Emakeele Seltsi aastaraamat 57 (2011), 66–102
doi:10.3176/esa57.05

ARVUTILINGVISTIKA JA KEELETEHNOLOOGIA TARTU ÜLIKOOLIS[1]

KADRI MUISCHNEK, MARK FIŠEL, HEIKI-JAAN KAALEP, MARE KOIT, KAILI MÜÜRISEP, HEILI ORAV, KADRI VARE, HALDUR ÕIM

Annotatsioon. Käesolev artikkel annab ülevaate sellest, mis seisus on eesti keelega seotud arvutilingvistika ja keeletehnoloogia Tartu Ülikoolis: milline on uurimistöö temaatika, millised on saavutused ning milliste probleemide ja uurimisülesannetega praegu tegeldakse.
Võtmesõnad: arvutilingvistika, keeletehnoloogia, arvutimorfoloogia, automaatne süntaksianalüüs, automaatne semantiline analüüs, dialoogi modelleerimine, masintõlge, eesti keel

1. Sissejuhatus

Kõigepealt paar terminoloogilist märkust. Lähedasi mõisteid *arvutilingvistika* ja *keeletehnoloogia* on teoreetilistes lähenemistes üritatud üksteisest lahus hoida. Arvutilingvistika all on mõistetud kas lingvistika või arvutiteaduse (või isegi kirjandusteaduse) haru (nt Dale jt 2000: iii) ning keeletehnoloogia all infotehnoloogia haru (nt Õim 2002). Ka on mõlemaid – nii arvutilingvistikat kui ka keeletehnoloogiat – kirjeldatud kui lingvistika ja informaatika hübriideriala, mis tegeleb inimkeele ehk loomuliku keele arvutitöötlusega, kusjuures arvutilingvistika tegeleb sellega pigem teoreetilisel ja keeletehnoloogia pigem rakenduslikul tasandil (nt Viks 2002).

[1] Artikli valmimist on toetanud Euroopa Regionaalarengufond Eesti Arvutiteaduse Tippkeskuse kaudu, Eesti Teadusfond (grant 9124) ning Haridus- ja Teadusministeerium (projektid SF0180078s08, SF0180056s08, EKKTT09-57 ja EKT5).

Laias laastus sama valdkonna kohta on ingliskeelses kirjanduses kasutusel veel kolmaski mõiste, mida inglise keeles märgib *Natural Language Processing* (NLP), eesti keeles *keele masintöötlus* (Viks 2002). Selle termini kasutuse põhjal võiks NLP eestikeelse vastena kasutada ka *keeletehnoloogiat*, sest mõlemad tähistavad oma keelealal keele arvutitöötluse ja arvutianalüüsi rakenduslikumat poolt.

Laias mõttes hõlmavad kõik kolm eelnimetatud mõistet ka 'kõnetehnoloogiat', mida siiski sageli käsitletakse omaette distsipliinina. Kuna Tartu Ülikoolis kõnetehnoloogiaga ei tegelda (seda tehakse Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris ning Eesti Keele Instituudi keeletehnoloogia osakonnas), siis käesolevas artiklis kõnetehnoloogiat ei puudutata.

Hoolimata sellest, et arvutilingvistika ja keeletehnoloogia on teoreetilisel tasandil eristatavad, on praktikas nendevaheline piir hägune. Nii on ala juhtiva ajakirja nimi Computational Linguistics, kuid seal avaldatud artiklitest julgelt pooli võiks keeletehnoloogiaalasteks pidada. Nii keeletehnolooge kui ka arvutilingviste ühendab maailmaorganisatsioon Association for Computational Linguistics (ACL), mille korraldatavatel konverentsidel kõlavad ka keeletehnoloogiaalased ettekanded. ACL-i kodulehel[2] kasutatakse kogu valdkonna, st nii teoreetilisemate kui ka rakenduslikumate aspektide hõlmamiseks väljendit *arvutilingvistika ja keeletehnoloogia* (inglise originaalis: *CL and NLP*).

Kuigi mõlemad – nii arvutilingvistika kui ka keeletehnoloogia – asuvad lingvistika ja informaatika vahel, on arvutilingvistika siiski lähemal pigem keeleteadusele ning keeletehnoloogia arvutiteadusele. Eelnevat näib kinnitavat seegi asjaolu, et Tartu Ülikoolis õpetatakse erialamoodulit „Arvutilingvistika" filosoofiateaduskonnas ja valikmoodulit „Keeletehnoloogia" matemaatika-informaatikateaduskonnas.

Arvutilingvistika ja/või keeletehnoloogia ajalugu Tartu Ülikoolis on palju pikem kui vastavate erialade õpetamise ajalugu või mitteformaalsete arvutilingvistika ja keeletehnoloogia uurimisrühmade ajalugu. Nagu käesoleva artikli 7. osast lugeda võib, tegeldi eesti keele automaattöötlusega juba 1950. aastate lõpus, ülevaate ala arengust 1960. aastate keskpaigast alates annab nt Haldur Õimu artikkel (Õim 2009).

Loomuliku keele automaattöötluseks vajalikud vahendid võib jagada kaheks: keeletöötlustarkvara ja keeleressursid. Viimaste all mõeldakse

[2] http://www.aclweb.org.

keelekorpusi, leksikaalseid andmebaase ja formaalseid grammatikaid. Käesolevas artiklis tuleb keeleressurssidest juttu ainult niipalju, kui see on otstarbekas iga keele töötlemise tasandi või rakenduse käsitlemisel. Eraldi ülevaadet TÜ keeleressurssidest siin anda ei püüta, see oleks juba omaette pikem jutt.

Mitmes selle artikli osas võib kohata väljendeid *reeglipõhine* ja *statistiline* või *masinõppel põhinev* töötlus või süsteem, näiteks on eesti keele jaoks olemas nii reeglipõhine kui ka statistiline morfoloogiline ühestaja. Nende kahe – ratsionalistliku ja empiirilise – lähenemise olemusest, arenguloost ning hetkeseisust nii Eestis kui ka maailmas annab põhjaliku ülevaate Mare Koidu artikkel „Ratsionalism ja empirism keeletöötluses: vastasseis või koostöö?" (Koit 2006). Ratsionalism väidab, et olemuslikke seaduspärasusi (nt $E = mc^2$) ei saa leida, tuginedes kogemusele. See eeldus on aluseks reeglipõhisele keeletöötlusele: keelemudel tuleb arvutile ette anda. Empirism seevastu on seisukohal, et keelestruktuure õpitakse ainult kogemusest. Seega on keeletöötlus korpustesse koondatud keeleandmete statistiline töötlus ja masinõpe (st arvuti ise tuletab keelemudeli).

Nagu juba öeldud, annab käesolev artikkel ülevaate sellest, kuhu on praeguseks jõudnud ja milliste probleemidega hetkel tegelevad TÜ arvutilingvistid ja keeletehnoloogid. Alustatakse arvutimorfoloogiast kui eesti keele puhul vältimatust analüüsietapist, mille väljund on sisendiks nii süntaktilisele kui ka semantilisele analüüsile ning paljudele rakendustele. Edasi tuleb juttu morfoloogilisest ühestamisest ning süntaktilisest, semantilisest ja pragmaatilisest analüüsist. Artikli viimane osa käsitleb masintõlget, mis on ühtaegu nii vanim kui ka noorim TÜ arvutilingvistika uurimisteemade seas.

2. Arvutimorfoloogia

Morfoloogia uurib sõnade struktuuri ehk ehitust. Eesti keele uurimise traditsioonis on morfoloogia sama mis vormimoodustus, st pööramine ja käänamine; sõnade moodustamine tuletamise ja liitmise teel on aga hoopis sõnamoodustus. Arvutimorfoloogia kui sõnade automaatanalüüsiga tegelev valdkond hõlmab siiski ka sõnamoodustust.

Arvutimorfoloogia eesmärk on nii morfoloogia uurimine arvuti abil kui ka selliste töövahendite (programmide, leksikonide jms) loomine, mille abil saab sõnavorme analüüsida ja sünteesida ning mida keeleuurija saab

kasutada töövahendina oma uurimisülesannete (nt süntaksi, sõnavara) lahendamisel.

Kuna arvutit kasutatakse kui tööriista, mis peaks aitama töödelda inimkeelt sellisena, nagu teda kasutatakse, siis on nii uurija kui ka praktiku vaatepunkt korpuslingvistiline ning arvutimorfoloogia probleemid ja lahendused on sellega tihedalt seotud.

2.1. Sõnastikupõhine morfoloogia

Esimene asi, mis iga keele arvutimorfoloogias ära tehakse, on sõnastikul põhinev sõnavormide analüüs ja süntees. Sellel etapil on probleemiks esiteks korraliku sõnastiku tegemine, st sellise, mis kataks keele leksikaalset rikkust piisavalt hästi, ja teiseks sõnamuutmise ja -moodustuse algoritmide väljatöötamine, mis koostöös leksikonis oleva infoga võimaldaks töödelda reaalses keeles ette tulevaid sõnavorme. Näiteks kui leksikonis on sõna *vend*, siis peaks olema algoritm, mis võimaldab sellega siduda vormid *venna*, *vennale*, aga ka *vennalik* ja *naisevend* (koos nende muutevormidega).

Sõnastikupõhise morfoloogilise analüüsi puhul leitakse morfeemid, millest sõna koosneb, kasutades mitmesuguste morfeemide (nt lõppude, liidete, tüvede) loendeid ja nende kombineerimise eeskirju, kusjuures kõik morfeemid – ka tüved – on varem tuntud. Analüüsi tulemusena leitakse nii leksikaalset infot kandev sõna tüvi kui ka grammatilised kategooriad, mida sõnavorm esitab. Morfoloogiline süntees on analüüsi pöördprotsess – sõnavorm genereeritakse algvormi ja grammatiliste kategooriate alusel. Seejuures vajab lahendamist küsimus, kuidas formaliseerida morfotaktikat ja morfofonoloogiat, st kuidas esitada morfeemide loendeid ja nende kombineerimise ning teisenemise reegleid nii, et sõnastiku esitus oleks lihtne ja hästi struktureeritud ning nii analüüsi kui ka sünteesi algoritm elegantsed.

Nt sõna *kollane* tüvelõpu muutust *-ne* > *-se* võib sõnastikus esitada sel moel, et mõlema lõpuga tüvi on tüvede loendis otseselt antud, aga võib esitada ka kahetasemelise morfoloogiamudeli (Koskenniemi 1983) kohaselt, st nii, et sõnastikus on ainult üks leksikoni vorm, nimelt süvaesituses (*kollaNe*). Sõnavormi sünteesi puhul muudavad eraldiseisvad reeglid *Ne* kas *ne*-ks või *se*-ks, lähtudes moodustatava vormi grammatilistest tunnustest; analüüsi puhul töötavad samad reeglid tagurpidi, teisendades *ne* ja *se Ne*-ks, et leida sõna tüve süvakuju.

Eesti keele morfoloogia kirjeldamiseks on kahetasemelist morfoloogiamudelit kasutanud Heli Uibo (2006) ja Jaak Pruulmann-Vengerfeldt (2010).

2.2. Reeglipõhine morfoloogia

Üks probleem, millega pärast sõnastikupõhise morfoloogia realiseerimist silmitsi ollakse, on asjaolu, et sõnastikud ikkagi ei kata keelt kogu tema rikkuses. Uutes tekstides tuleb alatasa ette varem mittekohatud sõnu, nn sõnastikuväliseid ehk tundmatuid sõnu oma muutevormides. Iga sellise sõnavormi puhul on kaks küsimust: mis on selle sõna algvorm ja mis vorm ta ise on? Nt kohates sõnavormi *jorsi*, võime oletada järgmisi analüüsivariante: 1) sõna *jors* ainsuse omastav, osastav või sisseütlev; 2) sõna *jorss* ainsuse omastav; 3) sõna *jorsi* ainsuse nimetav või omastav; 4) tegusõna *jorssima* käskiva kõneviisi ainsuse teine pööre või kindla kõneviisi oleviku eitus; milline on õige variant, selgub ainult kontekstist (vt 3. osa „Morfoloogiline ühestamine").

