EESTI KEELE INSTITUUDI TOIMETISED 12

TOIMIV KEEL

I

Töid
rakenduslingvistika alalt

Tallinna Pedagoogikaülikool
Eesti Keele Instituut

Eesti Keele Sihtasutus
Tallinn 2003

Sisu

Keeletehnoloogia Tartu Ülikoolis läbi aegade

Mare Koit, Haldur Õim
Tartu Ülikool

Sissejuhatus

Artikkel käsitleb keeletehnoloogiat Tartu Ülikoolis kahest vaatekohast: esiteks, erialaõpet ja teiseks, teaduslikku uurimistööd. Muidugi on need kaks aspekti teineteisega tihedalt seotud, sest ühelt poolt on stuudiumi üheks eesmärgiks ette valmistada selliseid spetsialiste, kes suudaksid edaspidi osaleda uurimis- ja arendustöös, aga teiselt poolt, kvaliteetne õpetus on võimalik ainult siis, kui õppetööd viivad läbi oma eriala aktiivsed teadlased.

Esmalt vajavad aga täpsustamist mõisted keeletehnoloogia ja arvutilingvistika[1], millest mõlemast edaspidi juttu tuleb. (Nende mõistete üksikasjaliku käsitluse on andnud Ülle Viks (2002).) Arvutilingvistika on hübriidala arvutiteaduse ja keeleteaduse vahepeal, tema eesmärk on töötada välja loomuliku keele automaattöötluseks sobivad keele kirjeldus-, analüüsi- ja sünteesivahendid ning need arvutil realiseerida (Õim 1983). Arvutilingvistika ja keeletehnoloogia on omavahel tihedasti seotud. Keeletehnoloogia on arvutilingvistika tehnoloogiline haru, mis tugineb teadmistele inimkeelest. Keeletehnoloogia tegeleb meetodite, tarkvara ja seadmetega, mis on spetsialiseeritud tekstide ja kõne töötlemiseks (http://www.dfki.de/lt/lt-general.html). Termin "keeletehnoloogia" (ingl *language technology*) tuli laiemalt kasutusele 1990ndate aastate alguses, kui Euroopa Liit käivitas ulatusliku multikultuurilise ja

[1] Varem on kasutatud ka terminit arvutuslingvistika või raallingvistika.

-keelelise Euroopa integreerimise programmi, kus keelebarjääride ületamise põhilise vahendina nähti just loomuliku keele arvutitöötlust (vt nt Danzin jt 1992).

1.
Erialaõpe

Eestis valmistatakse arvutilingviste/keeletehnolooge ette ainult Tartu Ülikoolis. Palju aastaid toimus see põhiliselt individuaalkorras, sh individuaalõppeplaanide alusel, enamasti kas keeleteaduse või matemaatika/informaatika üliõpilaste spetsialiseerumise teel. Niiviisi on omandanud oma hariduse mitmedki nende hulgast, kes Tartu Ülikoolis (ja Eestis) praegu keeletehnoloogiaga tegelevad.

1997./1998. õppeaastal avati Tartu Ülikoolis, täpsemalt filosoofiateaduskonnas eesti ja soome-ugri keeleteaduse osakonnas arvutilingvistika eriala (esialgu küll ainult bakalaureuseõppe tasemel). Parimad selle eriala lõpetajad peaksid olema suutelised alustama uurimistööd arvutilingvistika ja keeletehnoloogia alal, täiendades ennast magistri- ja doktoriõppes.

Õppekava ja mitmed uued kursused töötasime välja kõrghariduse toetusprogrammi HESP (ingl *Higher Education Support Program)* projektide kaasabil. Lääne ülikoolide eeskujul valisime õppekavasse keeleteaduse, informaatika, matemaatika ja arvutilingvistika aineid (de Smedt jt 1999, Köit jt 2002). Kuna eriala on avatud filosoofiateaduskonna koosseisu kuuluvas eesti ja soome-ugri keeleteaduse osakonnas, siis on õppekavas ülekaalus keeleteaduse ained. Õpetamine ise aga toimub kähe teaduskonna koostöös: keeleteaduse aineid õpetab filosoofiateaduskond, matemaatikat ja informaatikat matemaatika-informaatikateaduskond, aga arvutilingvistika ained on jagatud kähe teaduskonna vahel. Koos uute ainete ettevalmistamisega on koostatud ka veebikonspektid.

