

RIIKLIKU PROGRAMMI „EESTI KEELE KEELETEHNOLOOGILINE TUGI (2006–2010)” TEINE KONVERENTS

6.–7. aprillil toimus Tartus riikliku programmi „Eesti keele keeletehnoloogiline tugi (EKKTT) 2006–2010” teine aruandekonverents.

EKKTT programm käivitus 2006. aastal ja selle eesmärgiks on eesti keele keeletehnoloogilise toe arendamine tasemele, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises keskkonnas. Programmi raames rahastatakse keeletehnoloogia-alast teadus- ja arendustegevust ressursside kogumisest kuni keeletehnoloogiliste rakenduste (prototüüpide) loomiseni (vt www.keeletehnoloogia.ee).

EKKTT programm rahastab 23 projekti (sh 15 jätkuprojekti ja 8 uut projekti, mis algasid käesoleval aastal). 2008. aastaga lõppes kaheksa projekti. Konverentsil andsid oma tööst ja tulemustest ülevaate kõigi programmi raames rahastatavate projektide esindajad. Kokku peeti kahe päeva jooksul 25 projekte tutvustavat ettekannet ja kaks plenaarettekannet.

Konverentsi avasid lühisõnavõttudega Haridus- ja Teadusministeeriumi teadusosakonna juhataja Indrek Reimand ja EKKTT programmi juhtkomitee esimees Jaak Vilo (TÜ). Esimese konverentsipäeva ettekanded puudutasid eelkõige keeleressursside ja kõnekeelsete korpusete alast tööd.

Kutsutud esineja Tanel Alumäe (TTÜ Küberneetika Instituut) andis ülevaate Prantsusmaal ja Eestis kõnetuvastuse alal tehtavast, tuginedes oma töökogemusele Pariisis Prantsusmaa Teadusministeeriumi rahastatavas CNRS-i (*Centre National de la Recherche Scientifique*) arvutiteaduste laboris ning EKKTT projekti „Eesti keelse kõnetuvastuse meetodite uurimine ja arendamine” juhina. Eestikeel-

se kõnetuvastuse valdkonnas on seni loodud kaks eesti keele tuvastustehnoloogiat kasutavat rakendusprototüüpi: autosegmenteerija ning häälega juhitud kalkulaator.

Tiit Roosmaa (TÜ) tutvustas 2008. aastal alanud projekti „Eesti keeleressursside keskus”, mille eesmärgiks on luua eesti keele alast tarkvara ja keeletehnoloogilisi ressursse koondav keskus. Selle ülesandeks oleks nii keeletehnoloogiliste standardite väljatöötamine ja fikseerimine kui ka keeleressursside kasutamiseks vajalike juriidiliste lepingute ja litsentside koostamine selleks, et Eestis olemasolev keeleressurss jõuaks kõigi võimalike huvilisteni. Selle eesmärgi nimel on Tartu Ülikool Eesti esindajana liitunud CLARIN-i (*Common Language Resources and Technology Infrastructure*) võrgustikuga, mille kaudu on võimalik kasutada vastavat üleeuroopalist kogemust.

Einar Meister (TTÜ Küberneetika Instituut) andis oma ettekandes ülevaate projekti „Kõnekeele ressursid ja kõnetehnoloogia andmebaasid” raames kogutavast kolmest kõnekorpusest (uudiste korpus, aktsendikorpus ja loengukõne korpus) ning nende kasutamiseks ja haldamiseks mõeldud infrastruktuurist.

Tiit Hennoste (TÜ) võttis kokku 2008. aastal lõppenud projekti „Eesti kõnekeele korpus (TÜKK) kogumine ja translitereerimine” tulemused ning tutvustas äsja alanud projekti „Tartu ülikooli eesti kõnekeele audio- ja videokorpuse kogumine ja otsingutarkvara loomine” eesmäärke. Neid kahte projekti seob spontaansete dialoogide ja monoloogide pragmaatilis-suhtlusliku korpusete kogumine, mille tulemusi on võimalik rakendada arvutiga peetavas

eestikeelses infodialoogis. Uue projekti raames on plaanis alustada ka videokorpuse kogumist mitteverbaalse suhtluse uurimiseks.

