EKKTT 2009

Riikliku programmi
Eesti keele keeletehnoloogiline tugi
(2006–2010)
teine konverents

6.–7. aprill 2009
Dorpat Konverentsikeskus, Tartu

Riiklik programm
"Eesti keele keeletehnoloogiline tugi (2006–2010)"

Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006–2010)" (EKKTT) peaeesmärgiks on eesti keele keeletehnoloogilise toe arendamine tasemele, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises keskkonnas. Programmi alaeesmärkideks on keeletarkvara ja keeletehnoloogiliste ressursside arendamine ning keeletehnoloogia infrastruktuuri ajakohastamine.

EKKTT rahastab keeletehnoloogiaalast teadus- ja arendustegevust alates ressursside loomisest kuni keeletehnoloogiliste rakenduste prototüüpide loomiseni. Põhjendatud vajaduse korral võib programmi raames rahastada ka eeluuringuid. Programmi tulemusena tekkiv intellektuaalne omand on avalik omand. EKKTT programm käivitus 2006. aasta viimases veerandis. Igal aastal vaadatakse läbi uute projektide rahastamistaotlused ja juba käimasolevate projektide jätkutaotlused. 2009. aastal rahastatakse 23 projekti.

EKKTT programmi koduleht: http://www.keeletehnoloogia.ee/

Programmi juhtkomitee koosseis:

Juhtkomitee esimees

Jaak Vilo	Arvutiteaduse instituut, Tartu Ülikool

Juhtkomitee aseesimees

Einar Meister	Tallinna Tehnikaülikooli Küberneetika Instituut

Liikmed

Heiki-Jaan Kaalep	Eesti ja üldkeeleteaduse instituut, Tartu Ülikool
Kaili Müürisep	Arvutiteaduse instituut, Tartu Ülikool
Karl Pajusalu	Eesti ja üldkeeleteaduse instituut, Tartu Ülikool
Indrek Reimand	Haridus- ja Teadusministeerium
Urmas Sutrop	Eesti Keele Instituut ja Tartu Ülikool
Uuno Vallner	Majandus- ja Kommunikatsiooniministeerium
Kadri Vider	Haridus- ja Teadusministeerium

EKKTT koordinaator:

Eva Liina Asu-García	Eesti ja üldkeeleteaduse instituut, Tartu Ülikool

2009. aastal käimasolevad EKKTT projektid

Jätkuprojektid

EKKTT06-1	Eesti emotsionaalse kõne korpus	Hille Pajupuu	Eesti Keele Instituut
EKKTT06-2	Eestikeelne korpusepõhine kõnesüntees	Meelis Mihkla	Eesti Keele Instituut
EKKTT06-3	Leksikograafi töökeskkond	Ülle Viks	Eesti Keele Instituut
EKKTT06-4	Korpusepäring keeleveebis	Heiki-Jaan Kaalep	Filosoft OÜ
EKKTT06-5	Kõne analüüs ja variatiivsuse mudelid	Einar Meister	Tallinna Tehnikaülikool
EKKTT06-6	Kõnekeele ressursid ja kõnetehnoloogia andmebaasid	Einar Meister	Tallinna Tehnikaülikool
EKKTT06-7	Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine	Tanel Alumäe	Tallinna Tehnikaülikool
EKKTT06-8	Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid	Kristiina Praakli	Tartu Ülikool
EKKTT06-13	Eesti vanema kirjakeele elektroonilised kogud	Külli Habicht	Tartu Ülikool
EKKTT06-14	Eesti keele koondkorpus	Kadri Muischnek	Tartu Ülikool
EKKTT06-16	Eesti keele spontaanse kõne foneetiline korpus	Pire Teras	Tartu Ülikool
EKKTT07-21	TÜ eesti keele tesauruse (eesti wordneti) täiendamine	Heili Orav	Tartu Ülikool
EKKTT08-26	VAKO - Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine	Pille Eslon	Tallinna Ülikool
EKKTT08-28	Eesti keeleressursside keskus	Tiit Roosmaa	Tartu Ülikool
EKKTT08-29	Eesti fraseologismide elektroonilise alussõnastiku loomine	Katre Õim	Eesti Kirjandusmuuseum

Uued projektid

EKKTT09-57	Intelligentne kasutajaliides andmebaasidele	Mare Koit	Tartu Ülikool
EKKTT09-61	Tartu ülikooli eesti kõnekeele audio- ja videokorpuse kogumine ja otsingutarkvara loomine	Tiit Hennoste	Tartu Ülikool
EKKTT09-62	Eesti keele semantika ressursid ja vahendid	Neeme Kahusk	Tartu Ülikool
EKKTT09-63	Lihtlause semantiline analüüs 2	Haldur Õim	Tartu Ülikool

EKKTT09-64	Masintõlge 2	Heiki-Jaan Kaalep	Tartu Ülikool
EKKTT09-65	Automaatne parafraaside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpuste abil	Maarika Traat	Tartu Ülikool
EKKTT09-66	Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp	Peep Küngas	Tartu Ülikool
EKKTT09-67	Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs	Tiit Roosmaa	Tartu Ülikool

2008. aastal lõppenud EKKTT projektid

EKKTT06-9	Masintõlge 1	Heiki-Jaan Kaalep	Tartu Ülikool
EKKTT06-10	Mitmesõnaliste verbide ja nende kokku-lahku kirjutamise vigade äratundmine eestikeelsetes tekstides	Heiki-Jaan Kaalep	Tartu Ülikool
EKKTT06-11	Lihtlause semantiline analüüs 1	Haldur Õim	Tartu Ülikool
EKKTT06-12	Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid	Jaak Vilo	Tartu Ülikool
EKKTT06-15	Eestikeelne infodialoog arvutiga	Mare Koit	Tartu Ülikool
EKKTT06-17	Eesti kõnekeele korpuse kogumine ja translitereerimine	Tiit Hennoste	Tartu Ülikool
EKKTT06-18	Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid	Tiit Roosmaa	Tartu Ülikool
EKKTT07-23	Reeglipõhine keeletarkvara	Jan Villemson	Tartu Ülikool

Konverentsi programm

6. aprill

10.00	Algab registreerimine ja pakutakse kohvi ja teed
10.30	Avasõnad: Indrek Reimand ja Jaak Vilo
11.00	Plenaarettekanne: Tanel Alumäe “Kõnetuvastusest Eestis ja CNRS/LIMSI laboris Prantsusmaal”
12.00	Tiit Roosmaa “Eesti keeleressursside keskus”
12.30	Lõuna
13.30	Einar Meister “Kõnekeele ressursid ja kõnetehnoloogia andmebaasid”
14.00	Tiit Hennoste “Kõnekeele korpus“ ja “Kõnekeele audio- ja videokorpus”
14.30	Pärtel Lippus “Spontaanse kõne foneetiline korpus”
15.00	Hille Pajupuu “Emotsionaalse kõne korpus”
15.30	Kohvipaus
16.00	Meelis Mihkla “Eestikeelne korpusepõhine kõnesüntees”
16.30	Einar Meister “Kõne analüüs ja variatiivsuse mudelid”
17.00	Mare Koit “Eestikeelne infodialoog arvutiga” ja “Intelligentne kasutajaliides andmebaasidele”
17.30	Heiki-Jaan Kaalep “Korpusepäring keeleveebis”
18.00	Ettekandepäeva lõpp
19.00	Õhtusöök Atlantises

7. aprill

9.00	Plenaarettekanne: Kristjan Rebane (Arengufond) “Info- ja kommunikatsioonitehnoloogia arenguseire”
9.50	Heiki-Jaan Kaalep “Masintõlge 1” ja “Masintõlge 2”
10.20	Maarika Traat “Automaatne parafraaside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpuste abil”
10.40	Peep Küngas “Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp”
11.00	Kohvipaus
11.30	Kadri Muischnek “Eesti keele koondkorpus”
12.00	Külli Habicht “Eesti vanema kirjakeele elektroonilised kogud”
12.20	Heili Orav “TÜ eesti keele tesauruse (eesti wordneti) täiendamine”
12.40	Neeme Kahusk “Eesti keele semantika ressursid ja vahendid”
13.00	Lõuna
14.00	Ülle Viks ja Andres Loopmann “Leksikograafi töökeskkond”
14.30	Haldur Õim “Lihtlause semantiline analüüs 1” ja “Lihtlause semantiline analüüs 2”
15.00	Kaili Müürisep “Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid” ja “Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs”
15.30	Katre Õim “Eesti fraseologismide elektroonilise alussõnastiku loomine”
16.00	Kohvipaus
16.30	Pille Eslon, Erika Matsak, Helena Metslang, Vahur Rebas ja Annekatrin Kaivapalu “VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine”
17.00	Krsitiina Praakli “Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid”
17.30	Heiki-Jaan Kaalep “Mitmesõnaliste verbide ja nende kokku-lahku kirjutamise vigade äratundmine eestikeelsetes tekstides”
17.50	Jaak Pruulmann-Vengerfeldt “Reeglipõhine keeletarkvara”
18.10	Jaak Vilo “Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid”
18.30	Konverentsi lõpp

Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine

Vastutav täitja	Tanel Alumäe		
Teised põhitäitjad	Toomas Kirt		
Finantseerimine 2008	790 000	**Finantseerimine 2009**	575 000

Eesmärgid ja tähtsus

Projekti eesmärgiks on eesti keelele sobivate kõnetuvastuse meetodite uurimine, arendamine ja testimine ning erinevate tuvastussüsteemide prototüüpide loomine. Projekti raames luuakse eesti keelele sobiv kõnetuvastustehnoloogia ja arendatakse välja piiratud ning piiramatu sõnavaraga tuvastussüsteemide prototüübid. Kõnetuvastustehnoloogia väljatöötamine võimaldab hakata arendama suulisel kommunikatsioonil baseeruvaid kasutaja-sõbralikke liideseid, mis leiaksid rakendust infotehnoloogilistes süsteemides. Kõnetuvastustehnoloogia loomine tagab eesti keelele "suurte" keeltega võrdsed tingimused ja kasutusvõimalused infotehnoloogilises keskkonnas ning loob seega eeldused eesti keele säilimiseks ja arenguks infoühiskonnas.