Erialaõppe maht on 60 ainepunkti[2] (AP) ja õppekava sisaldab 4 plokki: lingvistika 20 AP, informaatika 3 AP, matemaatika 10 AP ja arvutilingvistika 27 AP (sealhulgas arvutilingvistika valikained 4 AP ja bakalaureusetöö 12 AP). Lingvistika plokki kuuluvad näiteks sellised ained nagu keeleteooria, fonoloogia ja morfoloogia, eesti keele lauseõpetus, süntaksimudelid, süntaksiteooriad, tekst ja diskursus, semantika, pragmaatika, lingvistilise kommunikatsiooni teooriad. Matemaatika plokk sisaldab hulgateooria, loogika, algebra, formaalsed keeled ja automaadid ning statistilise analüüsi. Arvutilingvistika plokki kuuluvad sissejuhatus arvutilingvistikasse, keeletehnoloogia, sissejuhatus korpuslingvistikasse, arvutimorfoloogia alused, arvutileksikoloogia, masintõlge, keeletarkvarasüsteemid. Informaatika plokk on praeguses õppekavas väga väikese mahuga, koosnedes vaid ainetest Prolog lingvistidele ja UNIX lingvistidele.

2002./2003. õppeaastast hakkavad Tartu Ülikoolis kehtima uued õppekavad, mille kohaselt 3-aastasele bakalaureuseõppele järgneb 2-aastane magistriõpe, millele omakorda võib järgneda 4-aastane doktoriõpe (mudel 3+2+4; siiani oli mudeliks 4+2+4 aastat). Uute õppekavade väljatöötamise vajaduse tingis Eestis läbiviidav kõrghariduse reform (vt nt http://www.ut.ee/reform). Bakalaureuseõppes omandatakse siis baasharidus, magistriõppes aga konkreetne eriala. Eesti ja soome-ugri keeleteaduse osakonnas saab edaspidi omandada kvalifikatsiooni „eesti ja soome-ugri keeleteaduse magister arvutilingvistika erialal". Lisaks sellele on matemaatika-informaa-Likateaduskonnas informaatika erialal õppides võimalik spetsialiseeruda keeletehnoloogiale. Uutes arvutilingvistika ja keeletehnoloogia õppekavades on palju ühiseid aineid, oluline erinevus on aga alushariduses. Arvutilingvistika puhul on selleks humanitaarteadused ja eriti keeleteadus, keeletehnoloogia puhul aga informaatika ja matemaatika. See toob paratamatult kaasa ka erinevused kvalifikatsioonis, kuid oletatavasti leidub rakendust mõlemat liiki spetsialistidele.

[2] **I ainepunkt** on 40 tundi üliõpilase tööd, sh kuni 50% auditoorset tööd.

1.1. Uus arvutilingvistika õppekava

Filosoofiateaduskonnas kehtima hakkav eesti ja soome-ugri keeleteaduse õppekava näeb ette võimaluse spetsialiseeruda arvutilingvistikale.

Bakalaureuseõppes tuleb sel juhul läbida

- kaks kohustuslikku valdkonna alusmoodulit: humanitaarteadused (16 AP) ja eesti filoloogia (16 AP),
- kohustuslik suunamoodul eesti ja soome-ugri keeleteadus (16 AP),
- kohustuslik erialamoodul arvutilingvistika (20 AP, sh bakalaureusetöö 4 AP),
- kaks valitavat moodulit (kumbki 16 AP), valida võib moodulite eesti keel, eesti keel ja kultuur muukeelsetele, soome keel ja kultuur, soome-ugri keeled, ungari keel ja kultuur, üldkeeleteadus hulgast,
- valik- ja vabaained (20 AP), neid võib valida ükskõik millisest õppekavast.