Projekti „Eesti keele spontaanse kõne foneetiline korpus” seisust tegi kokkuvõtte Pärtel Lippus (TÜ). Spontaanse kõne foneetiliselt märgendatud korpuses on 26 tundi ja 33 minutit lindistusi. Sõna- ja häälikutasandil on märgendatud kokku umbes 15 tundi kõnet. Korpust saab rakendada eesti keele kõnesünteesis ja kõnetuvastuses ning häälduse põhiparameetrite uurimisel.

Projekti „Eesti emotsionaalse kõne korpus” senistest põhitulemustest rääkis Hille Pajupuu (EKI). Selle korpuse esmaseks rakenduseks on emotsionaalse kõnesünteesi tarbeks akustilise baasi loomine. Praegu on korpuses 1091 lauset. Lausete emotsiooni määramiseks on läbi viidud taju- ja lugemisteste, kus osalejatel on tulnud määrata kontekstivabade lausete emotsioon heli ja kirjaliku teksti põhjal.

Meelis Mihkla (EKI) andis ülevaate projekti „Eestikeelne korpusepõhine kõnesüntees” alal saavutatust. Eestikeelse kõnesüntesaatori tarvis on koostatud foneetiliselt ja fonoloogiliselt representatiivne märgendatud kõnekorpus. Loodud on eestikeelsete teabetekstide ettelugemise süsteem nägemispuudega inimestele.

Einar Meister tutvustas projekti „Kõne analüüs ja variatiivsuse mudelid” seniseid põhitulemusi. Projekti eesmärgiks on kõne akustilise/foneetilise analüüsi meetodite uurimine ja arendamine ning erinevate kõnevariatsioonide foneetiliste kirjelduste ja kõnetehnoloogilisteks rakendusteks sobivate mudelite loomine. Seni on uuritud kõne mikroprosoodilisi nähtusi ning eesti ja vene emakeelega keelejuhtide eesti vokaalkategooriate taju, plaanis on jätkata kõne makroprosoodiliste nähtuste uurimisega.

Mare Koit (TÜ) tegi kokkuvõtte lõppenud projektist „Eestikeelne infodialoog arvutiga” ja tutvustas uut projekti „Intelligentne kasutajaliides andmebaasidele”. Esimese projekti raames on välja töötatud veebis kasutatav nn vestlusrobot, mis annab infot Tartu kinokavade kohta. See dialoogisüsteem integreerib ajaväljendite automaatse tuvastamise ja dialoogi juhtimise reeglid ning tekst-kõne-sünteesi. Uue projekti eesmärgiks on luua kasutajaliides, mis võimaldaks hõlpsalt kasutada erinevate ainevaldkondade andmebaase. Kasutajal oleks võimalus pöörduda andmebaaside poole eesti keeles ja saada vastuseks adekvaatset infot kas tekstina või tehiskõnes.

Esimese konverentsipäeva lõpetas Heiki-Jaan Kaalepi (TÜ) ettekanne projektist „Korpusepäring keeleveebis”, mille põhieesmärgiks on võimaldada tekstikorpust ja sõnastikke internetis mugavalt kasutada. Selle projekti raames on tehtud tasuta kättesaadavaks suur hulk erialasõnastikke ning loodud ühispäring 15 miljoni sõnalisele tasakaalus korpusele, nii et on võimalik otsida sõnu ka algvormi ja grammatiliste kategooriate järgi.

Teist konverentsipäeva alustas külalisineja Kristjan Rebane (Eesti Arengufond), kes rääkis Eesti Arengufondi info- ja kommunikatsioonitehnoloogia (IKT) arenguseirest, mis vaatleb maailmas toimuvaid IKT-alaseid arenguid ja Eesti võimalusi selles valdkonnas kümne aasta perspektiivis. Ettekandest jäi kõlama tõsiasi, et Eesti oma väiksuse juures ei hakka maailma tehnikatrende ja -turge kunagi olulisel määral suunama. Meil oleks aga vaja maailmas toimuvaga kursis olla ning panustada eelkõige haridusse ning teadusse.