Põhitulemused

Senise töö raames on teostatud mitmesuguseid uuringuid, keskendudes suure sõnavaraga eestikeelse kõnetuvastuse probleemistikule ning erinevate rakendusprototüüpide väljatöötamisele.

2006–2007. a läbi viidud uuringud ja eksperimendid kontsentreerusid põhiliselt morfeemipõhisele suure sõnavaraga kõnetuvastuse keelemudeli detailidele. Muu hulgas pakuti välja suhteliselt täpne meetod liitsõnade rekonstrueerimiseks morfeemidest koosnevast tuvastaja väljundist. Samuti uuriti keelemudeli teema-põhise adapteerimise temaatikat.

2008. a loodi Eesti Raadio (ER) uudistekorpuse käsitsi märgendamise kiirendamiseks uudiste automaatne transkriptsioonisüsteem. Süsteem kasutab varem välja töötatud eesti keele tuvastusmootorit. Konkreetse päeva uudistesaadete transkribeerimiseks luuakse sellele päevale kohandatud keelemudel, milleks kasutatakse ER-st saadud diktorite poolt kasutatud uudistetekste. Selline keelemudel peaks hõlbustama antud päeva uudistetekstides olevate sõnakombinatsioonide tuvastamist, kuid suutma rahuldavalt toimida ka tekstide puudumise korral. Süsteem on praeguseks täielikult tööle rakendatud.

Teiseks 2008. a teostatud uurimistöö põhisuunaks oli kõnelõikude automaatne grupeerimine kõnelejate järgi (ingl k *speaker diarization*).See ülesanne on äärmiselt oluline pikkade kõnesalvestuste (näiteks loengute ja raadiote jutusaadete salvestused) automaatselt transkribeerimisel, indekseerimisel ja struktureerimisel. Esialgu on sellealane uurimistöö sisaldanud üldist kontseptuaalse raamistiku uurimist ja meetodite implementeerimist.

Samuti on projekti raames implementeeritud kaks eesti keele tuvastustehnoloogiat kasutavat rakendusprototüüpi: autosegmenteerija ning häälega juhitav kalkulaator.

Eesti keeleressursside keskus

Vastutav täitja	Tiit Roosmaa		
Teised põhitäitjad	Krista Liin		
Finantseerimine 2008	497 000	**Finantseerimine 2009**	475 000

Eesmärgid ja tähtsus

Aastal 2008 käivitus Eesti keele keeletehnoloogilise toe riikliku raamprogrammi projekt, mille eesmärgiks on luua Eesti keeleressursside keskus.

Käesolevas projektis mõistame keeleressursside all nii keeletarkvara kui olemasolevaid keeletehnoloogilisi ressursse.

Loomuliku keele ressursid on erinevate soovijate/huviliste poolt kasutatavad ainult siis, kui olemasolevad keeleressursid on korralikult dokumenteeritud ja arhiveeritud ning avalikult kättesaadavad. Selliste, kohati keeleressursside loojatele tarbetutena tunduvate tegevuste toetamiseks on vaja teatud infrastruktuuri olemasolu, mis korraldaks ja koordineeriks sellealast tööd Eestis. Loodava keskuse tegevusvaldkond oleks alates keeletehnoloogiliste standardite väljatöötamisest/fikseerimisest kuni keeleressursside kasutamiseks vajalike juriidiliste lepingute/litsentside koostamiseni.

Sellise eesmärgi saavutamiseks on käivitunud ESFRI projekt CLARIN (*Common Language Resources and Technology Infrastructure*, http://www.clarin.eu), milles üheks kolmekümne ühest partnerist on Eesti ametliku esindajana ka Tartu Ülikool. Osalemine CLARINi võrgustikus annab meile unikaalse võimaluse kaasata oma probleemide lahendamisse üleeuroopaline kogemus. Põhjamaades käivitus 2002. aastal analoogiline põhjamaade ministrite nõukogu projekt nime all *Language Technology Documentation Centre*, (http://www.nordoknet.org/) mille toel loodi keskuste võrgustik Soomes, Rootsis, Taanis, Islandil ja Norras.

Loodav Eesti keeleressursside keskus püüab, kasutades projekti CLARIN kaudu liikuvat teadmust, teha kõik temast oleneva, et Eestis olemasolev keeleressurss ei jääks ainult loojate ja koostajate teada, vaid jõuks kõigi võimalike huvilisteni nagu näiteks keeleuurijad, õpetajad, tarkvarasüsteemide ja -rakenduste loojad, riigiametnikud jne.

Kõnekeele ressursid ja kõnetehnoloogia andmebaasid

Vastutav täitja	Einar Meister		
Teised põhitäitjad	Lya Meister, Rainer Metsvahi		
Finantseerimine 2008	750 000	**Finantseerimine 2009**	630 000

Eesmärgid ja tähtsus
Projekti eesmärgiks on eesti keele foneetilisteks ja kõnetehnoloogilisteks uuringuteks ning arendustöödeks vajalike kõnekorpuste salvestamine, digitaliseerimine, märgendamine ja arhiveerimine, samuti ühtse tehnoloogilise keskkonna loomine erinevate andmebaaside haldamiseks ja efektiivseks kasutamiseks.

Põhitulemused
Uudistekorpus
Korpus sisaldab ca 300 tundi Eesti Raadio lühiuudiste salvestusi ja üle 8000 lk digitaliseeritud uudistetekste. Uudistekorpuse märgendamiseks on välja arendatud töökeskkond vabavaralise programmi Transcriber (http://trans.sourceforge.net) baasil. Hõlbustamaks korpuse märgendamist, on automaatse kõnetuvastuse abil genereeritud signaalifailidele vastavad tekstifailid. Uudiste tuvastuseks adapteeritud keelemudel ja vastav tuvastusmoodul on loodud T.Alumäe projektis „Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine“.

Aktsendikorpus
Aktsendikorpus sisaldab eri emakeelega inimeste eestikeelse kõne salvestusi. Seni on salvestatud 135 keelejuhi kõnematerjal, kelle keeletaust on järgmine: soome (30 keelejuhti), vene (40), prantsuse (12), saksa (13), itaalia (4), hispaania (2), taani (2), hollandi (2), slovaki (2), inglise (2), läti (1), leedu (1), jaapani (1), šoti (1), iiri (1), aserbaidžaani (1); võrdlusmaterjalina on salvestatud 20 eesti emakeelega keelejuhi kõnenäited.

Loengukõne korpus
Korpus sisaldab üle 100 tunni eri ainevaldkondade akadeemiliste loengute salvestustusi (erinevate lektorite arv on 15) ja ca 4 tundi konverentsiettekandeid (20 isikut, keskmine ettekande pikkus 20 min). Loengukõne korpus on ettevalmistamisel märgendamiseks Transcriber'iga.

Infrastruktuuri kaasajastamine
On välja ehitatud ja sisustatud kõnesalvestusstuudio, kõnekorpuste tarvis on paigaldatud eraldi server (Dell PowerEdge R200, kõvaketta maht 2TB). Kõnekorpuste haldamiseks ja kasutamiseks kohandatakse korpuste haldussüsteem LAMUS (Language Archive Management and Upload System, välja töötatud Max Planck'i Psühholingvistika Instituudis http://www.lat-mpi.eu/tools/lamus/).

Kõik Küberneetika Instituudis loodavad korpused tehakse kättesaadavaks LAMUS-süsteemi kaudu.