Arvutilingvistika erialamoodul sisaldab neli ainet: matemaatika arvutilingvistidele I (s.o hulgateooria ja matemaatiline loogika), programmeerimine, andmeanalüüs humanitaarteadustes, keeleteooriad arvutilingvistidele, igaühe maht on 4 AP.

Bakalaureuseõppe lõpetanu omandab kvalifikatsiooni humanitaarteaduste bakalaureus (eesti ja soome-ugri keeleteadus)". Reeglina ei taga selline kvalifikatsioon veel pääsu tööturule (vähemalt mitte arvutilingvistina), vaid järgnema peab magistriõpe. Eeldatavasti jätkab magistriõppes 75% bakalaureuseõppesse astunutest.

Magistriõppekava koosneb erialaõpingutest (56 AP), magistritööst (20 AP) ja vabaainetest (4 AP). Magistriõppesse astumise eelduseks on bakalaureusekraad või sellele vastav haridustase. Arvutilingvistika erialale astumise eeldustingimuseks on arvutilingvistika bakalaureuseõppe erialamooduli läbimine. Seega võivad arvutilingvistika alal magistriõppesse astuda ükskõik millise eriala bakalaureused, kellel on sooritatud need neli ainet, mis moodustavad arvutilingvistika erialamooduli.

Magistriõppes koosnevad arvutilingvistika erialaõpingud omakorda kohustuslikest (22 AP) ja valikainetest (34 AP). Kohustuslikeks aineteks on sissejuhatus arvutilingvistikasse, korpuslingvistika, keeletehnoloogia, matemaatika arvutilingvistidele II (teemad algebra ning formaalsed keeled ja automaadid) ja magistriseminar. Valikainete loend on avatud, seda täiendatakse vastavalt vajadusele ja võimalustele (nt teistest ülikoolidest lektorite-erialaspetsialistide kutsumine).

Praeguses valikainete loetelus on nii keeleteaduse, informaatika kui ka arvutilingvistika aineid.

Keeleteaduse ained on näiteks fonoloogia ja morfoloogia, eesti keele lauseõpetus, semantika, lingvistilise kommunikatsiooni teooriad ja pragmaatika.

Informaatika ainetest kuuluvad loetellu tehisintellekt I ja tehisintellekt II, programmeerimiskeel Perl, andmebaasid; arvutilingvistika ainetest arvutimorfoloogia, arvutileksikoloogia, süntaksianalüsaator jt.

Paljud neist ainetest on samad, mida me seni kehtiva õppekava järgi juba oleme õpetanud ja praegu õpetame, kuid on ka uusi aineid, näiteks loomulike keelte statistilised mudelid ja sissejuhatus kõnetehnoloogiasse.

Magistriõppe lõpetanu omandab kvalifikatsiooni „eesti ja soome-ugri keeleteaduse magister (arvutilingvistika)". Seega on tegu spetsialistiga, kelle arvutilingvistiline haridus põhineb lingvistikale (nagu see on olnud ka siiani, seni kehtinud õppekava järgi õppides).

1.2. Uus võimalus - spetsialiseerumine keeletehnoloogiale

Seoses uute õppekavade loomisega tekkis võimalus hakata Tartu Ülikoolis ette valmistama ka teistsuguse suunitlusega arvutilingviste-informaatikuid-keeletehnolooge. Matemaatika-informaatikateaduskonnas kehtima hakkava uue informaatika õppekava järgi õppides saab valida keeleteaduse ja arvutilingvistika ainetest koosnevaid plokke, mida (filosoofiateaduskonnas käivituva eesti ja soome-ugri keeleteaduse

õppekava arvutilingvistika moodulitest eristamiseks) nimetasime keeletehnoloogia mooduliteks.

Informaatika bakalaureuseõpe annab üldteadmised matemaatika klassikalistest harudest ning baasteadmised arvutite tarkvarast, riistvarast, arvutivõrkudest ja süsteemidest, tehisintellektist, tarkvaratehnikast ja andmeturbest; samuti teatud hulga praktilisi oskusi tööks informaatika valdkonnas, sh programmeerimisoskuse. Võimalik on valida teoreetilise informaatika, tarkvarasüsteemide või keeletehnoloogia suund. Bakalaureuseõppe lõpetanu omandab kvalifikatsiooni täppisteaduste bakalaureus (informaatika)".

