

EESTI KEELE KEELETEHNOLOOGILINE TUGI (2006–2010)

19.–21. novembril 2007 toimus Tallinnas TTÜ Küberneetika Instituudis konverents, mille teemaks oli riiklik programm „Eesti keele keeletehnoloogiline tugi (2006–2010)”. Konverentsi korraldasid ülalmainitud programmi (lühidalt EKKTT) juhtkomitee, EKKTT koordinaator, keeleteaduse ja -tehnoloogia doktorikool ning TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia labor.

Riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006–2010)” peaeesmärgiks on eesti keele tehnoloogilise toe arendamine tasemele, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises keskkonnas võrdväärselt suurema kõnelejakonnaga keeltega (vt lisaks: <http://keele-technologia.cs.ut.ee>).

EKKTT konverentsil oli kolm põhi-eesmärki: esiteks, käimasoleva uurimis- ja arendustöö tutvustamine laiemale avalikkusele, teiseks, programmi seotud inimestele parema ülevaate andmine erinevate projektide käigust ja juba saavutatud tulemustest ning kolmandaks, erinevate uurimisrühmade koostöö soodustamine. Lisaks pakuti 2008. aastal esitatavate projektide uurimisrühmadele võimalust oma plaane tutvustada ja tagasisidet saada.

Kutsutud külalistena esinesid Peter Wittenburg (Nijmegeni Max Plancki Psühholingvistika Instituut) ja Markku Mäenpää (Eesti Arengufondi Investeeringute Divisjon). P. Wittenburg tutvustas 2008. aasta alguses käivituvat üleeuroopalist keeleressursside ja tehnoloogia infrastruktuuri projekti, M. Mäenpää käsitles ettevõtte käivitamise protsessi riskikapitali seisukohast ning pakkus näpunäiteid tulevastele ettevõtjatele.

Konverentsi esimene päev oli pühendatud korpustele. Pärast konverentsi avanud haridus- ja teadusministri Tõnis

Lukase sõnavõttu, Jaak Vilo (TU) ülevaateettekannet ja P. Wittenburgi esinemist astusid üksteise järel üles riikliku projekti raames korpustele tööd tegevate uurimisrühmade esindajad.

Tiit Hennoste (TÜ) pidas ettekande „Eesti kõnekeele korpuse kogumine ja translitereerimine”. Uurimisrühm tegeleb autentsetes situatsioonides toimunud vestluste ja monoloogide kogumise ja translitereerimisega. Sellise korpuse põhjal on võimalik uurida kõnekeele morfoloogiat, süntaksit, leksikat, semantikat ja pragmaatikat, mis teadupärast erinevad oluliselt kirjakeele omadest. Keeletehnoloogia seisukohast on kõnekeele korpus vajalik inimese ja arvuti võimalikult inimsuhtlusalheda-seks modelleerimiseks. Praeguseks on translitereeritud 1777 teksti, mille hulgas on nii vahetuid vestlusi, telefoni-vestlusi kui ka tele- ja raadiosaateid.

Pire Teras (TÜ) esindas projekti „Eesti keele spontaanse kõne foneetiline korpus”. Kuigi ka sel juhul on tegu eestikeelse kõne korpusega, on eesmärk erinev. Korpus on vajalik eesti keele häälduse põhiparameetrite analüüsimisel ning eesti keele kõnesünteesi ja kõnetuvastuse ülesannete täitmisel. Korpusesse salvestatakse argivestlusi, ametlikke dialooge ja monolooge. Selline foneetiline märgendamine on töö- ja ajamahukas. Praeguseks on lindistatud peaaegu 20 tundi kõnet, millest on märgendatud natuke üle veerandi.

Hille Pajupuu (EKI) pidas ettekande eesti emotsionaalse kõne korpuse loomisest. Korpus sisaldab ette loetud lauseid, mis jaotatakse neis peituva emotsiooni järgi viha-, rõõmu-, kurbuse- ja neutraalseteks lauseteks. Emotsioonide sisaldumist kontrollitakse tajukasetega. Korpuse põhieesmärgid on olla ühelt poolt korpuspõhise emotsionaalse tekst-kõne-sünteesi akustiline baas ja

teiselt poolt usaldusväärne andmekogu kõnes avalduvate emotsioonide uurimiseks. Kuna ilma emotsioonideta on kõne ebaloomulik ning oluliselt informatsioonivaesem, siis on emotsionaalse kõne korpus tänapäevase (korpuspõhise) tekst-kõne-sünteesi jaoks hädavajalik.

