELDAMINE
SEJ: NÕUSTUMINE	IL: KOKKUVÕTMINE
SEJ: MITTENÕUSTUMINE	IL: ÜLERÕHUTAMINE
SEJ: PIIRATUD NÕUSTUMINE	IL: PEHMENDAMINE
SEJ: KEELDUMINE	IL: HINNANG
SEJ: MUU	IL: TÄPSUSTAMINE
	IL: MUU