Sõnastiku puudulikkus ei ole selline praktiline probleem, mille saaks lahendada üha uute sõnade lisamisega, vaid tegeliku keelekasutuse olemusest tulenev püsiv olukord. Ühe miljoni tekstisõna suuruse tekstikorpuse lemmatiseerimisel leidsid Kaalep ja Muischnek (2002: 14), et 32 000 sõna ehk 3,2% korpuse tekstisõnadest esines terves korpuses täpselt üks kord. Korpuse sõnavara koosnes 60 000 sõnast, seega üks kord esinevad sõnad moodustasid sellest üle poole. Goodi-Turingu valemi (Good: 1953) kohaselt on tõenäosus, et kohtame oma tekstikorpuses varem mittekohatud sõna, võrdne tõenäosusega, et kohtame mingit sõna täpselt üks kord. Baayen (2001: 53–54) on näidanud, et üks kord esinevate sõnade hulk (*hapax legomenon*) korpuse põhjal koostatud leksikonis (ja seega ka korpuses endas) suureneb koos korpuse suurenemisega, kusjuures korpuse suurenedes (nt 100 miljoni tekstisõnani) *hapax*'i kasvutempo küll väheneb, kuid ainult natuke. Morfoloogilise analüsaatori või süntesaatori poolt kasutatavas leksikonis on ainult sellised sõnad, mida keegi kuskil on varem kohanud (nagu mis tahes sõnastikus, mille keegi on teinud), st leksikoni võib käsitada kui tekstikorpuse alusel loodut (tekstikorpust tõlgendame siinkohal kui tegelikult aset leidnud keelekasutust). Seega võib oletada, et ligikaudu 3% eestikeelse sisendteksti sõnadest ei saagi sõnastikupõhisel moel analüüsida.

2.3. Rakendused praktikas

Praktiliselt kasutatav morfoloogiline süntesaator ja analüsaator Estmorf koos Markovi peitmudelil põhineva statistilise ühestajaga Estyhmm (vt 3. osa „Morfoloogiline ühestamine") valmis TÜ ja OÜ Filosofti koostöös üle 10 aasta tagasi (Kaalep, Vaino 2000). Hiljem on programmi pidevalt täiendatud ja ümber tehtud. Peale keele arengu ning ilmnenud vigade ja puuduste kõrvaldamise tingib ümbertegemise arvutustehnika kiire areng: vanad programmid tahavad pidevat kohandamist, et nad uutel riistvara- ja tarkvaraplatvormidel töötaksid.

Morfoloogiaanalüsaatori väljund on paljude keeletöötlusprogrammide (nt süntaksianalüsaatori, masintõlkeprogrammi) osaks. Morfoloogilist analüsaatorit ja ühestajat on kasutatud mh eesti kirjakeele sagedussõnastiku tegemiseks (Kaalep, Muischnek 2002), õpikute sõnavara analüüsiks (Asser jt 2004), mitmesõnaliste üksuste tuvastamiseks tekstist (nt Kaalep, Muischnek 2006), korpuste märgendamiseks (nt www.keeleveeb.ee, www.cl.ut.ee/morfkorpus/). Morfoloogilist süntesaatorit on kasutatud sõnastikesse vormimoodustusinfo lisamiseks (Kaalep, Mikk 2008a, 2008b). Kõigi nende rakenduste puhul on ilmnenud aspekte, mis on toonud kaasa programmide muutmise sellises ulatuses, et Estmorfi asemel kasutatakse praegu nime Etmrf ja statistilise ühestajana programmi T3mesta (põhineb trigramme kasutaval Markovi peitmudelil; Estyhmm kasutas bigramme). Algselt põhines Estmorf Ülle Viksi „Väikesel vormisõnastikul" (Viks 1992). Praeguseks on tüvede leksikon kasvanud poole suuremaks.

Sõnastikku mittekuuluvate sõnade puhul kasutatakse oletamist, st arvestatakse mitmesuguseid tõenäosuslikke reegleid võimalike morfeemide ja nende kombineerimise viiside kohta. Tõenäosused põhinevad reaalsete tekstide peal tehtud statistikal, kuid on siiski inimese poolt sõnastatud (Kaalep, Vaino 2000). Kasutades praeguseks olemasolevaid korpusi, saaks tundmatute sõnade oletamise algoritmi kindlasti palju töökindlamaks ja põhjendatumaks teha.

3. Morfoloogiline ühestamine

Paljusid ettetulevaid sõnavorme saab analüüsida mitmel võimalikul moel, nt sõnavorm *kanda* võiks olla nii algvormide *kand* kui ka *kandma* vorm; sõnavormi *printeritest* algvorm võiks olla nii *printer* kui ka *printeritest*. St oleme silmitsi morfoloogilise mitmesuse probleemiga.

Morfoloogilise ühestamise (ingl *morphological disambiguation*) ülesanne on valida morfoloogiaanalüsaatori poolt pakutud morfoloogiliste tõlgenduste hulgast välja konteksti sobiv. Morfoloogiaanalüsaator ise analüüsib ainult üht sõna korraga, konteksti ei arvesta. Lausete (1) ja (2) sõnavormi *tee* analüüs annab erineva tulemuse, ühel juhul on tegemist verbi käskiva kõneviisi vormiga, teisel juhul nimisõnaga.

(1) Tee tuba korda!
(2) Tee on valmis.

Nimisõnal *tee* on omakorda kaks tähendust, aga kuna neid väljendatakse morfoloogiliselt ühesugusel moel, siis morfoloogiline ühestamise seisukohalt neil vahet ei tehta.

Lausetes (1) ja (2) on sõnavormi *tee* võimalik morfoloogiliselt ühestada lausekontekstile toetudes, kuid lauses (3) ei ole lausekonteksti põhjal võimalik öelda, milline morfoloogiline tõlgendus on õige, sest kõrvuti asuvad kaks mitmest sõnavormi.

(3) Tee sai valmis.

Morfoloogiline ühestamine on traditsioonilisele keeleteadusele küllaltki võõras probleem, sest enamasti lauset lugedes see inimesele muret ei valmista. Samas, kui toimus korpuse morfoloogiline ühestamine käsitsi, siis jäid ka kogenud lingvistid hätta (vt Kaalep jt 2000). Eelkõige oli raskusi määrsõna-, kaassõna- ja nimisõnaklassi piiride kindlaksmääramisega, st otsustamisega, millal on tegu veel nimisõna muutevormiga ja millal uue leksikaalse üksuse – sellest muutevormist leksikaliseerumise teel kujunenud määr- või kaassõnaga. Nt lause (4) sõnavormi *käes* on võimalik analüüsida nii nimi- kui ka kaassõnaks.

(4) Lipp lehvis tuule käes.

Eestikeelse teksti automaatseks morfoloogiliseks ühestamiseks kasutatakse kaht lähenemist: statistikal põhinevat ja reeglipõhist.

Statistiline morfoloogiline ühestaja T3mesta õpib varem käsitsi märgendatud korpusest, millises kontekstis milline morfoloogiline tõlgendus esineb. Statistilised ühestajad arvestavad tavaliselt kahe-kolmesõnalist vasakut konteksti; T3mesta töötab trigrammidega, st arvestab peale vaadeldava sõna kahesõnalist konteksti.

Katsed programmi T3mesta treenimisel TÜ morfoloogiliselt ühestatud korpuse peal näitasid, et statistiline ühestaja valib 5% sõnadele vale

märgendi (Veskis, Liba 2008), kusjuures üle 8% sõnadest jääb mitmeseks. Seejuures leidsid katsetajad, et veaprotsent oleneb tekstiklassist, mida ühestatakse. Treenimiseks kasutatud tekstide klassist ja/või korpuse suurusest oleneb ühestamise täpsus palju vähem. Kõige vähem oli vigu juriidiliste tekstide puhul, kõige rohkem ajakirjanduse puhul, ilukirjandus ja populaarteadus jäid nende äärmuste vahele. Nende tulemuste põhjal võib öelda, et suurema tekstikorpuse loomine ja kasutamine treenimiseks poleks põhjendatud. Samuti võib ennustada, et treenimiskorpusest puudunud tekstiklasside peal töötab T3mesta sama hästi kui treenimiskorpuses sisaldunud tekstiklasside ühestamisel. Ühesõnaga, statistiline ühestaja T3mesta paistab olevat treenitud piisavalt suure ja heterogeense treeningkorpuse peal; selle kui trigrammidel põhineva ühestaja peamine puudus on see, et ta arvestab ainult lokaalset konteksti ja tulemust võiks parandada pikema konteksti arvestamine.

Kadri Kajaste näitas oma magistritöös (Kajaste 2009), et ka teised statistilised ühestajad (TreeTagger ja TnT), mida on treenitud sama korpuse peal, ei anna paremat tulemust. Edasist paranemist võiks saavutada ka reeglipõhise ja statistilise märgendaja kombineerimise teel, ning vastavaid eksperimente K. Kajaste ka tegi, kusjuures märgenduse püüdis ta jätta võimalikult informatiivseks. Korrektsete märgendite osakaaluks sai Kajaste 95,7%. 95-protsendiline korrektsus on ka teiste keelte statistilise ühestajate puhul levinud arvuline näitaja.

Reeglipõhine morfoloogiline ühestaja (Puolakainen 2001) tugineb kitsenduste grammatika formalismile (Karlsson jt 1995) ning koosneb ühestaja mootorist ja reeglite baasist ehk grammatikast. Iga reegel kujutab endast keelereeglilaadset eeskirja, mis arvestab konteksti ja morfoloogilist infot. Reeglid kas valivad tõlgenduste hulgast ühe välja või kustutavad ühe tõlgenduse ja jätavad teised alles. Näiteks lauses (2) eemaldatakse sõnavormilt *tee* verbitõlgendus, sest (osa)lauses leidub kindel finiitne verbivorm *on*. Kokku on ühestamisreegleid üle 1300. Reeglid on kirjutatud nii, et nad pigem jätavad sõna analüüsi tulemuse mitmeseks, kui eemaldavad korrektse tõlgenduse. Eesti keele kitsenduste grammatika morfoloogiline ühestaja suudab ühestada 85–90% sõnadest, tehes seejuures ligikaudu 2% vigu. Kuna reeglipõhist ühestajat on arendatud peamiselt ilukirjandusliku tekstikorpuse toel, siis need statistilised näitajad erinevad žanriti.

Võrdluseks meenutagem, et T3mesta eksis 5% sõnade märgendi valikul. Siinjuures tuleb arvestada aga erinevaid märgendusskeeme.

Reeglipõhise ühestaja märgendus koosneb umbes 800 unikaalsest märgendikombinatsioonist. Selline märgendus on väga informatiivne ja heaks sisendiks süntaktilisele analüüsile, kuid seda on raske, et mitte öelda võimatu statistiliselt treenida. Statistiline ühestaja väljastab 330 eri märgendit, järgides Ülle Viksi vormisõnastikku (Viks 1992), kus sama lõpuformatiivi kasutavad verbivormid, nt (*ma*, *sa*, ..) *teeks*, on esitatud ühe märgendina.

Kõik need automaatsed morfoloogilised ühestajad eeldavad, et sisendiks on kirjaliku keele tekst koos kirjavahemärkidega (erinevalt reeglipõhise süntaktilise analüüsi reeglitest, mida on kohandatud ka suulise kõne ja murdetekstide analüüsiks, vt täpsemalt eespool). Samas on Tartu Ülikoolis koostatud küllaltki suur suulise kõne korpus, mille automaatne analüüs aitaks võita oluliselt aega ja ressursse. Suulise kõne morfoloogiline ühestamine oleks vajalik ka automaatse kõnetuvastuse väljundi edasisel töötlusel.

4. Süntaktiline analüüs

Nii nagu võõrkeeleõppijale on peale teise keele sõnade teadmise väga oluline keele grammatilise süsteemi tundmine, on ka automaatne süntaktiline analüüs vajalik paljude keeletehnoloogiliste rakenduste jaoks, alustades automaatsest grammatikavigade tuvastajast või sisukokkuvõtete tegijast ning lõpetades dialoogsüsteemide ja masintõlkega.

Süntaktilise analüüsi mõiste on väga lai, kuid eesti keele kontekstis hõlmab see traditsiooniliselt lauseliikmete funktsiooni kindlaksmääramist. Muu maailma arvutilingvistikas on rohkem levinud fraasi- ja sõltuvusstruktuuride määramine.