Külli Habicht (TÜ) tutvustas projekti „Eesti vanema kirjakeele elektroonilised kogud”. Selle projekti raames kogutakse eestikeelseid kirjalikke tekste, mis on ilmunud kuni XIX sajandi esimese pooleni. Loodav tekstikorpus tagab nii keeleuurijatele kui ka -huvilistele ligipääsu tekstidele, mida muidu säilitatakse muuseumides ja arhiivides. Selle korpuse abil on võimalik uurida eesti kirjakeele kujunemist läbi aegade ja vaadelda kirjalikus keeles toimunud muutusi, mille tundmine aitab mõtestada ka tänapäeva keeles toimuvaid protsesse.

Oma ettekandes „Eesti keele koondkorpus” andis Kadri Muischnek (TÜ) ülevaate kirjaliku eesti keele suuremahulisel kogumisel seni tehtud tööst ja tulevikuplaanidest. Praeguseks on kogutud 90 miljoni sõna mahus teksti, kuid plaanis on see korpus laiendada kuni 200 miljoni sõnani. Korpus sisaldab ajakirjanduse, aga ka ilukirjanduse, teaduse ja seaduste tekste, lisaks on näiteid Interneti-jututubadest. Sellist korpust on tarvis nii tänapäevaste statistiliste keeletehnoloogia süsteemide treenimiseks kui ka reeglipõhiste süsteemide testimiseks. Peale keeletehnoloogide kasutatavad seda ka keeleuurijad.

Heili Orav (TÜ) tutvustas ettekandes „TÜ eesti keele tesaauruse (eesti *word-net*) täiendamine” eestikeelse leksikaal-semantilise andmebaasi laiendamiseks ja täiendamiseks tehtavat tööd. Selline andmebaas sisaldab peale mõistete seletuste informatsiooni ka sõnadevaheliste seoste ning nende ingliskeelsete vastete kohta. Projekti lõpuks plaanitakse andmebaasi täiendada vähemalt 25 000 uue mõisteni. Andme-

baasi saab kasutada keeletehnoloogilisteks rakendusteks, nagu sisupõhine infootsing, automaatne sisukokkuvõte, masintõlge ja keeleõpe.

Heiki-Jaan Kaalepi (TÜ) ettekanne „Korpusepäring keeleveebis” andis ülevaate seni tehtust projektis, millel on mitu eesmärki: tekstikorpuste mugava kasutamise võimaldamine Interneti kaudu, veebis kättesaadavate erialasõnastike arvu suurendamine ja sõnastike ning korpusepäringute omavaheline sidumine. Seni on kõige suurem töö toimunud keeleveebis (www.keeleveeb.ee), milles on tehtud kättesaadavaks 25 erialasõnastikku ja loodud ühispäringu võimalus, kus kasutaja saab otsida sõna või terminit kuni 50 sõnastikust korraga.

Esimese konverentsipäeva lõpetuseks rääkis Einar Meister (TTÜ Küberneetika Instituut) projektist „Kõnekeele ressursid ja kõnetehnoloogia andmebaasid”, mille eesmärgiks on eesti keele foneetilisteks ja kõnetehnoloogilisteks uuringuteks ning arendustöödeks vajalike kõnekorpuste salvestamine, digitaliseerimine ja märgendamine, ühtsete kõnesignaali segmenteerimise ja märgendamise reeglite väljatöötamine ning erinevate andmebaaside haldamiseks ja efektiivselt kasutamiseks ühtse tehnoloogilise keskkonna loomine. Seni on valminud nüüdisaegne kõnesalvestusstudio, on kogutud üle 300 tunni Eesti Raadio lühiauudiste salvestusi ja digitaliseeritud üle 8000 lehekülje uudistekste. Aktsendiga kõne salvestamiseks on koostatud eestikeelne tekstikorpus ja spontaanse kõne tuvastuseks vajalike mudelite arendamiseks on alustatud loengute salvestamist.

Konverentsi teise, kirjaliku keele tehnoloogiale pühendatud päeva avatekkande pidas H.-J. Kaalep masintõlke teemal. Projekt keskendub statistilisele masintõlkele ja eelkõige eesti-inglise tõlkimissuunale. Seni on katsetatud olemasolevate ressurssidega,

parandatud erinevate paralleelkorpusete omavahelise kombineerimise kaudu nende kvaliteeti ja ekstraheeritud automaatselt masintõlke jaoks vajalikke väljendeid ja muutevorme sisaldav leksikon. Katsetatud on olemasoleva süsteemi Moses abil eestikeelsete tekstide inglise keelde tõlkimist. Edaspidi on plaanis uuendada korpusi, jätkata masintõlke katseid erinevate korpusete baasil ja eksperimenteerida eri faktoreid arvestavate fraasipõhiste masintõlke mudelitega.