Eesti kõnekeele korpuse (TÜKK) kogumine ja translitereerimine (2006–2008)
Tartu ülikooli eesti kõnekeele audio- ja videokorpuse kogumine ja otsingutarkvara loomine (2009–2010)

Vastutav täitja	Tiit Hennoste		
Teised põhitäitjad	Olga Gerassimenko, Riina Kasterpalu, Andriela Rääbis, Krista Strandson		
Finantseerimine 2008	455 000	**Finantseerimine 2009**	470 000

Eesmärgid ja tähtsus

Mõlema projekti eesmärgiks on tegelike spontaansete dialoogide ja monoloogide pragmaatilis-suhtlusliku korpuse kogumine. See võimaldab analüüsida keele kasutamist suhtluses, mida ei võimalda teist tüüpi korpused. Samuti võimaldab ainult see korpus analüüsida suulise keele süntaksit, milles on suulise ja kirjaliku keele keskne erinevus. TÜKK on maailma mastaabis harv suur tegeliku suhtluse korpus, mis on transkribeeritud põhjalikult ja varustatud väga põhjalike taustakirjeldustega suhtlsusituatsioonide ja kõnelejate kohta. TÜKK on allikas suulise kõne keeleteaduslikuks analüüsiks ja dialoogi modelleerimiseks. Analüüs on eelduseks nt kõnetuvastusele ja telefonipõhistele infosüsteemidele, interaktiivsetele kõnekeele õppeprogrammidele, suulise kõne erisõnastike koostamisele, mis on pea kõigi keeletehnoloogiliste rakenduste realiseerimiste eelduseks.

Põhitulemused

TÜKK koosneb kõnelindistustest (argi- ja institutsionaalne, monoloog ja dialoog, silmast-silma, telefoni- ja meediasuhtlus), nende transkriptsioonidest, taustakirjeldustest ja tarkvarast, mis võimaldab otsida ja analüüsida korpusest erinevaid keelelisi nähtusi. Lindistatud on 2,1 miljonit tekstisõna, transkribeeritud 1,35 miljonit tekstisõna.

Kahte projekti ühendavad tegevused:

- suuliste tekstide lindistamine, transkribeerimine ja taustakirjeldustega varustamine
- inimese-arvuti suhtluse modelleerimisel kasutatava Dialoogikorpuse koostamine
-1997-2004 kogutud analoogkorpuse digitaliseerimine ja transkriptsioonide täpsustamine
- taustakirjelduste korrastamine, mis on eelduseks selle maksimaalselt arvutipõhisele kasutusele
- täppistranskriptsiooni valdavate transkribeerijate koolitamine
- korpuse kogumise ja kasutamisega seotud juriidiliste probleemide lahendamine

Alanud projekti uued eesmärgid:

- videokorpuse kogumise ja transkribeerimise alustamine, sh mitteverbaalne suhtlus
- Jaak Vilo ligikaudse otsimise süsteemi kohandamine suulise keele sõnavariantide päringusüsteemiks.
- taustakirjelduse süsteemi viimine formaati, mis on vajalik automaatotsinguks

Suurem osa korpusest on kasutatav uurimistööks korpuse administraatoriga sõlmitava lepingu/konfidentsiaalsuskohustuse alusel. Korpust ja selle põhjal tehtud töid on töörühma liikmed tutvustanud rohkem kui 20 konverentsil ja avaldanud ligi 30 artiklit erinevates rahvusvahelistes väljaannetes.

Eesti keele spontaanse kõne foneetiline korpus

Vastutav täitja	Pire Teras		
Teised põhitäitjad ja täitjad	Pärtel Lippus, Tuuli Tuisk, Nele Salveste, Liis Raasik		
Finantseerimine 2008	774 000	**Finantseerimine 2009**	735 000

Eesmärgid ja tähtsus

Selle projekti eesmärgiks on luua teiste eestikeelse kõne korpustega ühilduv spontaanse kõne foneetiliselt märgendatud korpus, mida saab kasutada eesti keele häälduse põhiparameetrite analüüsimisel ning eesti keele kõnesünteesi ja kõnetuvastuse ülesannete täitmisel.

Eesti keele spontaanse kõne foneetilise korpuse jaoks tehakse spontaanse kõne kõrge kvaliteediga salvestusi. Foneetilisse korpusse salvestatakse esimeses etapis 40 keelejuhi kõne (20 naist ja 20 meest, umbes 30 minutit keelejuhilt). Salvestised on kas dialoogid argivestlustena või institutsionaalsed monoloogid loengute ja ettekannetena). Korpusesse valitakse 20–60-aastased eesti keelt emakeelena rääkivad keelejuhid, kellel on erinev sotsiaalne ja hariduslik taust.

Salvestatud kõne märgendatakse foneetiliselt erinevatel märgenduskihtidel. Segmentimis- ja transkribeerimisalused jms on lepitud kokku koostöös Eesti Keele Instituudi ning TTÜ Küberneetikainstituudi foneetika ja kõnetehnoloogia laboriga. Koostöö tulemusel on korpus kasutatav nii kõnetehnoloogiliste rakenduste arendamiseks kui eesti keele foneetika teoreetiliseks uurimiseks.

Põhitulemused

Hetkel on korpuses lindistusi 32 keelejuhilt kogukestusega 26 tundi ja 33 minutit, mis on juba praegu rohkem, kui esialgu kavandatud. Lindistuste suurem kogumaht tuleneb sellest, et mõni keelejuht osaleb mitmes lindistuses. Keelejuhtidest 11 on 20. aastates, 10 on 30. aastates, 8 on 40. aastates ja 3 on 50. aastates ja vanemad. Lindistustest 21 on dialoogid ja 5 monoloogid.

Lindistused märgendatakse 8 eri märgenduskihil (sõnad, häälikud, silbid, taktid, lausungid, häälelaad, paralingvistilised nähtused, muu). Praeguse seisuga on korpuses sõna- ja häälikutasandil kokku 433 891 segmenti (sõnatasandil 107 996 ja häälikutasandil 325 895 segmenti). Muudel tasanditel on kokku 164 401 segmenti. Kõiki segmente on korpuses kokku 595 292. Põhitasanditel (st sõna- ja häälikutasandil) on märgendatud kokku umbes 15 tundi kõnet.

Korpuse otsingumootori prototüüp on valmis. Olemasolev otsingumootor võimaldab teha päringuid sõnatasandi segmentide piires ja kuvab vastused viiesõnalises kontekstis, soovi korral annab ka vastava lõigu helifailist ja TextGridist. Päringut võib teha nii sõnatasandi (ortograafilise kirjaviisi) kui häälikutasandi (SAMPA transkriptsiooni) märgendite kohta. Otsingut on võimalik kasutada programmis Praat. Nüüd, kus on olemas uus server, on kavas teha internetipõhine kasutajaliides.

Korpuse põhjal on tehtud mitu uurimust, kus on uuritud kõne struktuuri, kvantiteedi ja intonantsiooni vastastikust mõju, sõnaalgulise *h*, vokaalidevaheliste lühikeste klusiilide, järgsilpide *e* häälдust. Ilmunud või ilmumas on kaks artiklit, valminud on kaks bakalaureuse- ja üks magistritöö.

Eesti emotsionaalse kõne korpus

Vastutav täitja	Hille Pajupuu		
Teised põhitäitjad	Rene Altrov, Kairi Tamuri		
Finantseerimine 2008	930 000	**Finantseerimine 2009**	958 000

Eesmärgid ja tähtsus

Projekti eesmärgiks on luua usaldusväärne akustiline baas emotsionaalse kõnesünteesi tarbeks. Eesti emotsionaalse kõne korpusele on juurdepääs Eesti Keele Instituudi kodulehelt www.eki.ee "Eesti keele keeletehnoloogiline tugi" projektide alt ja programmi Eesti keele keeletehnoloogiline tugi projektidest www.keeletehnoloogia.ee/projektid.

Põhitulemused

2008. aastal on jätkunud materjali lisamine korpusesse. Selleks on ajakirjandusest välja valitud tekstilõigud (114 lõiku, lõigus keskmiselt 6-7 lauset). On läbi viidud lugemistest, mille käigus on lõikude emotsionaalsust lastud hinnata kirjaliku teksti põhjal ilma heli kuulmata.

Tekstilõigud on salvestatud ja segmenteeritud lauseteks. Neist on 2008. aastal kokku pandud 4 tajutesti ja 4 lugemistesti ning need läbi viidud. Testides osalejatel on tulnud määrata kontekstivabade lausete emotsioon heli ja lugemise põhjal. Heli- ja lugemistestides osalesid erinevad testijad. Korpuse registreeritud testijaid on 159.

Korpuses on praegu 1091 lauset, neist on avalikkusele kättesaadavad tajutesti ja lugemistesti läbinud 579 lauset (4376 sõnet, 1245 sõna). Vt korpuses Aruanded / Testide tulemused.

Korpus on teiste teadaolevate emotsioonikorpustega võrreldes eriline ses mõttes, et selles eristatakse laused, mille emotsiooni kannab ainult heli, ja laused, mille emotsiooni taju võib mõjutada tekst. Selline eristus on saanud võimalikuks tänu kahe testi – tajutesti ja lugemistesti tulemuste võrdlusele. Korpusest leiavad usaldusväärset uurimismaterjali nii need, keda huvitab helis peituv emotsioon, kui ka need, keda huvitab kirjutatud tekstis peituv emotsioon (vt Aruanded / Lausete tulemused.)

Jätkub lausete segmenteerimine sõnadeks ja häälikuteks ning nende märgendamine. On lisandunud ortograafiatasand (koos märgendatud pausidega) ning sõna põhitooni tippude ja nõgude tasand.