4.1. Pindsüntaktiline analüüs[3]

Eesti keele akadeemiline grammatika (EKG II) käsitleb eesti keele süntaksit sõltuvusgrammatika põhimõtetest lähtuvalt, tuues küll sisse ka fraasi mõiste. Eesti keele automaatse süntaksianalüsaatori (Müürisep 2000) väljatöötamisel on tuginetud sellele grammatikale. Süntaksianalüsaator on reeglipõhine, rajaneb kitsenduste grammatika formalismile ja annab morfoloogiliselt analüüsitud ja ühestatud tekstile pindmise süntaktilise

[3] Pindsüntaktiliselt analüüsitud korpust vt lingil http://www.cl.ut.ee/korpused/syntaksikorpus.

kirjelduse. See tähendab, et analüüs ei näita, milline sõna millise sõnaga täpselt seotud on, kuid leitud on seoste liigid. Iga sõnavorm märgendatakse selle sõna süntaktilist funktsiooni näitava märgendiga. Kitsenduste grammatika morfoloogiline ühestaja ja süntaksianalüsaator kasutavad samu morfoloogilisi märgendeid ning esimese väljund sobib teise sisendiks.

Süntaktiline analüüs koosneb kahest etapist: kõigepealt lisatakse sõnavormile selle kõik võimalikud süntaktilist funktsiooni näitavad märgendid. Näiteks nominatiivis nimisõna saab olla subjekt, objekt, predikatiiv, eesatribuut, järelatribuut, adverbiaal, aga ka kuuluda adpositsiooni juurde. Järgmisel etapil hakatakse konteksti mittesobivaid märgendeid eemaldama (valikuvõimalusi kitsendama). Reeglid on samasuguse kujuga nagu kitsenduste grammatika reeglipõhise morfoloogilise ühestamise reeglidki. Kui kitsenduste grammatika morfoloogilise ühestaja ja süntaksianalüsaatori esimeses versioonis olid analüüsi etapid rangelt lahus ning morfoloogiline ühestamine ja süntaktiline analüüs toimusid täiesti eraldi moodulites, siis grammatika uues versioonis teevad nad koostööd.

Joonisel 1 on pindsüntaktiliselt analüüsitud lause *See oli osa vihkamise nädala eelsest kokkuhoiukampaaniast*.

```
$<s>
See
   see+0 //_P_ dem sg nom #cap //  **CLB @SUBJ
oli
   ole+i //_V_ main indic impf ps3 sg ps af #Intr // @+FMV
osa
   osa+0 //_S_ com sg nom // @PRD
vihkamise
   vihka=mine+0 //_S_ com sg gen #mine // @NN>
nädala
   nädal+0 //_S_ com sg gen //  @ADVL
eelsest
   eelne+st //_A_ pos sg el //  @AN>
kokkuhoiukampaaniast
   kokku_hoiu_kampaania+st //_S_ com sg el // @<NN
$.
   $. //_Z_ Fst //
$</s>
```

Joonis 1. Näide pindsüntaktiliselt analüüsitud teksti kohta

Nagu jooniselt näha, moodustab iga sõnavorm mitmerealise kohordi, kus esimesel real on sõnavorm ise ja järgmisel tema morfoloogilise analüüsi tulemus. Joonisel toodud näites on kõikidel morfoloogiliselt analüüsitud sõnavormidel ühene tulemus. Morfoloogiaanalüsaator on leidnud sõna tüve ja lõpud, //-sümbolite vahel on sõna morfoloogilise analüüsi tulemus ning @-sümboliga algab süntaktiline märgend (subjekt, finiitverb, predikatiiv jne). @CLB tähistab osalause piiri. Atribuutide märgendid näitavad küll põhja leidumise suunda (*vihkamise* @NN> on nimisõnaline eesatribuut), kuid atribuut ja põhi pole omavahel seotud (*kokkuhoiukampaaniast* @<NN on nimisõnaline järelatribuut, kuid pole teada, millist sõna ta täiendab). Samuti ei eristata fraasi- ja lauseadverbiaale (nt *nädala* @ADVL).

Kui sõnal võib olla lauses mitu funktsiooni, esitatakse need kõik. Enamasti on selline süntaktiline mitmesus põhjustatud morfoloogilisest mitmesusest (vt näide (5)), kuid näiteks adverbiaalatribuutide puhul võib juhtuda, et on raske otsustada, kas see sõnavorm talitleb üht sõna laiendava atribuudina või laiendab adverbiaalina kogu lauset. Mitme märgendiga jäävad ka sõnad, mida analüsaator pole suutnud lõpuni ühestada, nagu näiteks sõnavorm *koera* näites (5), mis võib olla morfoloogiliselt nii ainsuse genitiivi kui ka ainsuse partitiivi vorm ning süntaktiliselt nii hulgafraasi kuuluv sõna (partitiivsena) kui ka üks rinnastatud objektidest (genitiivsena).

(5) Päästjad tõid majast välja kaheksa inimest ja koera.

Grammatika koosneb 1300 süntaksireeglist ning varem käsitsi ühestatud kirjaliku keele tekstist suudab analüsaator üheselt analüüsida 85–90% sõnadest, tehes vigu ligikaudu 2%. Peamiselt jäävad mitmeseks adverbiaalid ja adverbiaalatribuudid, mille eristamine toimubki ju suuresti semantika põhjal. Nende eristamist statistika põhjal uuris Aivi Kaljuvee oma magistritöös (Kaljuvee 2008); masinõppimise teel on võimalik saavutada 85% korrektsus.

Kuigi süntaksianalüsaatori põhiversioon on arendatud normeeritud kirjaliku keelekasutuse analüüsiks, on tehtud ka katseid selle kohandamiseks normeerimata, kirjakeelest hälbiva keelekasutuse töötlemiseks.

Esiteks on katsetatud süntaksianalüsaatori kohandamist suulise keele korpuse[4] tekstide analüüsiks (Müürisep, Nigol 2008). Suulise keele süntak-

[4] http://www.cl.ut.ee/suuline1/suulisekorpus/.

tilist analüüsi lihtsustab suulise keele lausungi suhteliselt lihtne struktuur, mh sisestatud infiniittarindite vähesus.

Keerulisimaks probleemiks osutus osalausepiiride määramine, kuna erinevalt kirjalikust tekstist, mille töötlemisel said osalausepiiride määramise reeglid arvestada kirjavahemärkide olemasoluga, tuli suulise keele puhul piirduda osalausepiiride määramisel pauside ja sidesõnadega.

Teine suulise keele analüüsi raskendav iseloomulik tunnus on kõnekonaruste esinemine: sõnade ja fraaside kordamine, parandamine, katkestatud sõnad, pause täitvad partiklid. Nii võetigi suulise kõne tarbeks peale muudetud osalausepiiride määramise reeglite ja pisut täiendatud süntaksireeglite kasutusele ka kõnekonarusi tuvastav programm, mis otsib tekstist kõnekonaruste mustreid (näiteks korduseid) ja eemaldab need ajutiselt süntaksianalüsaatori sisendist. Eemaldatud sõnad lisatakse pärast analüüsi teksti tagasi spetsiaalse kõnekonarust tähistava märgendiga. Kõiki keelevääratusi ei ole aga selliselt võimalik üles leida ning automaatsel süntaktilisel analüüsil tekib 2,5–3,5% vigu. Samas on saadud analüüsitulemus väiksema mitmesusega kui kirjaliku keele analüüsi tulemus: mitme märgendiga jääb alla 10% sõnadest.

Teiseks on süntaksianalüsaatorit kohandatud murdetekstide analüüsiks. Tartu Ülikooli murdekorpuse[5] tekstid on küllaltki sarnased suulise keele korpusele, tegemist on suuliste intervjuudega, mis on osaliselt käsitsi morfoloogiliselt analüüsitud ja ühestatud, mis teeb võimalikuks ka süntaktilise analüüsi (Lindström, Müürisep 2009). Peale suulise keele joonte (raskused osalausepiiride määramisel, kõnekonarused) on murdetekstides paljud süntaktilised konstruktsioonid, aga ka sõnavara ja rektsioonid murdespetsiifilised ning nii jäi hoolimata grammatika täiendamisest süntaksianalüsaatori korrektsus 95–97,5% vahele. Süntaktilise mitmesuse näitajad olid samad kui suulise keele korralgi – 90–93% sõnadest said ühese analüüsitulemuse.

[5] http://www.murre.ut.ee/korpus.html.

4.2. Sügavam süntaktiline analüüs ja puude pank[6]

Pindsüntaktiline analüüs sobib mitmetele keeletehnoloogilistele rakendusprogrammidele, kuid keele edasistel analüüsietappidel on vaja leida lause (puukujuline) struktuur.

Lause puustruktuuri saab leida kahte moodi: fraasistruktuuripuu kirjeldab, millistest fraasidest lause koosneb, sõltuvuspuu on pigem suunatud graaf, mille tippudeks on sõnad ja kaarteks nende sõnade vahelised seosed.

Lausepuule sobivat formaati otsides pidi arvestama asjaoluga, et loodav süntaksipuude panga märgendus oleks ühilduv või teisendusrelatsioonis Põhjamaade paralleelse puude panga märgendamiseks valitava formalismiga, sest siis oleks puude panga loomiseks ja kasutamiseks võimalik kasutada mujal välja töötatud tarkvara, näiteks päringu- ja visualiseerimisvahendeid. Selleks võeti kasutusele Lõuna-Taani Ülikoolis loodud märgendussüsteem VISL (*Visual Interactive Syntax Learning*), mis kohandati eesti keele süntaksile. Eesti keele puude panga Arborest[7] esimene versioon genereeriti poolautomaatselt eesti keele kitsenduste grammatika korpusest fraasistruktuurireeglite abil (Bick jt 2004). Automaatselt genereeritud puudest olid korrektsed ainult kuni 40%.

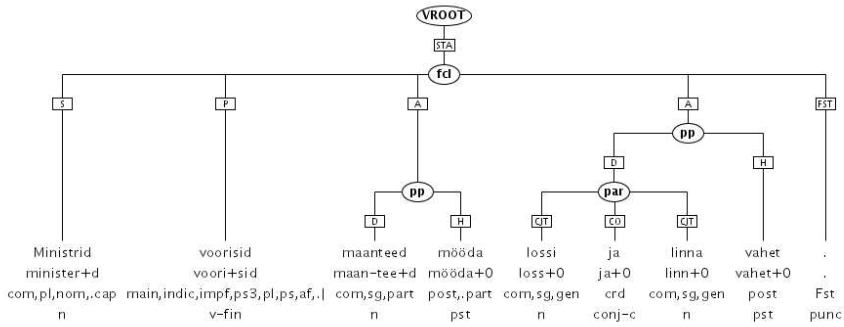
Lihtlause semantilise analüüsi (vt järgmine osa) arendamiseks vajaliku semantiliste rollide suhtes märgendatud lihtlausete korpuse loomiseks koostati uus, 40 fraasistruktuurireeglit sisaldav grammatika (Müürisep jt 2008). Reeglid teisendavad süntaktiliselt märgendatud laused fraasistruktuuripuu kujule, nii et fraasipuu ülemustipu küljes on fraasi tüüpi näitav märgend ning ülemustipu ja alluva vahelise kaare küljes on süntaktilist funktsiooni näitav märgend[8] (vt joonis 4).

---

[6] Puude pank (ingl *treebank*) – korpus, milles on esitatud lausete (puukujuline) süntaktiline struktuur.

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

[8] A – adverbiaal, CJT – konjunkt, CO – konjunktsioon, D – laiend, FST – lauselõpumärk, H – põhi, P – predikaat, S – subjekt, STA – lause; fcl – finiitne osalause, par – paratagma (rindtarind), pp – postpositsioonifraas.



Joonis 2. Fraasistruktuuripuu koos süntaktiliste märgenditega

Kahjuks on selline grammatika töökindel vaid lihtlausete puhul, keerulisemate struktuuride kirjeldamine ei olnud selles formalismis võimalik. Selle asemel võeti kasutusele sõltuvusstruktuuri ehitav grammatika, mis on otseselt integreeritud kitsenduste grammatika uue versiooni analüsaatorisse. Loodud on 50 reeglit, mille abil koostatud lihtlausete analüüsipuude korrektsus on üle 90%, kuid töö osalausete funktsioonide määramise ja infiniitsete verbikonstruktsioonide analüüsiga veel käib.