Kaili Müürisep (TÜ) tutvustas projekti „Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid”. Plaanis on luua süntaktilist analüüsi kasutatavate rakenduste prototüübid: grammatikakorrektor ja süntaksianalüüsil põhinev automaatsete sisukokkuvõtete tegija. Nii süntaksianalüüsaatori täiendamiseks kui ka eespool mainitud rakenduste loomiseks on vaja süntaktiliselt märgendatud tekstikorpust. Projekti raames on juba loodud ligikaudu poole miljoni sõnaline süntaktiliselt märgendatud korpus ning 370 lauset sisaldav puupank, tööd on alustatud komavigade korpuse loomise ja komavigade automaatse tuvastamise reeglite kallal.

Haldur Õim (TÜ) pidas ettekande teemal „Lihtlause semantiline analüüs”. Selle projekti eesmärgiks on luua eesti keele automaatse semantilise analüüsi programm, mille sisendiks on süntaksianalüüsaatori väljund ja semantiline andmebaas. Valdkonnaks valiti liikumisverbidega lihtlausead, semantiliseks esitusviisiks freimisemantika. Seni on poolkäsitsi semantiliselt analüüsitud 189 lauset nn Rätsepa korpusest, mis sisaldab 370 lihtlausead Huno Rätsepa tööst ”Eesti keele lihtlauseade tüübid” (1974).

Jaak Pruulmann-Vengerfeldt (TÜ) rääkis projektist „Reegli põhinev keeletarkvara”. Projekti eesmärgiks on Eesti Keele Instituudis varem väljatöötatud

morfoloogiataarkvara ja -reeglite uuendamine ja taaskasutatavaks muutmine. Plaanis on luua mugavad ja kiired vahendid reeglite täiendamiseks ja teendamiseks, et neid saaks kasutada muudes morfoanalüüsi või -sünteesi vajavates rakendustes. Esimese praktilise rakendusena on plaanis luua morfoanalüüsil põhinev vabavaraline speller.

Ülle Viks, Andres Loopmann ja Indrek Hein (EKI) tegid ettekande projektist „Leksikograafi töökeskkond”, mis jätkab kolme varasema projekti tööd. Eesmärkideks on luua leksikograafidele sobiv ja mugav veebipõhine töökeskkond ehk sõnastike haldussüsteem (EELex), koostada Eesti-X-keele sõnastik (EXS) ehk eesti lähtekeele andmebaas mis tahes kakskeelse sõnastiku koostamise tarbeks ning lõpuks anda projekti tulemused (sõnastike veebiversioonid ja sõnastike haldussüsteemi laiatarbeversioon) avalikku kasutusse. Praeguseks on loodud leksikograafi töökeskkonna tuum EELex. Koostatud on EXS-i esialgne versioon ja loodud ÕS-i avalik veebiversioon.

Jaak Vilo rääkis projektist „Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid”, mille eesmärk on muuta Interneti andmebaaside päringusüsteemid kasutajasõbralikumaks. Ennekõike tähendab see näiteks võimalust otsida tekstist sõnu, mis pole täpselt kirjutatud. Otsisõnad võivad sisaldada kirjavigu, häälduse järgi üleskirjutust, vana või uut kirjaviisi või teist tähestikku. Näiteks saab otsida venekeelseid sõnu eestikeelse kõlapildi järgse ülestähenduse järgi. On loodud kahe erineva ligikaudse otsimise algoritmi esimesed prototüübid. Töö käib nende edasiarendamise ja rakendamise suunas.

Maarika Terali ja Ingrid Rummo (TÜ) ettekande teemaks oli „Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid”. Projekti eesmärgiks on veebipõhise õpikeskkonna „Eesti

keele ja kultuuri kursused” infrastruktuuri täiustamine, uue materjali lisamine ning eesti keele õppijate keelevigade korpuse kogumine ja märgendamine. Vigade korpuses on paralleelselt esitatud veaga laused ja neile vastavad korrigeeritud laused. Vigade korpust saab tulevikus kasutada keeleõppe-materjalide või elektroonilise grammatikakorrektori loomisel, aga ka uurimistöös.