On jätkunud korpuse tarkvara täiustamine:

- on täiendatud tajutestide interaktiivset koostamist;
- on lisatud märgendusel põhinevad aruanded: foneemide ja sõnade päringud emotsionaalsete parameetrite ja positsiooni järgi lauses;
- on lisatud emotsioonisõnade päring lugemistestide andmete baasil;
- on alustatud süntesaatoripäringute implementeerimist. Implementeeritud on foneemide, sõnade ja suvalise ajaga määratud kõnelõigupäringud;
- on jätkatud serveri API laiendamist ja dokumenteerimist;
- on jätkatud korpuse liidese tõlkimist teistesse keeltesse (soome, inglise, läti).

Korpuse tehnilist dokumentatisooni vt http://193.40.113.40:5000/docs/

Korpuse arenduses on põhitähelepanu siirdunud planeeritud mahtude saavutamisele.

Eestikeelne korpuspõhine kõnesüntees

Vastutav täitja	Meelis Mihkla		
Teised põhitäitjad	Indrek Kiissel, Tõnis Nurk, Liisi Piits		
Finantseerimine 2008	1 225 000	**Finantseerimine 2009**	1 140 000

Eesmärgid ja tähtsus

Projekti eesmärgiks on suurte kõnekorpuste (40–200 minutit kõnet ühe diktori kohta) baasil sünteesida võimalikult loomulikku eestikeelset kõnet. Eestikeelne kõnekorpus on planeeritud professionaalsete diktorite etteloetud tekstide salvestistena (mees- ja naishääl). Korpusesse koondatakse foneetiliselt tasakaalustatud kõnematerjal. Projekti töömahukaim osa on kõnekorpuse märgendamine ja kõneüksusteks segmenteerimine. Kõneüksuste valiku hõlbustamiseks on plaanis rakendada fonoloogilise struktuuri vastavuse tehnikat, mille puhul kõnekorpus on esitatud fonoloogiliste puude võrgustikuna ja üksuste valikul leitakse sünteesitavale lausungile korpusest suurim sarnane struktuur. Kõnesünteesi arendamiseks kasutatakse arendussüsteeme Festival ja/või HTS ning testitakse sünteesimootorita korpussünteesi. Sünteeskõne loomulik rütm ja kõla modelleeritakse prosoodiageneraatoriga. Sidusa kõne korpuse baasil modelleeritakse erinevate statistiliste meetoditega (regressioon, klassifikatsioon ja regressioonipuud, närvivõrgud) häälekõrgust ja kõne ajalist struktuuri. Plaanis on siduda korpuspõhine süntees EKIs loodava emotsionaalse kõne korpusega, mis looks eelduse emotsionaalse kõnesünteesi rakendamiseks. Korpussünteesi on plaanis rakendada inimene–masin dialoogsüsteemide moodulina. Eelkõige vajavad kõnesünteesi nägemispuudega inimesed arvuti vahendusel info hankimiseks, sel eesmärgil luuakse erinevaid rakendusi ja liideseid korpuspõhise süntesaatori integreerimiseks Windows-põhiste rakendusprogrammidega.

Põhitulemused

1. Süntesaatori tarvis on koostatud ning salvestatud foneetiliselt ja fonoloogiliselt representatiivne kõnekorpus (54 minutit kõnet). Sünteesi akustilise baasina on kõnekorpus märgendatud ja esitatud fonoloogiliste struktuuridena. Kõnekorpuse ja genereeritava lausungi jaoks on loodud sisendteksti ja kõnelaine foneemmärgenduse alusel süsteem, mis genereerib automaatselt eri tasandite fonoloogilise liigenduse.
2. Kõneprosoodia vallas on uuritud statistiliste meetoditega kõne ajalise struktuuri ja põhitooni modelleerimist ning on välja töötatud vastav metodoloogia. Sõnadevaheliste kollokatsioonide ja kõnetempo vahelisi seoseid uurides pandi tähele, et eesti keeles on täheldatavad fraasikesksetele keeltele omased nähtused: sagedamini koosesinevate sõnade pikkus lüheneb.
3. Koostöös Põhja-Eesti Pimedate Ühinguga ja Eesti Pimedate Raamatukoguga loodi eestikeelsete teabetekstide ettelugemise süsteem nägemispuudega inimestele.
4. Eestikeelset korpussünteesi arendati Festivali keskkonnas ja testiti sünteesimootorita süsteemis, mis võimaldas hinnata kõneüksuste valikuprintsiipe ning lingvistilisi ja füüsikalisi sobituskaale signaalide ühendamisel. Festivali arendussüsteemi tarvis programmeeriti eesti keele jaoks vajaminevaid mooduleid: häälikusüsteem ja täht-häälik teisendus, mittesõnade interpretaator, fraseerimine, prosoodilised mudelid. Täiustati pauside ja funktsioonisõnade käsitlusprintsiipe.

Kõne analüüs ja variatiivsuse mudelid

Vastutav täitja	Einar Meister		
Teised põhitäitjad	Lya Meister, Jüri Kuusik		
Finantseerimine 2008	775 000	**Finantseerimine 2009**	550 000

Eesmärgid ja tähtsus
Projekti eesmärgiks on uurida ja arendada kõne akustilise/foneetilise analüüsi meetodeid ning luua erinevate kõnevariatsioonide foneetilised kirjeldused ja kõnetehnoloogilisteks rakendusteks sobivad mudelid.

Põhitulemused
Kõne mikroprosoodiliste nähtuste uurimine
Uuriti vokaali omakestuse rolli vokaalikategooria tajumisel vokaalipaarides /i/-/e/, /ü/-/ö/, /u/-/o/, /e/-/ä/ ja /o/-/a/. Kasutades sünteesstiimuleid, viidi läbi tajueksperimentide seeria 10 eesti emakeelega (5 meest, 5 naist) katsealusega. Esimeses eksperimendis määrati kõigi katsealuste individuaalsed vokaalikategooriate piirid; teise eksperimendi jaoks valiti kolm kategooriapiiri ümbruses paiknevat stiimulit, mille baasil sünteesiti erineva kestusega (60 kuni 140 ms) vokaalid.

Tajukatse tulemused näitasid, et vokaalikategooria taju ja vokaali kestuse vahel on oluline seos – pikemaid vokaale tajutakse madalama, lühemaid kõrgema vokaalina. Kõrge-keskkõrge vokaalipaaride (/i/-/e/, /ü/-/ö/, /u/-/o/) puhul on kestuse ja kvaliteeditaju seos tugevam kui keskkõrge-madal vokaalipaaride (/e/-/ä/, /o/-/a/) korral. Tulemuste erinevust selgitab vokaalide erinev foneetiline kaugus – kõrge-keskkõrge vokaalipaari vokaalide foneetiline kaugus on väiksem kui keskkõrge-madal vokaalipaaride puhul. Väiksema foneetilise kauguse puhul on vokaali omakestus lisatunnuseks, mis aitab vokaale paremini eristada; suurema foneetilise kaugusega vokaalide korral on spektraalne informatsioon piisav vokaalikategooria tuvastamiseks ja vokaali omakestus mõjutab taju vähem.

Aktsendiga kõne uurimine
Uuriti eesti ja vene emakeelega keelejuhtide eesti vokaalikategooriate taju. Tajukatse tarvis sünteesiti neljaformandilised eesti vokaaliprototüübid ja vahepealse kvaliteediga stiimulid, mis moodustasid vokaaliprototüüpide vahelises ruumis diskreetse rastri (16–18 stiimulit iga vokaalipaari vahel, kokku 14 vokaalipaari). Stiimulid esitati kuulajatele vokaalipaaride kaupa juhuslikus järjekorras ja neil tuli otsustada, millist vokaali kahest võimalikust nad tajusid. Katseisikutena osalesid 5 eesti emakeelega (2 meest, 3 naist) ja 9 vene emakeelega (4 meest, 5 naist) isikut.

Tajukatsed näitasid: 1) eesti vokaalide /i, e, u, o, a, ä/ kategooriapiire tajuvad vene emakeelega keelejuhid sarnaselt eestlastega; 2) vokaalide /ü, ö, õ/ piirid on vene emakeelega katsealuste tajuruumis tunduvalt hägusamad.

Esimesel juhul on tegemist eesti vokaalikategooriate assimileerumisega akustiliselt ja pertseptiivselt lähedaste vene vokaalivariantidega, teisel juhul on tegemist vene vokaalisüsteemi jaoks uute kategooriatega, mis osaliselt assimileeruvad vene /õ/-vokaaliga.

Eestikeelne infodialoog arvutiga

Vastutav täitja	Mare Koit		
Teised põhitäitjad	Mark Fišel, Olga Gerassimenko, Kristiina Jokinen, Riina Kasterpalu, Taavet Kikas, Andriela Rääbis, Krista Strandson, Margus Treumuth, Maret Valdisoo, Evely Vutt		
Finantseerimine 2008	550 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus

Seoses arvutite levikuga omandab järjest suurema tähtsuse tarkvara, mis vahendab inimese suhtlust arvutiga loomulikus keeles, sh kõne abil. Inglise jmt keele jaoks kasutatakse arvukalt kõnedialoogsüsteeme automaatsete telefoniteenuste osutamiseks erinevates valdkondades. Eesti keele jaoks selline süsteem seni puudub.