Joonisel 3 on toodud näide automaatselt tuvastatud sõltuvusstruktuuri kohta, sisendiks lause (6). Süntaktiliste funktsioonide märgenditele lisaks on näidatud, milline sõna millise juurde kuulub.

(6) Kuigi president andis stuudiosse saabudes oma rivaalile kätt, ei näinud televaatajad neid kordagi kõrvuti.

Analüüsitud on näite (6) esimest osalauset, mille sõltuvuspuu juurtipuks on finiitne verbivorm *andis* (lause 3. sõna), mis allub pealause finiitsele verbile (*näinud*, lause 11. sõna või kirjavahemärk). Subjekt, objekt ja adverbiaal alluvad predikaatverbile, gerundiivi laiendav adverbiaal, mis pindsüntaktilise analüüsi korral ei olnud eristatav lauseadverbiaalidest, viitab nüüd otseselt gerundiivile.

Selle sõltuvusstruktuure analüüsiva grammatika kirjutamine on veel pooleli, samuti puuduvad esialgsed hinnangud selliselt genereeritud puude korrektsusele. Sõltuvusseostega analüüsi saab teisendada joonisel 2 olevale fraasistruktuuri kujule.

```
“<Kuigi>”
   “kuigi” <kuigi+0> J sub <cap>  @J #1->3
“<president>”
   “president” <president+0> S com sg nom  @SUBJ #2->3
“<andis>”
   “and” <and+is> V main indic impf ps3 sg ps af <FinV> <NGPP> @FMV
#3->11
“<stuudiosse>”
   “stuudio” <stuudio+sse> S com sg ill  @ADVL #4->5
“<saabudes>”
   “saabu” <saabu+des> V main ger <Intr> <Ill>  @ADVL #5->3
“<oma>”
  «oma» <oma+0> P det pos refl sg gen  @NN> #6->7
”<rivaalile>”
  ”rivaal” <rivaal+le> S com sg all  @ADVL #7->3
”<kätt>”
   “käsi” <käsi+tt> S com sg part  @OBJ #8->3
```

Joonis 3. Sõltuvusseostega analüüsitud lause

5. Semantika

Ka semantika puhul ei hakka me kirjeldama n-ö asjade algust, mis läheb tagasi 1980. aastatesse, kui tehisintellektiuuringute, täpsemalt TARLUSe (= *TARtu Language Understanding System*) raames üritati ka lauseid-tekste semantiliselt analüüsida (Õim 2009; vt nt ka Roosmaa, Saluveer 1983). Võiks vaid märkida, et tollesse aega lähevad tagasi mõned teoreetilised uurimisliinid, mis ulatuvad tänapäeva, nt freimi mõiste käsitlus (Õim, Saluveer 1985), mis on ka meie praeguse (lause)semantika alus.

Praegused semantikaalased tööd võib jagada leksikaalse semantika ja lausesemantika vahel, kusjuures esimesega on tänapäevases mõttes tegeldud pidevalt viieteistkümne aasta ringis, teisega viis-kuus aastat.

Leksikaalse semantika rakenduslike harudena antakse lühike ülevaade ka sõnatähenduste ühestamisest ja ka pärisnimede liigitamisest.

5.1. Wordnet

TÜ arvutilingvistika uurimisrühmas on leksikaal-semantilist andmebaasi nimega Eesti Wordnet koostatud mitmes etapis. Esmalt alustati tesauruse loomist EuroWordNeti projekti raames[9], mille eesmärk oli luua Euroopa keeli kaasav mitmekeelne *wordnet*-tüüpi tesaurus, eeskujuks Princetonis 1980. aastatel loodud WordNet[10]. EuroWordNeti projekti lõppemise järel ei arenenud mõni aasta ka Eesti Wordnet edasi, kuid alates 2007. aastast tänu riiklikule programmile „Eesti keele keeletehnoloogiline tugi (2006–2010)" ja selle jätkuprogrammile „Eesti keeletehnoloogia (2011–2017)"[11] on uuema põlvkonna arvutitesaurus jõudsalt kasvanud. Praeguseks (märts 2012) on andmebaasis umbes 53 000 mõistet, mis katab eesti keele sõnavara suhteliselt mittetäielikult, st süsteemselt on läbi töötatud ainult teatud teemad (nt muusikariistad, transpordivahendid, arhitektuur jms).

Kui tavasõnastikus on märksõnaks tüüpiliselt üksiksõna, siis *wordnet*-tüüpi tesaurus on oma ülesehituselt mõistepõhine, st selle tesauruse põhiüksuse – sünohulga – moodustavad ühe mõiste väljendamiseks kasutatavad üks või mitu sünonüümset sõna. Näiteks leidub tesauruses nimisõnaline sünohulk: *teksased*, *teksad*, *džiinid*, *teksapüksid*; tegusõnaline sünohulk: *õpetama*, *harima*, *koolitama*; omadussõnaline sünohulk: *normaalne*, *vastuvõetav*, *talutav*, *inimlik*; ja määrsõnaline sünohulk: *sageli*, *sagedasti*, *tihti*.

Toodud näited on selged, kuid tesauruse koostamise igapäevatöös tuleb tihti tegeleda mõistete kindlaksmääramise ja defineerimisega ehk siis küsimustega, mis on mõiste ja kas mingi mõiste väärib Wordnetti sisestamist või mitte. Mõiste olemusest teoreetilisel tasandil on kirjutatud mitmeid artikleid (nt Õim 2011; Vainik, Kirt 2008) ja raamatuid (nt Erelt 2007), kuid kui minna praktikasse, on piirid mõistet esindavate sõnade vahel tihti ähmased. Üheks näiteks võib tuua sõnaliigilise ebakõla: *habemik*, *habemega*, *parrakas* on sõnavormidena sünonüümsed, kuid sõnastiku formaati silmas pidades peaksid need vormid kuuluma kahte erinevasse sünohulka – nimisõnalisse ja omadussõnalisse, st toimub üleeristamine. Üleeristamisega on seotud ka liitsõnad, mida eesti keeles võib moodustada piiranguteta ning mis tesaurusesse satuvad reaalsetest tekstidest sõnatähenduste ühestamise (vt osa 5.2) kaudu. Näiteks sünohulga *õpetaja*,

[9] www.illc.uva.nl/EuroWordNet/.

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

[11] http://www.keeletehnoloogia.ee/.

pedagoog alammõisteks on nii sattunud liitsõnad *rahvakooliõpetaja*, *vallakooliõpetaja*, *lemmikõpetaja* jne, ehkki *rahvakooli* ja *vallakooli* tähendused tesauruses puuduvad.

Sünohulgad Eesti Wordnetis on omavahel ühendatud 43 semantilise seosega (vt lähemalt Orav jt 2011), millest põhirõhk on olnud hierarhiaid moodustavatel suhetel – hüponüümia- ja hüperonüümiasuhtel. Teiste suhete lisamisega olukord nii lihtne pole. Näiteks mõistel *klassijuhataja* on ainult ülemmõiste *õpetaja*, *pedagoog*, kuid seotud peaks ta olema ka *klassi* ja ehk ka *juhatamisega*, ehkki seosetüüpi on keeruline määratleda. Kas *klassijuhataja* kuulub *klassi* juurde või vastupidi?

Eesti Wordneti koostamine on põhiliselt olnud käsitsitöö. Üks põhjusi selleks oli see, et algul polnud elektroonilisi arvutiressursse, mida oleks olnud võimalik lihtsal viisil tesauruse koostamisel kasutada, ja nüüdseks oleme suurendanud tesaurust valdkondliku sõnavaraga, mis ühtlasi on toonud kaasa selle, et mõisted on muutunud kohati väga erikeelseks ja spetsiifiliseks. 2012. a alguses alustasime uute mõistete lisamist ka värske tekstikorpuste põhjal tehtud sõnaloendiga[12], sest see tagab tesauruse parema kvaliteedi – üldkeele mõistete osa saab kindlamalt ja korrektsemalt kaetud.

Edasi on arenenud ka Eesti Wordneti formaat. Nimelt on projekti META-NORD[13] raames tesauruse andmebaas viidud MySQL-i formaati ja lingitud Princetoni Wordnet 3.0 versiooniga, mis annab meile võimaluse võrrelda oma andmeid ingliskeelse Wordnetiga (nt eristada baasmõisteid, st mõisteid, millel on suur hulk alammõisteid; võrrelda, mis valdkondade sõnavara erineb tugevalt jms) ja võimaldab lisada valdkonnamärgendeid.

Keele leksikaal-semantiline andmebaas, kus peale sõnade tähenduste eristamise on fikseeritud ka tähendustevahelised seosed, on oluline nii lingvistiliseks uurimistööks kui ka arvutilingvistilisteks rakendusteks, nagu sisupõhine infootsing, automaatne refereerimine, masintõlge, keeleõpe. Ka järgnevalt kirjeldatud rakendus – sõnatähenduste ühestamine – poleks eesti keeles võimalik ilma Eesti Wordnetita.

---

[12] http://www.cl.ut.ee/ressursid/sagedused1/.

[13] Euroopa Liidu ICT PSP (CIP-ICT-PSP.2010-4 Theme 6: Multilingual Web: Machine translation for the multilingual web) projekti „META-NORD - Euroopa avatud lingvistilise infrastruktuuri Balti- ja Põhjamaade haru" eesmärk on Põhjamaade keeletehnoloogiliste ressursside ja vahendite kindlaks tegemine, standardiseerimine ja kogumine (vt ka http://www.cs.ut.ee/metanord/ ja ttp://www.meta-nord.eu/).

5.2. Sõnatähenduste ühestamine

Sõnatähenduste ühestamine (ingl *word sense disambiguation*) on olnud üks arvutilingvistika uurimisülesandeid alates 1950. aastatest, kui hakati tegelema masintõlke ja tehisintellektiga. Sõnatähenduste ühestamine seisneb mitmetähenduslikkuse probleemi lahendamises, st üritatakse leida polüseemse sõna tähendus konkreetses kontekstis. Polüseemia on keeles levinud nähtus, näiteks on nimisõnal *asi* eesti keeles (Eesti Wordneti järgi) 12 tähendust; verbil *käima* on tähendusi tervelt 23. Siiski ei valmista polüseemia inimesele tavaliselt erilisi mõistmisraskusi.

Arvutilingvistikas ja eriti keeletehnoloogias on tähenduste ühestamisel vahendav roll: tuleb välja selgitada, mis tähenduses mitmetähenduslik sõna konkreetses tekstis ja kasutuses esineb, selleks et rakendusprogramm, nagu nt infootsing või masintõlkeprogramm, saaks oma töös just sellest tähendusest lähtuda.

TÜ arvutilingvistika uurimisrühmas on sõnatähenduste ühestamisega tegeletud alates 2001. aastast. Erinevad sõnatähenduste ühestamise meetodid kasutavad eri ressursse. Kuna TÜ-s luuakse leksikaal-semantilist andmebaasi Eesti Wordnet, siis on loomulik, et Kaarel Kaljurand kirjutas just sellele toetuva automaatse ühestamisprogrammi Semyhe[14]. Aastal 2001 hakati looma ka ühestatud sõnatähendustega korpust[15], mis kajastab nn loomulikku keelekasutust. Ühestatud sõnatähendustega korpus võib olla kas selline, kus ühestatud on ainult valitud hulk sõnu, või korpus, kus on märgendatud kõik sõnad. Kõikide sõnade märgendamine on küll ajamahukam ülesanne, ent siis on korpuses sisalduv informatsioon kõige sarnasem loomuliku keelega töötavale lõppsüsteemile. Ka TÜ ühestatud sõnatähendustega korpuses on valitud meetodiks just kõikide sõnade märgendamine. Korpuse märgendamiseks ühestati sõnatähendusi käsitsi, sisendiks oli morfoloogiliselt ühestatud tekst, milles substantiividele ja verbidele olid lisatud viited nende sõnade tähendustele tesauruses. Kui sõna juures oli mitu viidet, pidi inimene valima nende hulgast konkreetses kontekstis õige; iga teksti ühestas kaks inimest. Pärast väikest pausi hakati 2009. aastal korpust täiendama kahes liinis. Ühestatavate sõnade hulka lisandusid adjektiivid ja adverbid ning kui eelmisel etapil ühestati ilukirjandustekste, siis nüüd märgendati ka teisi tekstiliike, eelkõige

[14] http://math.ut.ee/~kaarel/NLP/semyhe30/.