Bakalaureuseõppes sisaldab keeletehnoloogia suunamoodul (16 AP) ained keeletehnoloogia, sissejuhatus arvutilingvistikasse, korpuslingvistika, sissejuhatus üldkeeleteadusesse ja andmebaaside teooria.

Informaatika magistriõpe annab põhjalikud teadmised mingist konkreetsest informaatika valdkonnast, mis lubavad teha selles valdkonnas arendustööd; oskuse anda erialaseid konsultatsioone; oskuse töötada meeskonnas ja osaleda projektides. Võimalik on valida teoreetilise informaatika, krüptoloogia või keeletehnoloogia suund. Lõpetanu omandab (olenemata valitud suunast) kvalifikatsiooni ,,informaatika magister". Magistriõppesse astumise eelduseks on bakalaureusetase informaatika (või sellele lähedasel) erialal ja 20 AP ulatuses eeldusaineid (objektorienteeritud programmeerimine, algoritmid ja andmestruktuurid, sissejuhatus matemaatilisse loogikasse, diskreetse matemaatika elemendid, algebra I, andmebaasid).

Kõigile magistrantidele on kohustuslik informaatika didaktika ja magistriseminar (kumbki 4 AP); vabaaineid saab kuulata 4 AP mahus.

Keeletehnoloogiale spetsialiseerujal on magistriõppe kohustuslikud ained tarkvaratehnika, automaadid, keeled ja translaatorid, graafid, andmebaaside teooria, tehisintellekt I, arvutimorfoloogia, süntaksiteooriad ja -mudelid, arvutileksikoloogia, semantika, loomulike keelte statistilised mudelid, kokku 32 AP. Neile lisandub 16 AP valikaineid avatud loete-

lust, mida nii nagu arvutilingvistika analoogilise loetelu puhul täiendatakse vastavalt vajadusele ja võimalustele. Neil I ahel valikainete loetelul on üsna suur ühisosa, kuid päriselt **nad** siiski ei kattu. Kui arvutilingvistika puhul on magistriõp- i II • valikainete moodulis küllalt suur osakaal keeleteaduse aine- I1 I, siis siin on selle asemel enam informaatika aineid, nt looril ise programmeerimise meetod, funktsionaalse programmee- i unise meetod, süsteemide modelleerimine, formaalsed keeled.

Seega on sellise õppekava järgi õppinu informaatik, kes on **aks** kuulanud keeleteaduse ja arvutilingvistika aineid niisugu ises mahus, et omab süstemaatilist ettekujutust loomuli- **ku** keele automaattöötluse ülesannetest ja oskab neid ülesandeid I koostöös keeleteadlastega ka lahendada.

L.3. Mõned probleemid

Arvutilingvistika ja keeletehnoloogia moodulites on palju id aineid, mida hakatakse õpetama koos nii arvutilingvistikale kui ka keeletehnoloogiale spetsialiseerujatele. Ka seni i diiva arvutilingvistika õppekava järgi õpetades on mõne aine MI i Li ja teks üheaegselt olnud nii arvutilingvistika erialale as- luud kui ka mõnd keeleteaduse eriala (nt germaani-romaani **keeli** või eesti keelt võõrkeelena) või koguni informaatikat õp- id üliõpilased. Uute õppekavade puhul on selline kuulaja- i **inev** taust lausa seadustatud. See esitab väljakutse õppe- i' 'Inliidele: kuidas esitada oma ainet nii, et see oleks kõigile mdav ja jõukohane, aga samas ei oleks mõnele liiga liht- **iiiani** oleme seda põhiliselt teinud erineva taustaga kuule erinevate iseseisva töö ülesannete andmise läbi. Näi- I i H istavad informaatika üliõpilased keeletöötlusprogrammi imal ajal keeleteaduse üliõpilased töötavad keelekor- i Kdaspidi muutuvad need probleemid aga üha aktuaalsemaks ia vajavad kompleksset lahendust. Näiteks kuidas i i .•da erineva ettevalmistusega kuulajatest meeskonnad, I id üksteist täiendades ühist ülesannet lahendada.