Projekti eesmärgiks on tarkvara väljatöötamine, mis võimaldaks eestikeelset küsimus-vastusdialoogi arvutiga. Sellise tarkvara loomiseks vajalik keeleressurss on dialoogiaktidega märgendatud dialoogikorpus, mille mahtu on dialoogiaktide automaatse analüüsi- ja sünteesiprogrammide väljatöötamiseks ja treenimiseks vaja suurendada vähemalt 200 000 tekstisõnani. Uuritakse dialoogiaktide automaatse tuvastamise erinevaid meetodeid, et valida eesti keelele sobiv formalism.

Põhitulemused

1. Märgendatud dialoogikorpuse maht on 2008.a lõpuks kasvanud 202 000 tekstisõnani.
2. Dialoogide märgendamise käigus on korrastatud dialoogiaktide tüpoloogiat, pidevalt täpsustatud ja täiendatud dialoogiaktide märgendusjuhendit.
3. On analüüsitud dialoogikorpuses leiduvate infodialoogide ülesehitust, sh telefonikõnede alustamist ja lõpetamist, alamdialoogide kasutamist, sagedamini esinevaid dialoogiakte ja nende väljendamist eesti keeles, eesmärgiga leida keelelisi märguandeid, mida dialoogsüsteem saaks kasutada dialoogiaktide automaatsel tuvastamisel.
4. Eesti dialoogikorpuse analüüsimise hõlbustamiseks on välja töötatud veebis kasutatav tarkvara, mis muuhulgas võimaldab otsida (parooliga kaitstud) korpusest ja loendada mitmesuguseid dialoogides esinevaid nähtusi: sõnu või sõnajärjendeid, transkriptsioonielemente, dialoogiakte, andes ette akti nime, osaleja tähise, sõne. Samuti teha automaatselt morfoloogilist analüüsi, määrata alamdialooge (partneri algatatud parandussekventse ja vastuse tingimuste täpsustamise alamsekventse). Tööpinki on lõimitud ka dialoogiaktide poolautomaatse märgendamise moodul. Tööpink asub aadressil http://www.dialoogid.ee/dialoogid/.
5. Eestikeelsetes dialoogides esinevate ajaväljendite analüüsiks on töötatud välja grammatika, mis võimaldab dialoogsüsteemil vastata mitmesugustele ajaga seotud küsimustele.
6. On loodud veebis kasutatav "vestlusrobot", mis annab infot Tartu kinokavade kohta. Süsteem kasutab dialoogi juhtimiseks regulaarset grammatikat ja võtab initsiatiivi, kui kasutaja on passiivne. Dialoogsüsteem integreerib ajaväljendite automaatse tuvastamise ja dialoogi juhtimise reeglid ning tekst-kõnesünteesi. Vt http://www.dialoogid.ee/alfred/

Intelligentne kasutajaliides andmebaasidele

Vastutav täitja	Mare Koit		
Teised põhitäitjad	Mark Fišel, Olga Gerassimenko, Kristiina Jokinen, Riina Kasterpalu, Andriela Rääbis, Krista Strandson, Margus Treumuth		
Finantseerimine 2008	-	**Finantseerimine 2009**	540 000

Eesmärgid ja tähtsus

Üldistatakse lõppenud projekti "Eestikeelne infodialoog arvutiga" täitmisel saadud kogemust sellise kasutajaliidese loomiseks, mis võimaldaks hõlpsat adapteerumist erinevate ainevaldkondadega ja seostamist erinevate andmebaasidega. Liidest saab minimaalsete täienduste tegemise teel häälestada uutele ainevaldkondadele ja siduda andmebaasidega, andes seega kasutajale võimaluse pöörduda andmebaaside poole eesti keeles ning saada vastuseks adekvaatset infot. Kasutaja sisestab oma päringu eesti keeles ja saab vastuse samuti eesti keeles, tekstina või soovi korral tehiskõnes. Dialoogihalduris realiseeritakse infodialoogi juhtimise üldine mudel, mis võtab arvesse erinevates praktilistes infodialoogides kehtivad üldised seaduspärasused.

Loodavat liidest saab kasutada ka nn võlur Ozi režiimis (kus arvuti rolli mängib inimene), see võimaldab hõlpsal viisil koguda andmeid liidese häälestamiseks uuele ainevaldkonnale, s.t määramaks, missuguseid kasutaja lausungeid ja missuguseid dialoogiakte peaks intelligentne liides hiljem suutma käsitleda ning kuidas nendele reageerima. Intelligentses liideses lõimitakse olemasolevad ja/või teiste keeletehnoloogiaprojektide toel loodavad eesti keele automaattöötluse vahendid: morfoloogiline ja süntaktiline analüüs ja süntees, õigekirjakontroll ja vigaste vormide korrigeerimine, nimega üksuste (pärisnimed, ajaväljendid jms) tuvastamine, tekst-kõnesüntees, võimalusel ka kõnetuvastus.

Korpusepäring keeleveebis

Vastutav täitja	Heiki-Jaan Kaalep		
Teised põhitäitjad	Rene Prillop, Tarmo Vaino		
Finantseerimine 2008	250 000	**Finantseerimine 2009**	300 000

Eesmärgid ja tähtsus

Eesmärgiks on võimaldada eesti keele uurijatel ja eesti keelt teise keelena õppijatel nii sõnastikke kui tekstikorpusi mugavalt üle interneti kasutada.

Ühendatakse olemasolevad eesti keele ressursid – TÜ tekstikorpus, Filosofti morfoloogiline analüsaator koos ühestajaga ning sõnastikud. Morfoloogilise analüsaatori ja ühestaja kasutamine võimaldab otsida korpusest sõnu, ilma et peaks muretsema sõna muutevormide rohkuse pärast.

Praegu on TÜ tekstikorpuse kasutamine mugav inimesele, kellel on vajadused ja oskused korpust põhjalikult töödelda (nt. statistika tegemiseks): kõik tekstid on mitte-kommertskasutuseks vabalt äratõmmatavad, misjärel neid saab oma arvutis töödelda mistahes viisil. Kasutaja, kelle põhiline huvi on leida korpusest kasutus-näiteid, aga enamasti ei soovigi korpust enda arvutisse kopeerida. Tema jaoks on põhjalikumaid arvutialaseid teadmisi eeldav kasutaja-liides asjatult keeruline ja seega tõsiseks takistuseks korpuse mõistlikul kasutamisel.

Korpuse loomiseks tehtavad kulutused läheksid osaliselt tühja, kui korpuse kasutamine oleks ka edaspidi raskendatud, kuivõrd korpuse kasvav maht nõuab uusi tehnoloogilisi lahendusi.

Sõnastikud ei sisalda (juba ruumipuudusel) kogu infot, mis on seotud sõnade tähenduste ja kasutuse kõigi aspektidega. Eriti oluline võib see olla inimesele, kelle jaoks eesti keel pole emakeel. Sõnastike sidumine korpustega peaks seda probleemi leevendama.

Põhitulemused

- Tehti tasuta kättesaadavaks 5 uut erialasõnastikku.
- Loodi päring 15 miljoni sõnalisele tasakaalus korpusele, kusjuures kogu päringut realiseeriv tarkvara kirjutati nullist alates. Päring võimaldab otsida sõnu ka algvormi ja grammatiliste kategooriate järgi.

Masintõlge 1 (2006–2008)
Masintõlge 2 (2009–2010)

Vastutav täitja	Heiki-Jaan Kaalep		
Teised põhitäitjad	Mark Fišel, Kaarel Veskis, Harri Kirik		
Finantseerimine 2008	600 000	**Finantseerimine 2009**	830 000

Eesmärgid ja tähtsus
Eesmärgiks on parandada olemasoleva masintõlke kvaliteeti. Selleks on kavas kolm suunda:
1. Kasutada keelespetsiifilist tarkvara, esmajärjekorras süntaksianalüsaatoreid
2. Kasutada tagasisidet ja alternatiivide võrdlemist
3. Kombineerida erineval moel treenitud programmiversioone ja valida väljund mitme variandi hulgast.

Põhitulemused
Loodi eesti-inglise statistilise masintõlke katseversioonid; proovida saab aadressil: http://ats.cs.ut.ee/smt/translate/

Versioonid erinevad esiteks selle poolest, milline on nende aluseks oleva paralleelkorpus, ja teiseks selle poolest, kas on kasutatud eesti keele morfoloogilist analüsaatorit (ning millist täpsemalt).

Mitmete katsete tulemusena, kus prooviti erinevaid korpusi ja erinevaid morfoloogilise analüüsi viise, on selgunud, et:
1. Eestikeelsete sõnade morfoloogiline analüüs, mille käigus sõnad on tükeldatud tüvedeks ja lõppudeks, aitab kaasa õigete ingliskeelsete fraaside leidmisele ja seega ka paremale tõlkele.
2. See ei aita parandada tõlkeprobleeme, mille põhjuseks on eesti ja inglise keele erinev sõnajärg. Nende lahendamiseks tuleks kasutada süntaksi analüüsi.