[15] http://www.cl.ut.ee/korpused/semkorpus/.

ajakirjandus- ja teadustekste. Sõnatähendusi on endiselt käsitsi märgendatud, kasutusele on võetud käsitsiühestamist hõlbustav programm KYKAP[16] ning ühestajatel on kasutada juhised, millal üht või teist tähendust valida (vt Kerner 2007). Korpuse maht (2011. aasta lõpuks ligikaudu 400 000 sõna) on kasvanud juba piisavalt suureks, et seda on võimalik kasutada väärtusliku treeningmaterjalina automaatse sõnatähenduste ühestamise programmi jaoks. Samuti sisaldab korpus olulist statistilist informatsiooni sõnatähenduste esinemise sageduse kohta.

5.3. Pärisnimede liigitamine tähenduse järgi

Loomuliku keele töötlemise problemaatikas on pärisnimedel eriline koht. Ühelt poolt jäävad nimed lingvistika vaateväljast kõrvale: sõnastikes neid ei esitata, süntaksis ja semantikas käsitletakse neid möödaminnes (kui üldse). Teiselt poolt pakuvad pärisnimed kui tähenduse jaoks ülimalt olulised üksused suurt huvi reaalsetes teksti sisuga seotud ülesannetes, nt infootsingus ja teksti põhjal järelduste tegemises. Kolmandaks avaldub keeleline loomingulisus nimede kasutamisel eriti ilmekalt: uued pärisnimed ilmuvad kasutusse ja vanad kaovad palju kiiremini kui üldnimed, nii et pole lootustki nimede töötlemisel loota üksnes varemloodud sõnastikule või ontoloogiale; ainus realistlik lähenemine on tekstipõhine.

Eesti ortograafiat arvestades on lihtne ära tunda, kas sõna esindab nime: pärisnimi kirjutatakse suure algustähega. Probleemiks on hoopis küsimus, mida see nimi tähistab: inimest, kohta, organisatsiooni, sündmust vms? Sisuliselt on tegemist tundmatu sõna tähenduse ligikaudse määramise ülesandega, kusjuures seda lahendades saab tugineda peamiselt tekstile endale. Tavaliselt eeldatakse, et tekstis on mingid tunnused, mis võimaldavad nimesid liigitada – nt perekonnanime ees on eesnimi või tiitel (*hr*, *dr* vms); omastavas käändes oleva kohanime järel on kohta tähistav üldnimi *tee*, *tänav*, *mägi* vms. Ülesandeks ongi need tunnused leida ja teha kindlaks nii tunnuste usaldusväärsus kui ka praktiline rakendatavus. Tunnused jagunevad laias laastus kolme gruppi: sõnapõhised (nt liitsõna puhul viimase komponendi liik: *Draamateater* on teater, st organisatsioon), loendipõhised (nt eesnimede loend) ja kontekstipõhised (seotud liigitatava sõna lähema ümbrusega, nt eesnimele järgneb perekonnanimi). Millised

[16] http://www.keeletehnoloogia.ee/projektid/eesti-keele-semantika-ressursid-ja-vahendid/projekti-tulemused/tarkvara.

tunnused täpselt valida ja millise kaaluga neid arvestada (sest mõned tunnused võivad olla omavahel vastuolus, nt *Vladimir* on nii mehenimi kui ka linnanimi), on vägagi keeruline küsimus, mistõttu on tavapärane, et kasutatakse statistilisi masinõppe meetodeid. Esimene eestikeelsetes tekstides pärisnimesid liigitav programm kasutas treeningkorpusena 84 000 sõna suurust pärisnimede suhtes märgendatud Delfi uudiste korpust (Tkachenko 2010) ja liigitas nimed isiku-, koha-, organisatsiooni- ja asutusenimedeks.

5.4. Lihtlause semantiline analüüs

Kolmas semantikaalane uurimissuund on seotud lausete automaatse semantilise analüüsi programmi loomisega (vt ka Õim jt 2009 ja ESA kontekstis üldteoreetilisemast aspektist Õim jt 2010).

Eesti keele puhul on semantiline analüüs lausetasandil saanud reaalselt võimalikuks seetõttu, et ühelt poolt on olemas (liht)lausete süntaksianalüüsi programm ning teiselt poolt on semantiline andmebaas (Wordnet) saavutanud taseme, kus nende ühitamisel on võimalik konstrueerida lause semantiline esitus. Alustuseks piisaks programmi prototüübi loomisest, mis suudaks morfoloogilise ja süntaktilise infoga varustatud lausetele lisada juurde ka lauseliikmete semantilised rollid (*Agent* 'agent', *Object* 'semantiline objekt' jm).

Esialgu oli mõttekas piirata semantilist analüüsi teatud kitsama ontoloogilise valdkonnaga. Selleks valiti liikumisega (k.a liigutamine, asetamine jne) seotud situatsioonid. Põhjus on selles, et see valdkond on tähelepanu keskmes olevaid alasid teoreetilises semantikas, aga ka mitmesugustes rakendustes (nt robotid). Liikumisega on vältimatult seotud ruumisuhted: füüsiline liikumine toimub alati teatud viisil füüsilises ruumis ja liikumise eri viise iseloomustatakse ruumisuhete kaudu (kus, kust, kuhu, mis suunas jne, nende suhete osutamiseks on eesti keeles omad spetsiifilised vahendid). Esialgu oleme analüüsinud ainult lihtlauseid, s.o lauseid, kus pole alistavaid sidesõnu ega alistusseoses olevaid osalauseid ega ka sisestatud infiniittarindeid. Lühidalt: vaatleme lauseid, kus on vaid üht sündmust esitav predikaat-argumentstruktuur, et vältida sündmustevaheliste seoste (nt põhjuslike) tuvastamise probleeme.

Selleks, et teha ükskõik millist keeletehnoloogilist programmi, on ennekõike vaja mingit andmehulka, millel hakata loodavat programmi katsetama. Lihtlause semantika uurimise huvides oli vaja koostada

võimalikult erinevate lihtlausete näidiskorpus. Selleks otsustati ära kasutada olemasolev uurimus eesti lihtlausete kohta ja moodustada lihtlausete kontrollitud korpus Huno Rätsepa raamatu „Eesti keele lihtlausete tüübid" (Rätsep 1978) näitelausetest. Korpuse maht on 30 000 tekstisõna ja ta sisaldab 4323 lihtlauset.

Valitud valdkonna jaoks (ruumiline liikumine ja liigutamine) töötati järgmisena välja semantiliste rollide loend, mis seoks lause süntaktilist ja semantilist esitust. Süntaktiliselt analüüsitud ja puukujuliselt esitatud lausetele (täpsemalt lauseliikmetele) tuli hakata omistama semantilisi rolle, mida nad konkreetses sündmuses täidavad: nt predikaadiga 'viima' moodustatud lauses võib olla *Agent* 'agent', *Object* 'semantiline objekt', *Instrument* 'liikumis- või liigutamisvahend', *Locfrom* 'lähtekoht', *Locto* 'sihtkoht' jne.

Nii saab joonisel 2 kujutatud lauses *Ministrid voorisid teed mööda lossi ja linna vahet* subjekt *ministrid* märgendi *Agent* 'agent', *teed mööda* märgendi *Path* 'rada' ja *lossi ja linna vahet* märgendi *Location* 'liikumiskoht'.

Rollide automaatseks määramiseks tuleb meid huvitava predikaadi iga võimaliku rolli juures esitada vähemalt kaht liiki informatsiooni: 1) rolli võimaliku täitja semantiline iseloomustus (eelkõige tema semantiline kategooria, nt elusolend, füüsiline objekt, vedelik jne, aga tihti on vaja ka täpsustavat lisainfot, nt mitte iga füüsiline objekt ei saa veereda, oluline on ka objekti kuju jne) ja 2) rolli täitva väljendi morfoloogiline iseloomustus, nimisõnade puhul nt võimalikud käänded selles rollis, kaassõnad, millega koos saab mingi nimisõna täita mingit kindlat rolli. Praegu tegeldakse selle huvitava probleemiga, kuidas lausete semantilises esituses kajastada nn varjatud argumente, st argumente, mida lauses endas süntaktilise üksusena ei pruugi esineda, aga semantilise analüüsi jaoks on nad vaja lisada, sest nad on obligatoorsed ja neid võidakse hiljem tekstis kuidagi spetsifitseerida (nt *kõndima*, *sammuma*, *jooksma* on vaikimisi seotud alati jalgadega: *Ta kõndis küll ettevaatlikult, kuid ühel kivil ikkagi libises ja väänas jala välja*).

Lause tähendus ei piirdu teatavasti selles vahetult öelduga, sellesse kuulub ka järelduv info (kui teadmine, mille lause vastuvõtja saab ja millega arvestab kui kehtivaga). Et meie valdkonnaks on liikumis- ja liigutamissündmused, siis tuleb täpselt tuvastada lauses need entiteedid, mis liiguvad (eri predikaatide puhul on need erinevad) ja fikseerida liikuva entiteedi asukoht pärast liikumissündmust. See ülesanne seostub aga

paraku sündmuses osalevate entiteetide ontoloogiaga (maailmateadmuslike omadustega). Näiteks lausest *Poiss viskas kivi tänavale* järeldub, et kivi on tänaval, kuni ei tule infot, et keegi/miski selle sealt mujale liigutas. Kuid lause *Poiss viskas kivi õhku* puhul selline järeldussituatsioon ei kehti. Niisugune sõnade, lausete, eriti aga teksti maailmateadmusliku või ontoloogilise aspekti arvestamine on vähemalt lausesemantikast alates, aga tekstisemantikas igal juhul vältimatu. Ja ontoloogiatele orienteeritud lähenemise roll keeletehnoloogias kasvab pidevalt, see on teatud mõttes reaalsete rakenduste alus (nt semantilise veebi areng).

6. Pragmaatika

Pragmaatikaalane uurimistöö on suunatud dialoogi modelleerimisele.

Dialoogi modelleerimise oluline eesmärk on võimaldada inimese suhtlust arvutiga loomulikus keeles ja inimestevahelise suhtluse reeglite kohaselt. See nõuab dialoogiteooria loomist ja algoritmide väljatöötamist, mille alusel saaks kasutaja pidada arvutiga loomulikku dialoogi.

Tööd selles valdkonnas algasid Tartu Ülikoolis juba 1980ndate teises pooles (vt nt Koit 1987) ning on kulgenud kahes suunas: teoreetiliste mudelite väljatöötamine ja praktiline eksperimenteerimine arvutil.

Meie nn konversatsiooniagendi mudel käsitleb suhtlust kui protsessi, mida mõjutavad osalejate arvamused, soovid ja kavatsused (Koit, Õim 2003; Koit jt 2009; Koit 2010). Seni oleme keskendunud arutluse ja argumenteerimise modelleerimisele, st osaleja võimalustele argumentide toel partneri arutlust endale soovitavas suunas juhtida. Mudeli proovimiseks on koostatud katseline programm, kuhu aga ei ole praegu lisatud keeletöötlusmooduleid. Edaspidi võiks sellist programmi rakendada näiteks suhtlustreeningul, kus arvuti (ehk konversatsiooniagent) saab inimesega suheldes kehtestada teatavad nõuded argumentide ja vastuargumentide valiku järjekorrale, millest oleks ehk kasu argumenteerimisoskuse arendamisel.

Kui oma teoreetilistes töödes oleme käsitlenud eeskätt mittekooperatiivset suhtlust, kus osalejate eesmärgid on vastandlikud, siis praktiliste rakenduste seisukohast on esmatähtis modelleerida kooperatiivset suhtlust, nagu see leiab aset nt infoandmis- või läbirääkimisdialoogides.

Inimestevahelise või inimese ja arvuti vahelise tegeliku suhtluse uurimise aluseks on dialoogikorpus. Eesti dialoogikorpust hakati Tartu

Ülikoolis koostama aastal 2001. Korpus sisaldab nii suulisi kui ka kirjalikke dialooge. Suuliste dialoogide allikaks on olnud Tartu Ülikooli suulise eesti keele korpus (Hennoste jt 2009). Seisuga märts 2012 on dialoogikorpuses kokku 1182 autentset inimestevahelist dialoogi, suurem osa neist (1137) on telefonikõned, kus inimesed helistavad ametiasutustesse, enamasti info saamiseks (ülejäänud 45 on silmast silma vestlused).