I IIIH' **probleem** on erikursuste vähene kuulajate arv. Selle Ui im i uks lahendus on veebipõhine õpe, s.t mitte lihtsalt

veebis kättesaadav õppematerjal nagu seni, vaid spetsiaalne kaugkoolituskeskkond, mis integreerib nii loengumaterjali, individuaalsed ülesanded kui ka teadmiste kontrolli. Tartu Ülikoolis on edukalt juurutatud koolituskeskkonda WebCT, millesse on juba paigutatud hulk kursusi (e-ülikooli projekti raames, http://www.ut.ee/e-ylikool/oppejoud/veebipohope.php). Seni ei ole nende kursuste hulgas aga ühtki arvutilingvistika kursust, mis tähendab, et peame hakkama selliseid kursusi oma olemasolevate veebikonspektide baasil ette valmistama.

2. Uurimistöö

Arvutilingvistika õppurid ning arvutilingvistikale/keeletehnoloogiale spetsialiseeruvad keeleteaduse ja informaatika üliõpilased osalevad meie projektides ja ühisseminarides, puutudes niiviisi varakult kokku tegelike ülesannetega. Seetõttu loodame neist saada kvalifitseeritud tööjõudu nii ülikoolidele, teadusasutustele kui ka keeletarkvarafirmadele.

Keeletehnoloogia-alasel uurimistööl on Tartu Ülikoolis juba päris pikk ajalugu.

1950ndatel aastatel tehti Tartu Ülikoolis Ülo Kaasiku juhendamisel matemaatikaüliõpilaste kaasabil masintõlkekatsetusi matemaatika tekstide tõlkimiseks vene keelest eesti keelde. Näiteks koostati tolleaegsele (Tartu Ülikooli arvutuskeskuse esimesele) elektronarvutile Ural-1 vene keele morfoloogilise analüüsi programm.

1960ndate aastate algul alustas Huno Rätsepa juhtimisel eesti keele kateedri juures tööd strukturaallingvistika töörühm, hilisema nimetusega generatiivse grammatika grupp, kuhu kuulusid nii keeleteaduse kui ka matemaatika õppejõud, aspirandid, üliõpilased ja kust said oma esimesed arvutilingvistikaalased teadmised nii mõnedki praegused arvutilingvistid.

1960ndatel aastatel hakati tegelema ka keelestatistikaga (Juhan Tuldava, Ülo Kaasik).

1970ndatel aastatel loodi Tartu Ülikooli kriminoloogialaboris Ilo Sildmäe juhendamisel juriidiliste tekstide automaatne otsisüsteem. Keeleeksperdiks oli selle süsteemi loomisel Haldur Õim. 1980ndatel aastatel alustati Haldur Oimu juhtimisel eesti keele automaattöötluseks arvutiprogrammide loomist. Katsetati eesti keele morfoloogilise ja süntaktilise analüüsi programme. Koostati eksperimentaalne eestikeelsete tekstide mõistmise süsteem, kus teadmised olid esitatud freimidena. Ainevaldkonnaks valiti varavastased kriminaalsed teod. Süsteem töötas tolleaegsel SM-tüüpi miniarvutil.

Praegu on Tartu Ülikool Eesti Keele Instituudi ning Tallinna Tehnikaülikooli foneetika ja kõnetehnoloogia labori kõrval üks kolmest keeletehnoloogia keskusest Eestis. Keeletehnoloogia-alane töö on Tartu Ülikoolis koondunud arvutilingvistika uurimisrühma (http://www.cl.ut.ee), mida võib pidada eespool mainitud töörühma järglaseks. Mitteformaalsesse rühma kuuluvad nii filosoofiateaduskonna koosseisus oleva eesti ja soome-ugri keeleteaduse osakonna üldkeeleteaduse õppetooli kui ka matemaatika-informaatikateaduskonna arvutiteaduse instituudi töötajad ja kraadiõppurid (praegu 14 teadurit ja õppejõudu ning 8 doktoranti). Rühmaga teeb tihedat koostööd *spin-off* firma Filosoft (http://www.filosoft.ee).