Automaatne parafraaside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpuste abil

Vastutav täitja	Maarika Traat		
Teised põhitäitjad	Hendrik Nigul, Krista Liin		
Finantseerimine 2008	-	**Finantseerimine 2009**	200 000

Eesmärgid ja tähtsus

Käeoleva projekti raames on plaanis luua veebiliidesega tööriist, mis võimaldab kasutajal sisestada sõna või fraasi ning päringule vastuseks saada kas tõlked valitud võõrkeeles või parafraasid lähtekeeles. Sarnane tööriist on olemas inglise keele ja mõnede teiste enamräägitavate keelte jaoks (vt http://linearb.co.uk). Meie tahaksime luua sama funktsionaalsusega tööriista, mis hõlmaks ka eesti keelt.

Sellist tööriista saab kasutada abivahendina tõlkimisel või ühekeelse teksti kirjutamisel. Viimasel juhul on tööriist abiks parafraseerimisel. Kirjutades võib kaunis sagedasti esineda olukord, kus mingit mõtet on raske kirja panna, kuna selle väljendamiseks vajalik sõna või fraas ei tule meelde. Plaanitud tööriist aitaks sellisel puhul, kuna sarnase tähendusega sõna või fraasi sisestamisel on mõni väljastatud parafraasidest suure tõenäosusega just see vajalik puuduv sõna või fraas. Ka tõlkimisel pakub tööriist laiema diapasooniga tõlgete valikut kui tavaline sõnaraamat, kuna väljundiks on sisendsõna või fraasi tõlked paljudes erinevates kontekstides. Väljastatud tõlgete ja parafraasidega koos väljastatakse ka väike tekstilõik, mis näitab, millises kontekstis vastav tõlge või parafraas esines. Tööriista abil leitud parafraase on võimalik kasutada eesti keele tesauruse/wordneti täiendamisel, kuid tööriistast on abi ka muud sorti leksikograafilises töös.

Kirjeldatud tööriista töö põhineb joondatud paralleelkorpuste kasutamisel. Masintõlkes on selliste korpuste kasutamine väga levinud, nende kasutamine parafraaside leidmiseks on aga kaunis värske idee. Korpustena plaanime alustuseks kasutada JRC-Acquis'd (http://langtech.jrc.it/JRC-Acquis.html), Acquis Communautaire tõlkemälu DGT-TMi (http://langtech.jrc.it/DGT-TM.html) ja OPUSt (http://urd.let.rug.nl/tiedeman/OPUS/), aga samas püüame ka ise materjali juurde muretseda. Loodame tõlkebüroodelt ja riigiasutustelt tõlkemälusid saada, et need siis meie andmebaasi lisada. Äsja lõppenud EKKTT projektis Masintõlge I tehti suur töö ära olemasolevate eesti keelt sisaldavate paralleelkorpuste kvaliteedi parandamise vallas − oma projekti raames plaanime kindlasti kasutada nimetatud projekti tulemusi. Korpuste täiendamisest, mis meil plaanis on, on aga huvitatud ka sellel aastal käivituva EKKTT projekti Masintõlge II täitjad, kellega loodame koostööd teha.

Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp

Vastutav täitja	Peep Küngas		
Teised põhitäitjad	Aleksandr Tkatsenko, Martin Luts, Margus Treumuth		
Finantseerimine 2008	-	**Finantseerimine 2009**	640 000

Eesmärgid ja tähtsus
Projekti eesmärgiks on edendada tehnoloogiat süntaksi- ja semantikapõhiste infootsingute toetamiseks nii veebis kui desktop- rakenduses. Projekti vahetulemused on sisendiks eesti keele morfoloogilisele ja semantilisele analüüsile.

Kavandatavad põhitulemused
Tulemused 2009

- Eestikeelse NER tehnoloogia realisatsioon koha, tähistavate nimede ja nimetaoliste väljendite tuvastamiseks. Nimega kohaüksuste ühestamisel (disambiguation) algoritmides pööratakse tähelepanu eesti keele spetsiifikale ja konteksti/semantika arvestamisele. Eesti kohanimede ressurssidena kasutatakse mh Maa-ameti hallatavaid riiklike registreid, sh kohanimeregistrit ja keskset aadressregistrit, võimalusel ka suuremate ruumiandmete tootjate erafirmade koha- ja aadressandmeid;
- Eesti üldontoloogia esimene versioon ning NER tehnoloogia tulemuste kontseptuaalseks modelleerimiseks vajalik ontoloogia ning seosed eelpool nimetatud ning andmeteenuste semantiliseks kirjeldamiseks kasutatud ontoloogiate vahel;
- Tarkvaramoodul (TM) eestikeelse ja inglisekeelse NER tehnoloogia ja ontoloogiate sidumiseks ning nende ontoloogiatega seotud semantiliste üksuste rakendamiseks infootsingutes;
- Lõppkasutajale orienteeritud süntaksi- ja semantikapõhise infootsisüsteemi prototüüp, kus seotakse eesti ja inglise keele NER tehnoloogiad, otsimootorid/sisu pakkujad; infootsisüsteem rakendab TMi ja andmeteenuste semantilisi kirjeldusi infopäringute teostamiseks.

Tulemused 2010

- Eesti keelele kohandatud infootsingu tarkvaralahendus ning selle praktiline juurutamine lõppkasutajale suunatud infokeskkonnas (eesti.ee, delfi.ee, neti.ee, visitestonia.com vms);
- Eestikeelse NER tehnoloogia realisatsioon isiku, aja, sündmuse või organisatsiooni tähistavate nimede ja nimetaoliste väljendite tuvastamiseks.

Eesti keele koondkorpus

Vastutav täitja	Kadri Muischnek		
Teised põhitäitjad	Kristel Uiboaed, Kaarel Veskis		
Finantseerimine 2008	640 000	**Finantseerimine 2009**	673 000

Eesmärgid ja tähtsus

Korpused ehk elektroonilised tekstikogud on keetarkvara väljatöötamisel vältimatult vajalikud. Statistikapõhiste süsteemide treenimiseks vajatakse väga suuri tekstihulki, ka reeglipõhiste süsteemide testimiseks ja keelekirjeldustes varem esitamata seaduspärasustel põhinevate reeglite kirjutamiseks vajatakse tekstikorpusi. Keelekorpus on üks põhilisi keelematerjali allikaid ka inimkeele teaduslikul uurimisel.

Käesoleva projekti eesmärgiks on täita riikliku programmi „Eesti keele keeletehnoloogiline tugi“ seletuskirja punktis 3.2.1. Kirjaliku keele korpused püstitatud põhieesmärk – eesti keele koondkorpuse arendamine 200 miljoni sõnani.

Projekt jätkab riikliku programmi „Eesti keel ja rahvuslik mälu“ projekti „Eesti keele koondkorpus“ raames tehtud tööd.

Põhitulemused

Korpuse planeeritud maht – 200 miljonit sõna – on kokku kogutud. „Üleplaaniliselt“ on alustatud ka SL Õhtulehe arhiivi korpuse kujule teisendamisega, teised „suured“ ajalehed (Eesti Päevaleht, Postimees, Eesti Ekspress ja Maaleht) on korpuses juba olemas. Äramärkimist väärib nn „uue meedia“, st interneti keelekasutuse tekstide 22 miljoni sõna suurune kogu. Allkorpuste loendit ja täpsemat kirjeldust vt http://www.cl.ut.ee/korpused/

Eesti vanema kirjakeele elektroonilised kogud

Vastutav täitja	Külli Habicht		
Teised põhitäitjad	Valve-Liivi Kingisepp, Pille Penjam, Külli Prillop, Kristel Ress		
Finantseerimine 2008	200 000	**Finantseerimine 2009**	50 000

Eesmärgid ja tähtsus

Projekti põhieesmärgiks on eesti vanema kirjakeele elektrooniliste kogude täiendamine ja vana materjali laiemasse kasutusse toomine. Korpusesse sisestatakse vanu ja haruldasi tekste, millele uurijatel ja huvilistel on raske ligi pääseda ja mida võiks ilma erialaste eelteadmisteta olla keerukas keeleliselt analüüsida.

Tartu Ülikoolis tegutsev vana kirjakeele uurimisrühm on projekti raames tegelnud 18. ja 19. sajandi tekstide korpusesse sisestamise, lemmatiseerimise ja morfoloogilise märgendamisega. Tekstikorpus koos päringusüsteemiga on kasutatav uurimisrühma veebilehel (http://www.murre.ut.ee/vakkur/Korpused/korpused.htm). Korpusepäringu hõlbustamiseks on välja arendatud kasutajaliides, mida pidevalt täiustatakse. Tekstide lemmatiseerimiseks on kasutusel Külli Prillopi loodud originaaltarkvara "Vakker", mis võimaldab vanu tekste paindlikult märksõnastada.