Kõik kirjalikud dialoogid korpuses on inimese ja arvuti vahelised vestlused ja neid on kahesuguseid: ühed (praegu 97 dialoogi) on kogutud nn võlur Ozi meetodil, kus arvuti rolli mängib kasutaja eest varjatult teine inimene (Vutt jt 2002; Pärkson 2011), ja teised (150 dialoogi) on saadud arendatavate (lennu-, teatri-, kino- ja hambaraviinfot andvate) dialoogsüsteemide reaalse kasutuse käigus (Treumuth 2004; Treumuth jt 2006; Treumuth 2008, 2011).

Korpuses on märgendatud dialoogiaktid, mida mõistetakse kui tegevusi, mida inimene keele abil suhtluses teeb (Hennoste, Rääbis 2004: 15). Selleks on Tartu Ülikoolis loodud aktitüpoloogia (Hennoste, Rääbis 2004), mis lähtub vestlusanalüüsi põhimõtetest. Vestlusanalüüsi kohaselt moodustavad mõned dialoogiaktid naabruspaare, kus esiliikme ütlemine teeb relevantseks järelliikme (nt küsimus ootab vastust). Ka arvuti kui dialoogis osaleja peab suutma eristada naabruspaari esiliiget, mis nõuab teatud reaktsiooni, nn üksikaktidest, mis reaktsiooni ei oota (nt vastuvõtuteade). Tüpoloogias on kokku 126 akti (89 naabruspaari- ja 37 üksikakti). Dialoogiaktide märgendamine on seni toimunud käsitsi, kasutades abivahendina programmi, mis võimaldab dialoogiaktide puust valida sobiva aktimärgendi ja paigutada see märgendatavas litereeritud tekstis vajalikku kohta. Programmi kasutamine aitab ühtlasi ära hoida teksti käsitsisisestamisel paratamatult tekkida võivaid vigu.

Dialoogiaktide automaatseks tuvastamiseks on katsetatud mitut statistilist meetodit: tehisnärvivõrgud, otsustuspuud, tõenäosuslikud sufiksipuud ja Bayesi meetod (Fišel, Kikas 2006; Kikas 2007; Fishel 2007; Koit 2011). Seni parim tulemus on saavutatud tõenäosuslike sufiksipuudega (keskmine tuvastamistäpsus 56%). Selle meetodi eeliseks teiste meetodite ees on hea arusaadavus, kuna dialoogiaktid tuvastatakse treeningkorpusest leitud alamsõnede ja -sekventside põhjal. Dialoogiaktide automaatse tuvastamise teeb keeruliseks ühelt poolt aktide suur arv tüpoloogias ja teiselt poolt andmete hõredus – on üle 30 akti, mis esinevad Eesti dialoogikorpuses alla 10 korra. Ka lingvistidest eksperdid ei ole märgendamisel alati üksmeelsed

(seda mõõdab nn κ-koefitsient, mille seni saavutatud parim väärtus on 0,8). Arendamisel on dialoogiaktide poolautomaatse märgendamise tarkvara, mis annab igale lausungile kuni viis kõige tõenäolisemat aktimärgendit, mille hulgast lingvistist kasutaja saab valida sobiva.

Dialoogiaktide automaatset tuvastamist vajab ka dialoogsüsteem, mis peab aru saama kasutaja lausungist (esmajoones on tähtis edukalt tuvastada olulisemaid infoakte – direktiive ja küsimusi). Dialoogsüsteemi arendamiseks oleme analüüsinud korpuses leiduvate infodialoogide ülesehitust, sh telefonikõnede alustamist ja lõpetamist, alamdialoogide kasutamist, sagedamini esinevaid dialoogiakte ja nende väljendamist eesti keeles, samuti võrrelnud inimestevahelistes ning inimese-arvuti imiteeritud (nn võlur Ozi meetodil kogutud) dialoogides rakendatud suhtlusstrateegiaid (vt nt Eskor 2005; Koit jt 2008; Gerassimenko jt 2010; Koit 2010). Eesti dialoogikorpuse analüüsimise hõlbustamiseks on välja töötatud veebis kasutatav (parooliga kaitstud) tarkvara (Treumuth 2004), mis muu hulgas võimaldab otsida korpusest ja loendada mitmesuguseid dialoogides esinevaid nähtusi: sõnu või sõnajärjendeid, transkriptsioonielemente, dialoogiakte, andes ette akti nime, osaleja tähise, suvalise sõna. Samuti saab teha automaatselt morfoloogilist analüüsi, enne eemaldades suuliste dialoogide üleskirjutustest analüsaatorit segavad transkriptsioonimärgid (tööpinki on integreeritud suulise keele analüüsiks kohandatud morfoloogiaanalüsaator Estmorf), määrata alamdialooge (partneri algatatud parandussekventse ja vastuse tingimuste täpsustamise alamsekventse).

Loodud on veebis kasutatav nn asünkroonsete dialoogsüsteemide raamistik (Treumuth 2011), kus on üldistatud varasemate küsimus-vastussüsteemide Reisiagent (Treumuth 2004) ja Teatriagent (Treumuth jt 2006) kogemusi (esimene andis infot Tallinna lennujaamast väljuvate lennukite väljumisaegade kohta, teine aga teatrietenduste kohta Eesti teatrites). Raamistik kujutab endast modulaarset tarkvara uute veebipõhiste dialoogsüsteemide loomise hõlbustamiseks. Eeskätt on arvestatud eestikeelse suhtlusega, kuid enamik mooduleid on püütud teha keelest sõltumatuks, et raamistikku saaks üle kanda ka teistele keeltele. Suhtlusmudeli aluseks on võetud kaks uudset põhimõtet: 1) asünkroonne suhtlus, kus arvuti ei pea passiivselt ootama kasutaja järgmist vooru, vaid võib jätkata info andmist (nii, nagu see toimub inimeste vahel näiteks interneti jututubades); 2) inimabi võimalus: süsteemi administraator saab vajaduse korral sekkuda info andmisse, abistades dialoogsüsteemi raskemate päringute analüüsil ja

neile vastuste leidmisel. Süsteem kasutab regulaarset grammatikat, vastamaks kasutaja küsimustele, ja võtab initsiatiivi, kui kasutaja on passiivne. Kasutaja lausungi mõistmine põhineb selles leiduvatel võtmesõnadel, piiratud ainevaldkonnas annab see rahuldava tulemuse. Süsteem suudab käsitleda ajaväljendeid (nt homme õhtul, kahe nädala pärast, maikuus jne), mis infopäringutes sageli esinevad. Selleks tehakse esmalt lausungi morfoloogiline analüüs, siis leitakse sellest sõnajärjendid, mida võrreldakse andmebaasis paiknevate normaliseeritud ajaväljenditega ja moodustatakse spetsiaalsete reeglite kohaselt vajaliku granulaarsusega (tund, päev, kuu) formaliseeritud ajaväljend (Treumuth 2008). Üheks mooduliks on tekstkõnesüntees, vastuslause moodustamisel kasutatakse valmislauseid või lausešabloone, vajaduse korral ka morfoloogilist sünteesi. Kõnetuvastust ei ole lisatud, st süsteemile saab esitada üksnes kirjalikke küsimusi (päringuakna kaudu), vastuseid annab süsteem kirjutatud tekstina või soovi korral tehiskõnes. Uue dialoogsüsteemi loomiseks raamistiku abil tuleb lihtsalt sisustada teadmusbaas, st koostada ainevaldkonna teadmisi esitavad reeglid (regulaaravaldised). Muud moodulid on universaalsed. Raamistiku abil on praeguseks loodud kaks dialoogsüsteemi, esimene oskab anda eesti keeles infot Tartu kinokavade kohta ja soovitada filme[17], teine annab hambaravialast nõu.

Edasine töö seisneb nii dialoogimudelite teoreetilises uurimises kui ka tulemuste praktilises rakendamises kasutajaga eesti keeles suhtlevate dialoogsüsteemide arendamisel.

7. Masintõlge

Masintõlke (MT) ülesanne on modelleerida tõlkimist loomulike keelte vahel. Formaalsemalt, eesmärk on sisendteksti põhjal genereerida väljundtekst, nii et selle tähendus sihtkeeles oleks sama, mis sisendteksti tähendus lähtekeeles. Sarnane ülesanne on tõlkeabiprogrammidel, mis on mõeldud abiks inimtõlkijatele ning mis oskavad pakkuda fraaside või sõnade erinevaid tõlkevariante, et kiirendada inimesest tõlkija tööd. TÜ-s on aga siiani tegeldud automaatse masintõlkega.

Eestis ja TÜ-s algas MT-alane uurimistöö umbkaudu samal ajal kui ülejäänud maailmas (Koit 2003). 1950ndate lõpus tekkinud uurimisrühm

[17] http://math.ut.ee/~treumuth/.

tegeles Ülo Kaasiku juhtimisel matemaatiliste tekstide tõlkimisega vene keelest eesti keelde. Praktilise arvutirakendusena koostati aga ainult mõned MT alamülesannete programmid, näiteks vene keele morfoloogiline analüsaator.

Järgmiseks poolsajandiks jäi MT ala Eestis puutumatuks. Selle aja jooksul jõudis muu maailma MT-alases töös esialgne lähenemine, mida tänapäeval nimetatakse otsetõlkeks, areneda suuremaks paradigmaks nimega reeglipõhine MT. Kui otsetõlke puhul kasutati enamasti ainult sõnastikku (et sisendi sõnu ükshaaval asendada) ja paari heuristilist reeglit (et nende sõnade järjekorda vajaduse korral muuta), siis reeglipõhine MT kasutab juba lingvistilist (st morfoloogilist, süntaktilist, semantilist jms) analüüsi ning sünteesi. Mõlemat tüüpi MT-d ühendab aga see asjaolu, et tõlkimisprotsessi kirjeldavad inimeksperdid, määrates tõlkereeglite hulga ehk tõlkemudeli.

Kui TÜ-s hakati aastal 2004 taas masintõlkega tegelema, võeti töö aluseks statistilise masintõlke paradigma. See on pärit 1990ndate algusest (Brown jt 1993) ning on tänapäeval kõige levinum lähenemisviis, mida kasutab mh ka Google (Kaalep, Koit 2010). Selle erinevus reeglipõhisest MT-st seisneb selles, et tõlkemudelit ei kirjelda mitte inimeksperdid, vaid see leitakse automaatselt statistilise õppimise printsiibil. Selleks kasutatakse õppimisandmeid ehk suurt hulka tõlkenäiteid (üldjuhul paralleelkorpust) ja tulemuse soravuse saavutamiseks ka sihtkeele näiteid (ükskeelset tekstikorpust). Õppimise (nn treenimise) eesmärk on leida selline tõlkemudel, mis nii tõlke- kui ka sihtkeele näidetega kõige paremini kooskõlas oleks.

Esimeseks ülesandeks oli koguda õppimisandmed ehk treeningkorpus. Lähtudes elektrooniliste tekstide kättesaadavuse lihtsusest, loodi 2004. aastal Eesti ja Euroopa Liidu seadusandlike aktide baasil paralleelkorpus[18], mille kogumaht oli 7,8 miljonit sõna inglise keeles ja 5 miljonit eesti keeles. Praegu esineb eesti keel ühe keelena ka mujal maailmas tehtud suurtes mitmekeelsetes paralleelkorpustes, kusjuures eestikeelsete tekstide maht avaldab muljet: mitmekeelne paralleelkorpus OPUS[19] sisaldab 36,5 miljonit sõna subtiitreid, 9 miljonit sõna EMEA (Euroopa meditsiiniagentuuri) tekste, 3 miljonit sõna KDE (vabatarkvara K Desktop Environment)

[18] http://www.cl.ut.ee/korpused/paralleel.

[19] http://opus.lingfil.uu.se/.

dokumentatsiooni (Tiedemann 2009); paralleelkorpus JRC-Acquis[20] 25,5 miljonit sõna EL-i seadusandlikke tekste (Steinberger jt 2006); Europarl[21] 9,5 miljonit sõna Euroopa Parlamendi sõnavõttude tõlkeid (Koehn 2005).