H.-J. Kaalep andis ülevaate lõppenud projektist „Masintõlge 1” ja tutvustas uut projekti „Masintõlge 2”. Lõppenud projekti raames loodi eesti-inglise masintõlke katseversioonid. Edasise

töö eesmärgiks on parandada olemasoleva masintõlke kvaliteeti, kasutades keelespetsiifilist tarkvara (esmajärjekorras süntaksianalüsaatoreid) ning kombineerides erineval moel treenitud programmiversioone.

Järgesid kahte uut projekti tutvustavad ettekanded. Maarika Traat (TÜ) rääkis projektist „Automaatne parafraside leidmine ning sõnade ja lühifraaside tõlkimine paralleelcorpuste abil”, mille raames on kavas luua eesti keele jaoks veebiliidesega tööriist, mis võimaldab kasutajal sisestada sõna või fraas ning saada päringule vastuseks kas tõlked valitud võõrkeeles või parafrasid lähtekeeles. Tööriista töö põhineks joondatud paralleelcorpuste kasutamisel ning koostööd loodetakse teha projektiga „Masintõlge 2”.

Peep Küngas (TÜ) esitles projekti „Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp”. Projekti raames plaanitakse edendada süntaksi- ja semantikapõhiste infootsingute toetamiseks vajaminevat tehnoloogiat nii veebis kui ka arvutiprogrammides. Kavandatavaks põhitulemuseks on eesti keelele kohandatud infootsingu tarkvaralahendus.

Kadri Muischnek (TÜ) andis ülevaate projekti „Eesti keele koondkorpus” tulemustest. Projekti raames kogutava kirjaliku eesti keele korpuse planeeritud maht – 200 miljonit sõna – on kokku kogutud. Korpus koosneb ajakirjandus- ja ilukirjandustekstidest, teadusartiklitest, Eesti ja Euroopa seaduste tekstidest, Riigikogu stenogrammidest ja nn uue meedia ehk interneti tekstidest.

Külli Habicht (TÜ) rääkis projektist „Eesti vanema kirjakeele elektroonilised kogud”, mille eesmärgiks on eesti vanema kirjakeele elektrooniliste kogude täiendamine ja vana materjali laiemasse kasutusse toomine. Projekti raames on tegeldud XVIII ja XIX sajandi tekstide korpusesse sisestamise, lem-

matiseerimise ja morfoloogilise märgendamise. Tekstikorpus koos pärin-gusüsteemiga on kasutatav interneti kaudu. Korpusepäringu hõlbustamiseks on välja arendatud kasutajaliides.

Heili Orava (TÜ) juhitud projekt „TÜ eesti keele tesauruse (eesti Wordnet'i) täiendamine” sai alguse 2007. aastal ning tegeleb eesti keele tesauruse täiendamisega, et seda saaks kasutada teiste keeletehnoloogiliste rakenduste jaoks. Konverentsi ajaks sisaldas tesaurus umbes 22 500 mõistet.

Neeme Kahusk (TÜ) tutvustas uut projekti „Eesti keele semantika ressursid ja vahendid”. Projekti eesmärkideks on semantiliselt ühestatud ja märgendatud sõnatähendustega korpuse korrastamine ja suurendamine ning sõnatähenduste käsitsi ühestamist ja leksikaal-semantilise andmebaasi koostamist ja haldamist hõlbustava tarkvara loomine. Lisaks on kavas üldontoloogia ja valdkonnaontoloogiatega põhise mitmekeelse infootsingu raamistiku, ressursside ja rakenduste loomine.

Ülle Viks ja Andres Loopmann (EKI) andsid ülevaate projekti „Leksikograafi töökeskkond” senistest tulemustest. Loodud on leksikograafidele sobiv interaktiivne töökeskkond ehk sõnastike haldussüsteem EELEX, mida kasutades on valminud neli sõnastikku. Jätkunud on „Eesti-x-keele sõnastiku” ehk eesti lähtekeele andmebaasi edasiarendamine uute kakskeelsete sõnaraamatute jaoks. On loodud „Eesti kirjakeele seletussõnaraamatu” avalik veebiversioon ja täiendatud ÕS-i struktuuri-põhist päringut. Samuti on olemas „Eesti-x-keele sõnastiku” esialgne avalik demoversioon, mille abil on võimalik koostada oma kakskeelne sõnaraamat.