Tegemist on eesti keele kujunemise ja kogu kultuuriloo seisukohalt olulise valdkonnaga, sest kirjakeele varasemat kasutust tundmata on raske põhjendada tänapäeva kirjakeeles toimuvaid muutusi. Vanade tekstide laiemasse kasutusse andmine võimaldab kokku hoida uurijate töövaeva ning hõlbustab diakroonilise mõõtme sissetoomist eesti kirjakeele sõnavara, vormistiku ja lausestusega tegelevatesse uurimustesse.

Põhitulemused

2008. aastal täideti nelja põhilist ülesannet:

- Prioriteetseks ülesandeks oli eesti vana kirjakeele korpuse täiendamine 19. sajandi esimese poole kirjakeele tekstidega. Korpust täiendati 400 000 tekstisõna võrra.
- Jätkati Heinrich Stahli teoste „Hand- vnd Hauszbuch" (1632–1638) ja „Leyen Spiegel" (1641–1649) lemmatiseerimist ja grammatilist märgendamist. Aasta jooksul lemmatiseeriti tekste 50 000 tekstisõna mahus.
- Tegeldi vana kirjakeele korpuse kasutajaliidese täiustamisega. Vana kirjakeele töörühma veebilehel tehti lisaks 16. sajandi tekstidele üldkasutatavaks ka 17. ja 18. sajandi korpuse terviktekstid.
- Tegeldi eesti vana kirjakeele korpuspõhise uurimisega. Olulistest töödest mainitagu Pille Penjami doktoriväitekirja „Eesti kirjakeele *da*- ja *ma*-infinitiiviga konstruktsioonid" (2008) ja tänapäeva lugejale tõlgitud mahukat allikapublikatsiooni „Georg Müller. Jutluseraamat" (2008), mille koostajateks olid Külli Habicht, Valve-Liivi Kingisepp, Jaak Peebo, Külli Prillop ja Kai Tafenau.

TÜ eesti keele tesauruse (eesti wordneti) täiendamine

Vastutav täitja	Heili Orav		
Teised põhitäitjad	Kadri Kerner, Sirli Parm, Lauri Eesmaa		
Finantseerimine 2008	390 000	**Finantseerimine 2009**	514 000

Eesmärgid ja tähtsus

1996. a. algasid esimesed katsetused eesti wordnet-tüüpi tesauruse loomiseks, tegelik töö algas 1998. a., kui osalesime EuroWordNet-2 projektis koos inglise, hollandi, itaalia, hispaania, saksa, prantsuse, tšehhi *wordnet*'idega. Praeguseks koostatakse maailmas wordnet-tüüpi tesaurusi u 50 keele jaoks.

Meie projekti eesmärk on täiendada eesti keele tesaurust ehk eesti wordneti nii kvantitatiivselt kui kvalitatiivselt, et seda saaks rakendada teiste keeletehnoloogiliste rakenduste jaoks.

Põhitulemused

Käesolevat projekti 2007. a. alustades oli eesti wordnetis ligi 15 000 mõistet (e sünohulka). Praeguseks (märts 2009) oleme suurendanud mõistete hulka u 22 500 mõisteni. Uudsena on lisatud adverbe. Palju oleme kontrollinud ja parandanud olemasolevat.

Eesti keele semantika ressursid ja vahendid

Vastutav täitja	Neeme Kahusk		
Teised põhitäitjad	Martin Luts, Siiri Pärkson, Kadri Kerner		
Finantseerimine 2008	-	**Finantseerimine 2009**	325 000

Eesmärgid ja tähtsus

1. Korrastada semantiliselt ühestatud ja märgendatud (ühestatud sõnatähendustega) korpus ja suurendada seda 500 000 sõnani (Riikliku programmi alaeesmärk 2.2.4).

Ühestatud sõnatähendustega korpus sisaldab praegu 100 000 sõna. Sõnatähenduste ühestamist on tehtud mitmes järgus ja kasutatud on erinevaid eesti wordneti versioone. Olemasolev osa korpusest viiakse vastavusse viimase versiooniga eesti wordnetist ja töötatatakse välja vahendid kindlustamaks selle korpuse vastavust viimasele wordneti versioonile. Korpust suurendatakse 500 000 sõnani.

2. Luua sõnatähenduste käsitsi ühestamist hõlbustav tarkvara (Riikliku programmi alaeesmärk 2.1.9)

Olemasolev osa ühestatud sõnatähendusega korpusest on loodud ilma spetsiaalset tarkvara kasutamata: leksikograafid redigeerivad puhast teksti, kirjutades sinna vajalikku kohta sõnatähenduse numbri. Projekti käigus luuakse sõnatähenduste ühestamist hõlbustav tarkvara, mis võimaldab tõsta ühestamise kiirust ja parandada selle kvaliteeti.

3. Luua leksikaal-semantilise andmebaasi loomist ja haldamist hõlbustav tarkvara (Riikliku programmi alaeesmärk 2.2.12).

Eesti wordneti tegemiseks kasutatakse praegu programmi "Polaris", mille toetus on lõppenud seoses tootjafirma pankrotistumisega 1999. aastal. "Polarisega" eri aegadel ja erinevates arvutites tehtud wordneti kirjetes on kasutatud erinevaid kooditabeleid, see teeb koondbaasi kokkuajamise raskeks. "Polaris" töötab ainult MS Windows platvormil, tema edasine kasutamine tulevaste Windowsi versioonidega ei ole garanteeritud. Konfigureerimisvõimaluste piirratuse ja lähtekoodi puudumise tõttu ei ole võimalik lisada uusi semantilisi suhteid ega muud informatsiooni, mis töö käigus võib vajalikuks osutuda.

Käesoleva projekti käigus luuakse programm, mis täidaks samu funktsioone, mis "Polaris", kuid oleks (1) avatud lähtekoodiga (2) kasutatav mitmel platvormil (3) ulatuslikumalt konfigureeritav.

4. Üldontoloogia ja valdkonnaontoloogiate-põhise mitmekeelse infootsingu raamistiku, ressursside ja rakenduste loomine (Riikliku programmi alaeesmärgid 2.1.6 ja 2.1.13)

Üldontoloogia ja valdkonnaontoloogiate-põhise mitmekeelse infootsingu raamistiku, ressursside ja rakenduste tulemid on kasutatavad semantilistes otsimootorites, infoportaalides (nt neti.ee, google.ee) kui ka mitmekeelsust nõudvates valdkonnaspetsiifilistes infootsingutes (nt turismivaldkonnas: visitestonia.com, tallinn2011 kultuuripealinna raames, expo2010 eestit tutvustavas portaalis; nt euroopa liidu infosüsteemide semantilise liidestamise programmis IDABC, semic.eu jt). Selle ülesande täitmiseks (1) valitakse sobiv ontoloogiakeel, (2) luuakse ontoloogia ressursid (üldontoloogia ja mõned valdkonna-ontoloogiad) (3) luuakse ontoloogiat kasutav näidisrakendus.

Leksikograafi töökeskkond

Vastutav täitja	Ülle Viks		
Teised põhitäitjad	Andres Loopmann, Indrek Hein, Sven-Olav Paavel, Ain Teesalu. Seotud isikud: Margit Langemets, Kaur Männiko		
Finantseerimine 2008	1 570 000	**Finantseerimine 2009**	1 980 000

Eesmärgid ja tähtsus

Projektil on kolm põhieesmärki:

1. Luua leksikograafidele sobiv interaktiivne töökeskkond e sõnastike haldussüsteem EELex, st töövahendid, mis ühilduvad kehtiva rahvusvahelise märgistusstandardiga (XML) ja rakendavad nii universaalseid kui ka eesti keele põhiseid keeletehnoloogia vahendeid: keeleressursse ja keeletarkvara.
2. Koostada eesti lähtekeele andmebaas uute kakskeelsete sõnaraamatute jaoks ehk Eesti–X-keele sõnastik.
3. Anda projekti tulemused avalikku kasutusse: (a) süsteemi kuuluvate sõnastike avalikud veebiversioonid, (b) sõnastike haldussüsteemi laiatarbeversioon.

Leksikograafi töökeskkond EELex muudab sõnastikutöö lihtsamaks, kiiremaks ja kvaliteetsemaks. EELexis koostatud või sinna üle viidud sõnastikud on standardse märgendusega universaalsed taaskasutatavad keeleressursid, mida vajavad nii leksikograafid ja keeletehnoloogid kui ka tavakasutajad.

Põhitulemused

1. Töökeskkonna arendus

1.1. EELex on saanud juurde uusi funktsioone: lisatud on sõnaartikli ja sõnastiku tööriistad, uued päringuvõimalused, automaatne klaviatuurivahetus vastavalt keelele, XML failide struktuurianalüüsi vahendid.

1.2. Süsteemiga on integreeritud reeglipõhise morfoloogia tarkvara.

1.3. Lisandunud on uued rakendustüübid koos vastavate funktsioonidega: suur ükskeelne üldsõnastik ja terminoloogiasõnastikud.

1.4. Sõnastikurakendused. Praeguse seisuga on EELexi süsteemi kasutades valminud 4 sõnastikku, tööversioonina on leksikograafide käsutusse antud 11 sõnastiku haldussüsteemid, testimisel on 2 uue sõnastiku haldussüsteemid ja eeltöötluse faasis on 3 sõnastiku andmebaasid.