Uurimistöös on kesksel kohal eestikeelsete tekstide automaattöötlus ja selle baasiks olevate eesti keele ressursside loomine. Põhilised uurimissuunad on

- eesti keele morfoloogia ja süntaksi formaliseerimine,
- eesti keele semantika formaliseerimine (sh leksikaalsemantilise andmebaasi loomine),
- pragmaatika: eestikeelse (suulise) dialoogi modelleerimine.

Tööstusliku rakenduseni on jõudnud vanemteadur Heiki-Jaan Kaalepi juhtimisel loodud eesti keele morfoloogiaanalüsaator ESTMORF (Kaalep 1996) ja -süntesaator; nende alusel on firma Filosoft koostanud eesti keele õigekirjakontrolle'a ja

poolitaja, mis on lülitatud kontoripaketi MS Office koosseisu. Praegu on käsil ka lõplikel automaatidel põhineva kahetasemelise morfoloogiamudeli (Koskenniemi 1983) rakendamine eesti keelele (Uibo 2002). Selle rakenduse valmimine võimaldab edaspidi hõlpsasti kohandada eesti keelele mitmesugust kommertstarkvara, mis eeldab kahetasemelist morfoloogiat (nt firma Xerox tooted, http://www.xrce.xerox.com).

Eesti keele süntaksi modelleerimisel oleme aluseks võtnud Fred Karlssoni kitsenduste grammatika (Karlsson jt 1995). Süntaksianalüsaator rakendab sisendtekstile morfoloogilise analüüsi moodulit ESTMORF, saadud tulemusele seejärel morfoloogilist ühestajat (Puolakainen 2001) ja lõpuks määrab lauseliikmed, s.t teeb pindmise süntaktilise analüüsi (Müürisep 2000). Eesti keele kitsenduste grammatika sisaldab praegu üle tuhande morfoloogilise ühestamise kitsenduse, paarsada süntaktiliste märgendite lisamise reeglit ja üle tuhande süntaktilise kitsenduse. Kitsenduste grammatikat käsutava süntaksianalüsaatori täpsus on praegu 83% ja vigu tekib alla 3,5%. Need näitajad, küll mitte veel ideaalsed, on juba piisavalt head süntaksianalüüsi eeldavate keelemoodulite testversioonide loomiseks. Siiani on valminud eesti keele nimisõnafraaside tuvastaja (http://www.eki.ee/keeletehnoloogia/projektid/EstNPTool/) ning sisukokkuvõtete genereerija testversioonid (Müürisep 2001). Ees seisab süntaksianalüsaatori baasil grammatikakorrektori loomine.

Eesti keele morfoloogiliseks ühestamiseks on koostatud teinegi programm (Kaalep jt 2000), selle aluseks on Markovi peitmudel. Kitsenduste grammatikal põhineva ühestaja täpsus on 85-90% ja vigade protsent 2, statistiline ühestaja teeb umbes 3% vigu. Kähe ühestaja tulemused pole siiski päriselt võrreldavad, sest morfoloogiliste märgendite hulgad ei kattu täiesti (reeglipõhise ühestaja märgendid on detailsemad, sest süntaksianalüsaatori eelmooduliks olemine nõuab võimalikult rikkaliku info säilitamist). Perspektiivis on morfoloogilisel ühestamisel kähe meetodi — reeglipõhise ja statistilise — ühendamine.

Lause semantilisele analüüsile peab eelnema süvasüntaktiline analüüs, mis lisaks lauseliikmetele määrab ka nendevahelised seosed. Seda ülesannet ei saa täita praegune kitsenduste grammatikal põhinev süntaksianalüsaator. Kavas on katsetada funktsionaalset sõltuvusgrammatikat *(Functional Dependency Grammar),* mis on kitsenduste grammatika edasiarendus ja võimaldab esitada lause süntaktilise struktuuri lauseliikmete vaheliste seoste kaudu (Järvinen jt 1997).