Oma teises ettekandes võttis N. Kahusk kokku lõppenud projekti „Lihtlause semantiline analüüs 1” ja tutvustas uue projekti „Lihtlause semantiline analüüs 2” eesmäärke. Esimese raames loodi eestikeelse lause semantiliseks

analüüsiks vajalikud ressursid – freimileksikon ja semantiliselt märgendatud korpus – ning semantilise analüüsi tarkvara, mille olulisimaks esindajaks on semantilisi järeldusi tegev programm. Uus projekt jätkab tööd lihtlause semantilise modelleerimise ning vastava tarkvara väljatöötamisega.

Kaili Müürisep (TÜ) tegi samuti ettekande kahest projektist. Lõppenud projekti „Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid” raames töötati välja grammatikakorrektori ja eesti keelsete tekstide sisukokkuvõtja prototüübid. Keeleressurssidest on loodud automaatseks analüüsiks vajalikke grammatikaid ning pindsüntaktiliselt märgendatud korpus ning puudepank. Uue projekti „Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs” eesmärgiks on luua grammatikakorrektori tööversioon ning arendada suulise keele süntaksianalüsaatorit ja murdetekstide, interneti keele ja õppijakeele süntaktilist analüüsi.

Katre Õim (Eesti Kirjandusmuuseum) rääkis projektist „Eesti fraseoloogismide elektroonilise alussõnastiku loomine”, mida alustati 2008. aastal ning mille eesmärgiks on elektrooniline eesti fraseoloogismide mõisteline alussõnastik. Praeguseks on välja töötatud alussõnastiku makro- ja mikrostruktuur ning „Eesti kõnekäändude ja fraseoloogismide andmebaasist” (EKFA) on alussõnastiku jaoks esialgu välja valitud 20 671 fraseologismi.

Pille Esloni, Helena Metslangi ja Annekatrin Kaivapalu (TLÜ) ühisettekanne tutvustas projekti „VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine”. Alates 2008. aastast on kogutud eesti keele kui teise keele või võõrkeele kasutajate kirjalikke tekste. Korpuses on käsitsi märgendatud õppijakeele

vead. Töötatakse välja tarkvararakendusi (sh vealeidja prototüüpi), mis võimaldavad käsitsi märgendamisel üle minna automaatsele.

Kristiina Praakli (TÜ) andis ülevaate projektist „Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid”. Selle raames on koostatud õppijakeele paralleelkorpus ja tekstikorpus, mis sisaldab TÜ-s eesti keelt teise keelena või võõrkeelena õppivate üliõpilaste kirjalikke tekste. Töö eesmärgiks on luua andmebaas, mis pakub autentset keelematerjali õppijakeele kirjaliku kasutamise uurimiseks, ja arendada õppijakeele korpuse põhjal välja eesti keele õppimist toetav veebikeskkond.

Konverentsi lõpetasid kolme lõppenud projekti tulemusi kokkuvõtavad ettekanded. H.-J. Kaalepi projekti „Mittesõnaliste verbide ja nende kokkulahekirjutamise vigade äratundmine eestikeelsetes tekstides” raames loodi kolm omavahel seotud keeleressurssi: verbikesksete püsiühendite andmebaas, morfoloogiliselt ühestatud tekstikorpus ning programm, mis andmebaasi ja korpuse alusel märgendab verbikeskseid väljendeid.

Projektiga „Reeglipõhine keeletarkvara”, mida konverentsil esindas Jaak Pruulmann-Vengerfeldt (TÜ), loodi universaalne reegliformaat ja selle kasutamise tarkvara. Samuti töötati välja programmid, mida on võimalik kasutada Eesti Keele Instituudis varem väljatöötatud morfoloogiataarkvara ja -reeglite uuendamiseks.

Jaak Vilo juhitud projekti „Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid” põhitulemusena töötati välja struktureeritud sõnastike andmebaasi genereerimise vahend ning päringusüsteemi prototüübid.

Järgmine EKKTT konverents on kavas korraldada 2010. aasta lõpus.

EVA LIINA ASU