Oma selle konverentsi kolmandas ettekandes rääkis H.-J. Kaalep projektist „Mitmesõnaliste verbide ja nende kokku ja lahku kirjutamise vigade äratundmine eestikeelsetes tekstides”. Projekti eesmärk on leida seaduspärasused, mille põhjal automaatselt ära tunda eestikeelses tekstis esinevaid verbikeskseid püsiühendeid: ühend- ja väljendverbe ning tugiverbiühendeid. Praeguseks on loodud 400 000-sõnaline korpus, kus on märgendatud verbikesksed püsiühendid. Korpusest pärit andmetega on täiendatud varem olemas olnud mitmesõnaliste verbide andmebaasi. Loodud on mitmesõnaliste verbide märgendaja esialgne versioon.

Pille Eslon (TLÜ) rääkis teemal „Eesti vahekeele korpuse hetkeseis ja arengusuunad”, millega seoses on plaanis esitada projekt riiklikule programmile 2008. aastal. Vahekeele ehk õppijakeelekorpus on rakendatav nii uurimistöös kui ka võõrkeeleõppe eesmärkidel. Edasise tööga soovitakse liikuda kahes suunas: suurendada ressursi ja kasutada seda keeletehnoloogiliste rakenduste loomiseks.

Konverentsi kolmandat, suulise keele tehnoloogiale pühendatud päeva alustas E. Meister ettekandega „Kõne analüüs ja variatiivsuse mudelid”. Selle projekti eesmärgiks on kõne akustilise/foneetilise analüüsi meetodite uurimine ja arendamine ning erinevate kõnevariatsioonide foneetiliste kirjelduste ning kõnetehnoloogilisteks rakendusteks sobivate mudelite loomine.

Seni on uuritud eesti vokaalide omakestuse ning põhitooni variatsiooni ning kestuse mõju vokaali kvaliteedi tajumisele. Analüüsitud on ka loengukõne temporaalset struktuuri ja vene aktsendi akustilisi parameetreid ning aktsendi tajut. Tööd on tehtud ka kõnesignaali kvaliteedi automaatse hindamise vallas.

Meelis Mihkla (EKI) ettekanne oli teemal „Eestikeelne korpuspõhine kõnesüntees”. Projekti eesmärgiks on luua kõrgekvaliteediline tekst-kõnesüntesaator, mis teisendaks eestikeelse sisendteksti loomuliku kõlaga kõneks. Sellist süsteemi saab rakendada inimese-arvuti-inimese suulise kõne suhtlussüsteemide osana ja pimedate ning vaegnägijate arvutikeskkonnas suhtlemiseks ning teabe vahendamiseks. On koostatud kõiki eesti keele difoone, sagedasemaid sõnu ja fraase, enamiku liitmorfeeme ja palju numbreid sisaldav tekstikorpus, mis salvestati professionaalsete raadiodiktorite esituses. Korpus märgendati erinevatel fonoloogilistel tasanditel. Palju tööd on tehtud ühikute valikualgoritmide väljatöötamise ja kõneprosoodia statistilise modelleerimise alal.

Tanel Alumäe (TTÜ Küberneetika Instituut) rääkis eestikeelsete kõnetuvastuse meetodite uurimisest ja arendamisest. Projekti eesmärk ka erinevate tuvastussüsteemi prototüüpide loomine. Senine töö on keskendunud eesti keele statistilise kõnetuvastusmeetodi loomisele ja täiustamisele. Eri-ist tähelepanu on pööratud morfeemide kui tuvastusühikute kasutamisele eesti keele statistilise tuvastusmudeli puhul, keelemudeli temaatilise adapteerimise võimalustele ja eesti keele vältevastanduse modelleerimisele.

Mare Koit (TÜ) pidas ettekande „Eestikeelne infodialoog arvutiga”. Samanimelise projekti eesmärk on töötada välja tarkvara, mis võimaldaks küsimuse-vastuse dialoogi arvutiga inim-

suhtluse reeglite kohaselt. Dialoogitarkvara arendamiseks on loodud dialoogikorpused, mis sisaldavad nii suulisi kui ka kirjalikke, nii inimeste kui ka inimeste ja arvuti dialooge. Valdav osa kogutud materjalist on juba dialoogiaktidega märgendatud. On loodud mitmesugust tarkvara: veebis kasutatav dialoogikorpuse tööpink, dialoogiaktide automaatse tuvastamise meetodite tes-

timise tarkvara ja dialoogiaktide poolautomaatse märgendamise tarkvara algversioon.

Konverents lõppes Einar Meistri juhitud diskussiooniga riikliku programmi tuleviku seisukohalt olulistel teemadel.

MAARIKA TRAAAT