Olulisemaid tulemusi on saavutatud leksikaalses semantikas. 1996. a alustati tööd eesti keele leksikaal-semantilise andmebaasi (Eesti WordNet) koostamisel (Vider jt 2000). Praegu sisaldab see umbes 11 tuhat mõistet (tähendust). Leksikaalse andmebaasi alusel saab edaspidi luua teadmusbaasi, mis on keele mõistmise oluline komponent. On valminud ka (statistikapõhine) semantilise ühestaja katseversioon (http://psych.ut.ee/~kaarel/semyhe/), mis käsutab kõnealust semantilist andmebaasi.

Pragmaatikaalased tööd on keskendunud eestikeelse infoküsimisdialoogi modelleerimisele, käsutades eesti suulise kõne korpuse materjali. Tiit Hennoste juhtimisel on välja töötatud dialoogiaktide süsteem, mida käsutatakse loodava dialoogikorpuse märgendamisel (Hennoste jt 2002). Eesti suulise kõne korpuse maht on praegu 300 000 litereeritud sõna, märgendatud dialoogikorpuse maht 6000 sõna. On tegeldud ka loomuliku dialoogi modelleerimise teoreetiliste probleemidega (Köit, Õim 2000).

Aastatel 1991-1995 loodi arvutilingvistika uurimisrühmas **eesti** kirjakeele baaskorpus, mis sisaldab eestikeelseid tekste (ilu- ja ajakirjandust, teadus- ja populaarteadustekste, esseid ja biograafiaid, hobi- ja harrastustekste, propagandatekste, entsüklopeedilisi tekste, dokumente ja vaimulikke tekste) aastatest 1983—1987 kogumahuga 1 mln sõna(vormi), iga teksti maht on umbes 2000 sõna. Baaskorpuse loomisel võeti eeskujuks **Browni** ja **LOBi** korpused (Johansson jt 1978). Korpuses **on** märgendatud lõigud, laused, lühendid, pärisnimed jms (I **lennoste** jt 1998), käsutades märgendussüsteemi TEI *(Text*

Encoding Initiative, http://etext.virginia.edu/TEI.html). Lisaks on loodud kümnendite kaupa liigendatud korpused aastatest 1890-1990, igaüks sisaldab umbes 500 000 sõna vastava perioodi ilu- ja ajakirjandust. Olemasolev kasutajaliides võimaldab otsida eesti kirjakeele korpusest sõnade grammatiliste vormide jm märgijadade konkordantse. Loomisel on nn segakorpus, mille kavandatav maht on vähemalt 100 mln sõna. Praegu kuuluvad korpusesse Riigikogu stenogrammid (13 mln sõna), ajalehed Postimees ja Eesti Ekspress (vastavalt 4,4 mln ja 5,5 mln sõna).

Korpust võib vabalt käsutada, aga ainult mitteärilistel eesmärkidel.

Morfoloogilise ühestaja ja süntaksianalüsaatori treenimiseks ja testimiseks on loodud morfoloogiliselt ja süntaktiliselt märgendatud korpused, praeguse mahuga vastavalt 200 000 ja 80 000 sõna. Märgendamine jätkub.

Täiendamisel on püsiühendite andmebaas (Kaalep jt 2002), kuhu kuuluvad ühend- ja väljendverbid, fraseologismid, populaarsed tsitaadid, mitmesõnalised terminid või lihtsalt sageli koos käsutatavad sõnad. Andmebaasi praegune maht on üle 20 000 kirje (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html).

TÜ arvutilingvistid on osalenud või osalevad mitmes rahvusvahelises projektis, sh Euroopa Liidu projektid GLOSSER, MULTEXT-EAST, TELRI-I, TELRI-II, CONCEDE, EuroWordNet, ning täitnud ja täidavad mitmeid Eesti Teadusfondi, Eesti Informaatikakeskuse ja Haridusministeeriumi sihtprogrammi Eesti keel ja rahvuskultuur keeletehnoloogia projekte.

Lähemate aastate plaanis on edasi arendada keeletarkvara, sh morfoloogilist ja semantilist ühestajat ning süntaksianalüsaatorit ja -süntesaatorit, uurida ja modelleerida dialoogide formaalset struktuuri, täiendada märgendatud korpusi. Kõiki neid tulemusi saab käsutada paljudes keeletehnoloogilistes rakendustes, alates kirjutaja abivahenditest (nt grammatika- ja stiilikorrektor tekstitoimetis) kuni masintõlkeni või eestikeelse dialoogini arvutiga.