Õppimisandmete olemasolu võimaldas juba teha tõlkimise katseid; esimesed tulemused on dokumenteeritud 2007. aastal (Fishel jt 2007), tõlkimise suund on eesti keelest inglise keelde. Tõlkeväljundil on kaks peaprobleemi: tõlkimata jäänud või valesti tõlgitud sõnavormid ning väljundsõnade või -fraaside vale järjekord. Esimese probleemi efektiivseimaks lahendusteeks on tõlgitavate lausete valdkonnale võimalikult sarnaste tekstide (ja üldse suurema paralleelkorpuse) kasutamine süsteemi treenimisel. Teist probleemi põhjustab kasutatava MT-meetodi piirang, nimelt saab see palju paremini hakkama üksteise naabruses asuvate ning mitte nii hästi üksteisest kaugemal olevate keelendite ümberjärjestamisega. See tähendab, et MT-süsteemi keelemudel suudab küll tõlkida eestikeelseid fraase inglise keelde, kuid ei suuda sageli toime tulla terve lause fraasijärje muutmisega sihtkeelepäraseks.

Masintõlke katsete kõrval tegeldakse ka õppimisandmete ja nende kvaliteedi uurimisega. Heiki-Jaan Kaalepi ja Kaarel Veskise artikkel (2007) uurib TÜ ja JRC-Acquis' korpuste kvaliteeti nende võrdlemise kaudu. Selle töö edasiarendus (Fishel, Kaalep 2008) pakub meetodit, kuidas paralleelkorpusi kombineerida või üksteisega võrreldes hinnata, arvestades pisierinevuste paratamatut olemasolu korpuste tekstides.

Harri Kirik ja Mark Fishel (Kirik 2008; Kirik, Fishel 2008) uurivad sõnade segmenteerimist eesti-inglise MT kvaliteedi parandamisel. Seejuures tuletavad nad segmenteerimise juhendamata masinõppe abil puhta, st morfoloogiliselt märgendamata korpuse baasil ilma lingvistilisi abivahendeid kasutamata; selleks kasutatakse programmi Morfessor (Creutz, Lagus 2005). Katsete tulemusena on selgunud, et sisendi sõnade segmenteerimine tüveks ja lõpuks ning ka sõnatüvede arvestamine sõnavastavuste leidmisel parandavad tõlkekvaliteeti ning osaliselt lahendavad morfoloogiaga seotud probleeme, eriti paranes tundmatute (ehk õppimisandmetes mitteesinevate) sõnavormide tõlkimine.

Uuringute teoreetilised tulemused on realiseeritud praktilises MT demosüsteemis, mis on kättesaadav internetis[22]. Süsteem pakub korraga

[20] http://optima.jrc.it/Acquis/.

[21] http://www.statmt.org/europarl/.

[22] http://masintolge.ut.ee.

mitu tõlkevarianti, mille aluseks on erinevad tõlkimisalgoritmid. Kasutajal on võimalus ära märkida talle aktsepteeritavana tunduvad väljundid ja pakkuda paremat tõlkevarianti; sellisena täidab interneti MT-süsteem korraga nii demonstreerimise kui ka tagasiside korjamise eesmärki.

Praegu käib töö selles suunas, et arendada ka inglise-eesti ning teiste keelte, nt eesti-prantsuse MT-d, samuti parandada kasutatavate paralleelkorpuste (ka mujal kui TÜ-s tehtute) kvaliteeti.

8. Lõpetuseks

Kokkuvõtvalt võib öelda, et arvutilingvistika ja keeletehnoloogia on eesti keele toena infotehnoloogilises maailmas päris heal järjel, eriti kui juurde arvata ka keeleressursid ja kõnetehnoloogia – kõnesüntees, kõnetuvastus, kõnelejatuvastus, millest siin juttu ei olnud. Nende näitajate poolest on eesti keel väidetavalt vähemalt 50 edenenuma keele hulgas maailma umbes 6000 keele seas (Arengukava 2011: 30).

Aga samas on eri keeletasandite (foneetikast pragmaatikani) küpsusjärk väga erinev ja seda ka tasandite sees. Nt semantika kohta võib öelda, et semantiline andmebaas (Wordnet) on teiste Euroopa keeltega täiesti võrreldaval tasemel, aga lausete-tekstide semantiline analüüs teeb alles esimesi samme (seda küll ka teistes, palju suuremates keeltes).

Teisalt võib osutada erinevustele tasandite või üldisemalt ka rakenduste arengu (õigemini arendamise) iseloomus, mis päris kõnekalt kajastab meie keeletehnoloogia tervikarengu tagamaid ja seoseid mujal maailmas selles valdkonnas toimuvaga. On alasid, nagu morfoloogiline analüüs, mis on olnud maailma mastaabis täiesti arvestataval tasemel ja kõikvõimalike rakenduste loomulik komponent juba pikki aastaid, aga mida on seni pidevalt olnud vaja täiendada ning uuendada, ka lausa põhimõtete poolel.

Ja on alasid, kus tõsine töö on käivitunud suhteliselt hiljuti, aga areng on olnud kiire, et mitte öelda jõuline, on vahetunud kontseptsioonid – näiteks selle kohta sobib ehk enim süntaktiline analüüs: algul lineaarne kitsenduste grammatikal põhinev pindsüntaktiline analüüs, siis hierarhiliste puustruktuuride konstrueerimine, seejärel sõnadevahelisi otsesõltuvusi fikseeriv esitus. Nii on see üldjoontes kulgenud ka mujal maailmas ja arengu pealesurujateks on olnud nii teised keeletasandid (semantika, pragmaatika) kui ka rakendused, mis nt infootsingus või sisukokkuvõtete tegemisel aina

rohkem nõuavad mitte märksõnade või statistika järgi väljanopitud lauseid, vaid fakte, sisulisi andmeid. Rakenduste vallas on kahtlemata samasuguse jõulise arengu läbi teinud masintõlge. Rakendus, millest 10 aastat tagasi keegi ei rääkinudki, on nüüd tasemel, mis on täiesti võrreldav maailmas tipus olevate süsteemidega, ja areng kestab intensiivselt.

Nii et ühtekokku: tööd jätkub veel tükiks ajaks, ehk on isegi õigem öelda, et see ei lõpe kunagi.

Kirjandus

Arengukava 2011 = Eesti keele arengukava 2011–2017. Eesti Keele Sihtasutus. Tallinn.

Asser jt 2004 = Hiie Asser, Heiki-Jaan Kaalep, Siret Linnas, Jaan Mikk, Kadri Muischnek, Merje Songe, Heli Uibo. Õpikute keerukuse analüüs arvutitel. – Toimiv keel II. Töid rakenduslingvistika alalt. Koost. Helle Metslang. Toim. Maria-Maren Sepper, Jane Lepasaar. (= Tallinna Pedagoogikaülikooli eesti filoloogia osakonna toimetised 3.) Tallinn: Tallinna Pedagoogikaülikooli Kirjastus , 72–84.

Baayen, Harald 2001. Word Frequency Distributions. Doedrecht, Boston, London: Kluwer Academic Publishers.

Bick jt 2004 = Eckhard Bick, Heli Uibo, Kaili Müürisep. 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, Dec 10-11, 2004.

Brown jt 1993 = Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. – Computational Linguistics 19 (2), 263–311.

Creutz, Mathias, Krista Lagus 2005. Inducing the morphological lexicon of a natural language from unannotated text. – Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Finland.

Dale jt 2000 = Robert Dale, Herman Moisl, Harold Somers (ed.). Handbook of Natural Language Processing. New York: Marcel Dekker.

EKG II = Mati Erelt, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare 1993. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Peatoim. Mati Erelt, toim. Tiiu Erelt, Henn Saari, Ülle Viks. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Erelt, Tiiu 2007. Terminiõpetus. Tartu: Tartu Ülikooli Kirjastus.

Eskor, Liina 2005. Dialoogiaktid ja suhtlusstrateegiad: Eesti dialoogikorpuse analüüs. – Keel ja Kirjandus 10, 711–727.

Fishel, Mark 2007. Complex taxonomy dialogue act recognition with a Bayesian classifier. – Proceedings: DECALOG'2007 Workshop on the Semantics and Pragmatics of Dialogue. Rovereto, Italy; May 30 – June 1, 2007, 161–162.

Fishel, Mark, Heiki-Jaan Kaalep 2008. Experiments on processing overlapping parallel corpora. – Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco; http://www.lrec-conf.org/proceedings/lrec2008/.

Fishel jt 2007 = Mark Fishel, Heiki-Jaan Kaalep, Kadri Muischnek. Estonian-English statistical machine translation: the first results. – Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. [University of Tartu, 24–26 May 2007] Ed. Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, Mare Koit. Tartu; http://hdl.handle.net/10062/2589.

Fišel, Mark, Taavet Kikas 2006. Dialoogiaktide automaatne tuvastamine. – Keel ja arvuti. Toim. Mare Koit, Renate Pajusalu, Haldur Õim. (= Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6.) Tartu: Tartu Ülikooli Kirjastus, 233–245.

Gerassimenko jt 2010 = Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson. Direktiivsed aktipaarid eestikeelsetes infodialoogides ja nende automaatne tuvastamine. – Eesti Rakenduslingvistika Ühingu aastaraamat 6. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Keele Sihtasutus, 67–86. http://dx.doi.org/10.5128/ERYa6.05.

Good, Irving John 1953. The population frequencies of species and the estimation of population parameters. – Biometrica 40, 237–264. http://dx.doi.org/10.2307/2333344.

Hennoste, Tiit, Andriela Rääbis 2004. Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu: Tartu Ülikooli Kirjastus.

Hennoste jt 2009 = Tiit Hennoste, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson. Suulise eesti keele korpus ja inimese suhtlus arvutiga. – Eesti Rakenduslingvistika Ühingu aastaraamat 5. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper, Reili Argus. Tallinn: Eesti Keele Sihtasutus, 111–130. http://dx.doi.org/10.5128/ERYa5.07.

Kaalep, Heiki-Jaan, Mare Koit 2010. Kuidas masin tõlgib? – Keel ja Kirjandus 10, 726–738.

Kaalep, Heiki-Jaan, Jaan Mikk 2008a. Creating specialised dictionaries for foreign language learners: a case study. – International Journal of Lexicography 21 (4), 369–394. http://dx.doi.org/10.1093/ijl/ecn017.

Kaalep, Heiki-Jaan, Jaan Mikk 2008b. Põhikooli ainesõnastikud. – Keel ja Kirjandus 10, 790–802.

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

Kaalep, Heiki-Jaan, Kadri Muischnek 2006. Multi-word verbs in a flective language: the case of Estonian. – Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context. Trento, Italy, 57–64.

Kaalep, Heiki-Jaan, Kadri Muischnek 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. – Eesti Rakenduslingvistika Ühingu aastaraamat 5. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper, Reili Argus. Tallinn: Eesti Keele Sihtasutus, 157–172. http://dx.doi.org/10.5128/ERYa5.10.

Kaalep, Heiki-Jaan, Tarmo Vaino 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Toim. Tiit Hennoste. (= Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikool, 87–99.

Kaalep, Heiki-Jaan, Kaarel Veskis 2007. Comparing parallel corpora and evaluating their quality. – Proceedings of Machine Translation Summit XI, Copenhagen, Denmark, 275–280.

Kaalep jt 2000 = Heiki-Jaan Kaalep, Kadri Muischnek, Kaili Müürisep, Andriela Rääbis, Külli Habicht. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. – Keel ja Kirjandus 9, 623–633.

Kajaste, Kadri 2009. Eestikeelsete tekstide morfoloogiline ühestamine. Magistritöö. Tartu Ülikooli matemaatika-informaatikateaduskond.

Kaljuvee, Aivi 2008. Määruste ja täiendite eristamine statistiliste meetoditega. Magistritöö. Tartu Ülikooli matemaatika-informaatikateaduskond.

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

Kerner, Kadri 2007. Sõnatähenduste ühestamise tulemuste parandamise meetodeid. Magistritöö. Tartu Ülikooli eesti ja üldkeeleteaduse instituut; http://hdl.handle.net/10062/2929.

Kikas, Taavet 2007. Dialoogiaktide tuvastamine eestikeelsetes dialoogides sufiksipuude abil. Magistritöö. Tartu Ülikooli arvutiteaduse instituut; http://dspace.utlib.ee/dspace/handle/10062/2755.

Kirik, Harri 2008. Juhendamata morfoloogia statistilises masintõlkes. Bakalaureusetöö. Tartu Ülikooli arvutiteaduse instituut.

Kirik, Harri, Mark Fishel 2008. Modelling linguistic phenomena with unsupervised morphology for improving statistical machine translation. – Proceedings of the SLTC'08 Workshop on Unsupervised Methods in NLP, Stockholm, Sweden.

Koehn, Philip 2005. Europarl: a parallel corpus for statistical machine translation. – MT Summit X, Phuket, Thailand, September 13–15, 2005, Conference Proceedings: the tenth Machine Translation Summit; 79–86.

Koit, Mare 1987. Eestikeelse dialoogi modelleerimisest arvutil. – Dialoogi mudelid ja eesti keel. Vastutav toim. Haldur Õim. (= Tartu Riikliku Ülikooli toimetised 795. Töid eesti filoloogia alalt 12.) Tartu, 38–53.

Koit, Mare 2003. Masintõlge ja kus temast kasu on? – Arvutimaailm 4, 51–55.

Koit, Mare 2006. Ratsionalism ja empirism keeleteaduses – vastasseis või koostöö? – Teoreetiline keeleteadus Eestis II. Toim. Ilona Tragel, Haldur Õim. (= Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 7.) Tartu: Tartu Ülikooli kirjastus, 41–54.

Koit, Mare 2010. Eesti dialoogikorpus ja argumenteerimisdialoogi arvutil modelleerimine. – Keel ja Kirjandus 4, 241–262.

Koit, Mare 2011. Automatic recognition of dialogue acts in complex typology. – Proceedings of INISTA: International Symposium on INnovations in Intelligent SysTems and Applications, Istanbul, 15–18 June 2011. Ed. Selim Akyokuş, Adil Alpkoçak, Bülent Bolat, Fırat Doğan, Tülay Yıldırım. Istanbul: IEEE, 485–489. http://dx.doi.org/10.1109/INISTA.2011.5946122.

Koit, Mare, Haldur Õim 2003. Eestikeelse dialoogi modelleerimine. – Keel ja Kirjandus 10, 721–735.

Koit jt 2008 = Mare Koit, Olga Gerassimenko, Riina Kasterpalu, Andriela Rääbis, Krista Strandson. Developing a dialogue system: how to grant a customer's directive? – TSD 2008. Proceedings: Text, Speech and Dialogue. 11th International Conference; Brno, Czech Republic; 8–12 September 2008. Ed. Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. (= Lecture Notes in Computer Science 5246.) Springer, 593– 600.

Koit jt 2009 = Mare Koit, Tiit Roosmaa, Haldur Õim. Knowledge representation for human-machine interaction. – Proceedings of KEOD: International Conference on Knowledge Engineering and Ontology Development, Funchal-Madeira (Portugal), 6–8 October 2009. INSTICC Press, 396–399.

Koskenniemi, Kimmo 1983. Two-level Morphology: A General Computational Model for Wordform Recognition and Production. (= Publications of the Department of General Linguistics, University of Helsinki, 11.)

Lindström, Liina, Kaili Müürisep 2009. Parsing corpus of Estonian dialects. – Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing. May 14, 2009, Odense, Denmark, Nodalida 2009. Ed. Eckhard Bick, Kristin Hagen, Kaili Müürisep, Trond Trosterud. (= NEALT Proceedings Series 8.) Northern European Association For Language Technology, 22–29; http://hdl.handle.net/10062/14288.

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. (= Dissertationes mathematicae Universitatis Tartuensis 22.) Tartu: Tartu Ülikooli Kirjastus.

Müürisep, Kaili, Helen Nigol 2008. Where do parsing errors come from: the case of spoken Estonian. – TSD 2008. Proceedings: Text, Speech and Dialogue. 11th International Conference; Brno, Czech Republic; 8–12 September 2008. Ed. Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. (= Lecture Notes in Computer Science 5246.) Springer, 161–168.

Müürisep jt 2008 = Kaili Müürisep, Heili Orav, Haldur Õim, Kadri Vider, Neeme Kahusk, Piia Taremaa. From syntax trees in Estonian to frame semantics. – Proceedings of the Third Baltic Conference on Human Language Technologies. Kaunas, Lithuania; 4-5. okt. 2007. Ed. František Čermak, Rūta Marcinkevičienė; Erika Rimkutė, Jolanta Zabarskaitė. Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 211–218.

Orav jt 2011 = Heili Orav, Kadri Kerner, Sirli Parm. Eesti Wordneti hetkeseisust. – Keel ja Kirjandus 2, 96–106.

Pruulmann-Vengerfeldt, Jaak 2010. Praktiline lõplikel automaatidel põhinev eesti keele morfoloogiakirjeldus. Magistritöö. Tartu Ülikooli matemaatika-informaatikateaduskond; http://dspace.utlib.ee/dspace/handle/10062/15199

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. (= Dissertationes mathematicae Universitatis Tartuensis 27.) Tartu: Tartu Ülikooli Kirjastus.

Pärkson, Siiri 2011. Võlur Ozi eksperimentide kogumine ja partneri algatatud paranduste analüüs. – Eesti Rakenduslingvistika Ühingu aastaraamat 7. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Rakenduslingvistika Ühing, 197–214. http://dx.doi.org/10.5128/ERYa7.12.

Roosmaa, Tiit Madis Saluveer 1983. Semantic analysis of Estonian texts by computer. – Symposium on grammars of analysis and synthesis and their representation in computational structures. Summaries. Tallinn, 65–67.

Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. (= Eesti NSV Teaduste Akadeemia Emakeele Seltsi Toimetised nr. 12.) Tallinn: Valgus.

Steinberger jt 2006 = Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. – Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 24-26 May 2006.

Tiedemann, Jörg 2009. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. – Recent Advances in Natural Language

Processing V. Ed. Nicolas Nicolov, Galina Angelova, Ruslan Mitkov. (= Current Issues in Linguistic Theory 309.) Amsterdam, Philadelphia: John Benjamins Public Gompany, 237–248.

Tkatchenko, Aleksandr 2010. Named Entity Recognition for the Estonian Language. Master's thesis. University of Tartu, Faculty of Mathematics and Computer Science, Institute of Computer Science; http://www.stacc.ee/files/tkachenko_master_thesis.pdf.

Treumuth, Margus 2004. Eesti dialoogikorpus ja selle töötlemise tarkvara. Magistritöö. Tartu Ülikooli matemaatika-informaatikateaduskond; http://dspace.utlib.ee/dspace/handle/10062/1172.

Treumuth, Margus 2008. Normalization of temporal information in Estonian. – TSD 2008. Proceedings: Text, Speech and Dialogue. 11th International Conference; Brno, Czech Republic; 8–12 September 2008. Ed. Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. (= Lecture Notes in Computer Science 5246.) Springer, 211–218. http://dx.doi.org/10.1007/978-3-540-87391-4_28.

Treumuth, Margus 2011. A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. (= Dissertationes mathematicae Universitatis Tartuensis 72.) Tartu University Press; http://dspace.utlib.ee/dspace/handle/10062/17522.

Treumuth jt 2006 = Margus Treumuth, Tanel Alumäe, Einar Meister. A natural language interface to a theater information database. – Language Technologies, IS-LTC 2006: Proceedings of 5th Slovenian and 1st International Conference. Ed. Tomaž Erjavec, Jerneja Žganec Gros. 9–10 October, Ljubljana, Slovenia, 27–30.

Uibo, Heli 2006. Eesti keele morfoloogia modelleerimisest lõplike muundurite abil. – Keel ja arvuti. (= Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6.) Tartu: Tartu Ülikooli Kirjastus, 13–35.

Vainik, Ene, Kirt, Toomas 2008. Kuidas me mõistame mõisteid? – Eesti Rakenduslingvistika Ühingu aastaraamat 4. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Keele Sihtasutus, 225–245. http://dx.doi.org/10.5128/ERYa4.14.

Veskis, Kaarel, Erkki Liba 2008. Automatic Tagger Evaluation. NLP course assignment report. March 16, 2008; http://lepo.it.da.ut.ee/~hkaalep/arvutimorf_09/kaarel_veskis_and_erkki_liba_report.pdf.

Viks, Ülle 1992. Väike vormisõnastik. I, Sissejuhatus & grammatika. II, Sõnastik & lisad. Toim. Henno Rajandi. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Viks, Ülle 2002. Mis kasu on keeleteadusel keeletehnoloogiast. – Arvutimaailm 8, 11–14.

Vutt jt 2002 = Evelin Vutt, Maret Kullasaar, Mare Koit. Developing a natural language dialogue system: Wizard of Oz studies. – First International

IEEE Symposium Intelligent Systems, proceedings. Varna, Bulgaria, 10–12 September, 2002, 202–207.

Õim, Haldur 2002. Keeletehnoloogia maailmas ja Eestis. – Arvutimaailm 8, 8.

Õim, Haldur 2009. Filoloogi mälestused sellest, kuidas eesti keel ja arvuti Tartus kokku said. – Pool sajandit arvutit Tartu Ülikoolis. Koost. Peep Uba. Toim. Andres Jaeger. Tartu Ülikooli matemaatika-informaatikateaduskond, 87–95.

Õim, Haldur 2011. Kognitiivne pööre. – Humanitaarteaduste metoloogia. Uusi väljavaateid. Koost. ja toim. Marek Tamm. Gigantum Humeris. Tallinn: TLÜ Kirjastus, 111–127.

Õim, Haldur, Madis Saluveer 1985. Frames in linguistic descriptions. – Quaderni di Semantica. An International Journal of Theoretical and Applied Semantics VI (2). Special Issue: Round Table on Frame Semantics I, 295–305.

Õim jt 2009 = Haldur Õim, Heili Orav, Piia Taremaa. Lihtlause semantika: teoreetiline kontseptsioon ja arvutianalüüsi võimalused. – Keel ja Kirjandus 7, 489–504.

Õim jt 2010 = Haldur Õim, Heili Orav, Piia Taremaa. Lausesemantikast üldkeeleteaduse ja eesti keele kontekstis. – Emakeele Seltsi aastaraamat 55 (2009). Peatoim. Mati Erelt. Tallinn: Teaduste Akadeemia Kirjastus, 201–223.

Development of computational linguistics and language technology at the University of Tartu

Kadri Muischnek, Mark Fišel, Heiki-Jaan Kaalep,
Mare Koit, Kaili Müürisep, Heili Orav,
Kadri Vare, Haldur Õim

The article gives an overview of the current research in computational linguistics and language technology at the University of Tartu: research subjects, achievements, and problems. Only research on written language is described, and language resources are mentioned only in brief. In computational morphology, the tools for morphological analysis and synthesis have been robust enough to be used in various practical tasks during the last decade. At present, the subject of most active research is disambiguation. In syntax, a shallow parser has been implemented, and current research is focused on dependency parsing. In semantics, a great deal of effort has been (and is still being) devoted to WordNet and a related task – word sense disambiguation. Recently, tools for named entity classification have been built. Studies in clause semantics focus on spatial movement situations, modelled in frame semantics. In pragmatics, the focus of interest is on modelling dialogues, and in particular, on recognizing dialogue acts (e.g. requests, greetings) as the building blocks of dialogues. In machine translation, the aim is to find language-specific ways to improve statistical machine translation.

Keywords: computational linguistics, language technology, computational morphology, automatic syntactic analysis, automatic semantic analysis, dialogue modelling, machine translation, Estonian language

Kadri Muischnek
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
kadri.muischnek@ut.ee

Mark Fišel
Institute of Computational Linguistics
University of Zurich
Binzmühlestr. 14
8050 Zürich
fishel@ut.ee

Heiki-Jaan Kaalep
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
heiki-jaan.kaalep@ut.ee

Heili Orav
eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
heili.orav@ut.ee

Mare Koit
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
mare.koit@ut.ee

Kadri Vare
eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
kadri.vare@ut.ee

Kaili Müürisep
arvutiteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
kaili.muurisep@ut.ee

Haldur Õim
eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Liivi 2
50409 Tartu
haldur.oim@ut.ee