Viited

Danzin, A. and the Strategic Planning Study Group 1992. Towards a European Language Infrastructure. Report to the Commission of the European Communities. 31 March.

Hennoste, Tiit; Mare Köit; Maret Kullasaar; Andriela Rääbis; Evely Vutt 2002. Eesti dialoogikorpuse loomise probleemid. - Tähendusepüüdja/Catcher of the Meaning. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim Tiit Hennoste ja Renate Pajusalu. Lk 143-160.

Hennoste, Tiit; Mare Köit; Tiit Roosmaa; Madis Saluveer 1998. Structure and Usage of the Tartu University Corpus of Written Estonian. - International Journal of Corpus Linguistics. Amsterdam: John Benjamins Publishing Co. Vol 3(2). Lk 99-114.

.lohansson, S.; Leech, G.; Goodluck, H. 1978. Manual of information to accompany the Lancaster—Oslo/Bergen corpus of British English, for use with digital computers. - Oslo. Manuscript.

Järvinen, Timo; Pasi Tapanainen 1997. A Dependency Parser for English. — Technical Reports, No. TR-1. Department of General Linguistics, University of Helsinki.

Kaalep, Heiki-Jaan 1996. ESTMORF. AMorphological Analyzer for Estonian. - Estonian in the Changing World. University of Tartu. Lk 43-98.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Püsiühendite leidmine teksti abil. - Tähendusepüüdja/Catcher of the Meaning. Tartu: TÜ üldkeeleteaduse õppetooli toimetised 3. Toim Tiit Hennoste ja Renate Pajusalu. Lk 172-184.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. — Arvutuslingvistikalt inimesele. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim Tiit Hennoste. Lk 87-99.

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

Köit, Mare; Roosmaa, Tiit; Õim, Haldur 2002. Teaching Computational Linguistics at the University of Tartu: Experience, Perspectives and Challenges. - Effective Tools and methodologies for teaching NLP and CL. Proceedings of the Workshop. 7 July 2002, University of Pennsylvania, Philadelphia, PA, USA. Published by the Association for Computational Linguistics. Lk 84-89.

Köit, Mare; Roosmaa, Tiit; Õim, Haldur 1996. Teaching computational linguistics: one vision. - Estonian in the Changing World. Ed. Tiit Hennoste. University of Tartu. Lk 115-122.

Köit, Mare; Oim, Haldur 2000. Developing a model of natural dialogue. - From spoken dialogue to full natural interactive dialogue-theory, Empirical analysis and evaluation. LREC 2000 Workshop proceedings. Ed. L. Dybkjser. Athen. Lk 18-21.

Koskenniemi, Kimmo 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. — University of Helsinki, Dept of General Linguistics. Publications No. 11. Helsinki.

Müürisep, Kaili 2001. Parsing Estonian with Constraint Grammar. Online proceedings of NODALIDA'01. Uppsala, http://stp.ling.uu.se/nodalidaOl/pdf/myyrisep.pdf

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.

de Smedt, Konraad; Gardiner, Hazel; Ore, Espen; Orlandi, Tito; Short, Harold; Souillot, Jacques; Vaughan, William (eds) 1999. Computing in Humanities Education. A European Perspective. — SOCRATES/ERASMUS Thematic Network Project on Advanced Computing in the Humanities. University of Bergen. 242 pp.

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

Viks, Ülle 2002. Mis käsu on keeleteadusel keeletehnoloogiast? - Arvutimaailm, nr 2, lk 11-14. Tallinn: Eesti Informaatikakeskus.

Vider, Kadri; Kahusk, Neeme; Orav, Heili; Õim, Haldur; Paldre, Leho. Eesti keele tesaurus. - Arvutuslingvistikalt inimesele. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim Tiit Hennoste. Lk 127-152.

Õim, Haldur 1983. Inimene, keel ja arvuti ehk kompuuterlingvistika. Tallinn: Valgus (Mosaiik 33).