2. Eesti-X-keele sõnastiku andmebaas

Jätkunud on Eesti-X keele sõnastiku (EXS) edasiarendamine. Väikese mahuga kakskeelse sõnastiku tarbeks on tehtud märksõnavalik (u 20 000 üksust). Käsil on EXS-i toimetamine EELex-i vahenditega: liitsõnamaterjali toimetamine, homonüümide korrastamine jne.

3. Avalik kasutus

3.1. Sõnastike veebiversioonid. Leksikograafi töökeskkonnas kehtivate XML standardite alusel on loodud avalik veebiversioon Eesti kirjakeele seletussõnaraamatust (esialgu on ligipääs parooliga, avalik sügisest). Täiendatud on Õigekeelsussõnaraamatu struktuuripõhist päringut:
(http://www.eki.ee/dict/QS2006.tegemisel/full.html).

3.2. EELexi tarkvara. Loodud on esialgne demoversioon avalikust Eesti-X keele sõnastikust (http://exsa.eki.ee/), mis annab kasutajale võimaluse koostada oma kakskeelne sõnastik EELexi keskkonnas.

Lihtlause semantiline analüüs 1

Vastutav täitja	Haldur Õim		
Teised põhitäitjad	Heili Orav, Neeme Kahusk, Kadri Vider, Piia Taremaa, Kaarel Kaljurand		
Finantseerimine 2008	650 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus
Projekti eesmärgiks oli lause semantilise analüüsi modelleerimine ja võimaluste piires vastava arvutiprogrammi loomine; selleks vajalike keeleressursside loomine.

Põhitulemused
1) semantiliseks analüüsiks vajalikud ressursid, milleks on 1a) freimileksikon ja 1b) semantiliselt märgendatud korpus;
2) semantilise analüüsi tarkvara, mille olulisimaks esindajaks praeguse seisuga on semantilisi järeldusi tegev programm.

1a) **Freimileksikon**
Freimileksikoni moodustavad freimid. Freim on struktuur, mis esitab meid huvitava üksuse - tüüpiliselt predikaadi - tähendust. Selleks on selle kirjes fikseeritud hulk "auke", "slotte". Nendeks slottideks on semantilised rollid (nt Agent, Instrument), mida saab konkreetsetel juhtudel täita konkreetse keelematerjaliga.
Freimid on organiseeritud hierarhiliselt: freimi peasõnal on ülemmõiste, millele vastab oma freim, ja nii kuni teatud absraktsete kategooriateni välja. Nt "jooksma" abstraktne ülemfreim on AGENTIIVNE LIIKUMINE. Teised samalaadsed kategooriad on mitteagentivne liikumine, agentiivne liigutamine ja mitteagentiivne liigutamine. Neid abstraktseid kategooriaid on vaja näiteks järelduste tuletamise programmis, mille ülesandeks on „tunda ära“, missugused sündmuse osalised liiguvad, ja vastata küsimusele „kus on X pärast liikumissündmust?“. Nii liigub agentiivse liikumise puhul obligatoorselt Agent (aga liikuda võib ka nt Instrument, kui see olemas on), mitteagentiivse liikumise puhul liigub Objekt (nt pall veeres...); agentiivse liigutamise puhul liigub Objekt (kas ka Agent ja Instrument, sõltub tegevusest e verbist , vrd viskama ja viima), mitteagentiivse liigutamise puhul liigub igal juhul Objekt (Tuul lükkas vaasi põrandale).

1b) **Semantiliselt märgendatud korpused**
Projekti keeleainestiku lähtekohaks on nn Rätsepa korpus. Korpus on analüüsitud ja ühestatud morfoloogiliselt ning süntaktiliselt, ühestatud on ka verbide ja noomenite tähendused. Korpus on teisendatud TIGER-XML kujule. Sisaldab 388 lauset. Lisaks Rätsepa korpusele on kasutuses ka teine korpus, mis on suuremas osas pärit Tartu Ülikooli korpuste kogust (http://www.cl.ut.ee/korpused/). See korpus on ühestatud morfoloogiliselt ja süntaktiliselt, samuti lause põhiverbide tähenduse osas. Ka see korpus on teisendatud TIGER-XML kujule. Sisaldab 736 lauset.
Frameneti freimid
Lisaks liikuma ja liigutama freimile oleme märgendanud korpustes ka freime Frameneti järgi (http://framenet.icsi.berkeley.edu). See võimaldab (1) võrrelda eesti freime Frameneti omadega, (2) siduda hiljem eestikeelseid lauseid samade lausetega mingis teises keeles, (3) kasutada juba Frameneti jaoks loodud töötlusvahendeid.

2) **Semantilise analüüsi tarkvara**

Järelduste tegemise programm (Kaarel Kajurand) on 2008. a põhitulemusi. Pogramm ("lexicon.pl") ise on saadaval leheküljelt www.keeletehnoloogia.ee/. Programmi sisuks on asukohamuutust kirjeldavate järelduste automaatne genereerimine freimidega märgendatud tekstikorpuse põhjal. Loodud programm võtab sisendiks semantiliste freimidega märgendatud tekstikorpuse ning genereerib iga korpuses sisalduva lause kohta selles lauses esinevate freimielementide (nt "Agent", "Object", "Instrument") asukohamuutuse info, esitatuna freimielementide "Locfrom" ja "Locto" kaudu, kujul

"Locfrom - >Locto". Programmi töö järelduste tegemisel põhineb täielikult leksikonil, mis koosneb Prolog-keeles kirjapandud reeglitest. Reeglid määravad, kas etteantud tüübiga freimielement (nt "Agent", "Instrument", "Object") muudab asukohta etteantud tüübiga freimis (nt "agentiivne liikumine", "miteagentiivne liigutamine", vt eespool).

Lihtlause semantiline analüüs 2

Vastutav täitja	Haldur Õim		
Teised põhitäitjad	Heili Orav, Neeme Kahusk, Piia Taremaa, Siim Orasmaa		
Finantseerimine 2008	-	**Finantseerimine 2009**	640 000

Eesmärgid ja tähtsus
See projekt on jätkuprojektiks projektile "Lihtlause semantiline analüüs", 2006-2008, ning põhieesmärgiks on riikliku programmi järelejäänud 2 aasta jooksul lahendada ülesanded, mis eelmise projekti käigus jäid poolikuks ning millede lahendamine tekitaks teatud tervikmudeli - nii kontseptuaalses kui tehnilises mõttes - eesti keele lihtlause ja hiljem liitlausete või ka teksti semantilise analüüsi alusena.

Põhitulemused (plaanina)
1. Lause prediaadist sõltuvate argumentide semantiliste rollide automaatne määratlemine. Lause argumentstruktuuri saame süntaktilise analüüsi mooduli väljundist, kuid semantilisele struktuurile üleminekuks tuleb argumentidele omistada semantilised rollid, mida nad vastavas sündmuses täidavad: nt predikaadiga 'viima' moodustatud lauses võib olla Agent, Object, Instrument, Locfrom (kust), Locto (kuhu) jne. Rollide automaatseks määratlemiseks tuleb meid huvitava predikaadi iga võimaliku rolli juures esitada vähemalt kaht liiki inormtsiooni: 1) rolli võimaliku täitja semantiline iseloomustus (eelkõige semaniline kategooria/ kategooriad, nt Elusolend, Füüsiine Objekt, Vedelik jne, aga tihti on vaja ka lisainfot, nt mitte iga Füüsiline objekt ei saa veereda, oluline on objekti kuju jne) ja 2) rolli täitva väljendi morfoloogiline iseloomustus, nimisõnade puhul nt võimalikud käänded, kaassõnad. 2. Lausest järelduste tegemise programmi täiustamine. Lause tähendus ei piirdu teatavasti selles vahetult öelduga, sellesse kuulub ka järelduv info (kui teadmine, mille lause vastuvõtja saab ja millega arvestab kui kehtivaga). Eriti oluliseks muutubki see aspekt, kui üksiklihtlausetelt minna edasi liitlausete ja lõpuks eriti teksti analüüsi juurde. Et meie valdkonnaks on liikumis- ja liigutamissündmused, siis tuvastab praegune järeldusprogramm lauses need entiteedid, mis liiguvad (erinevate predikaatide puhul on need erinevad) ja fikseerib liikuva entiteedi asukoha pärast liikumissündmust. Kuid järelduste kaudu on võimalik tuvastada muudki infot, muudel alustel. Nii on võimalik ühe lause poolt esitatava sündmuse raames teha ühtede rollide esinemise põhjal järeldusi teiste rollide obligatoorsuse, võimalikkuse või võimatuse kohta (mis arvutianalüüsis on rollide määratlemises vägagi oluline piiranguid seadev info). Näiteks on omavahel seotud Agendi (kui tahtliku tegija), Goal-i (eesmärgi) ja Instrumendi rollid: kaks viimast eeldavad esimese olemasolu. Teine tüüp meid huvitavaid järeldusi seostub vastavas 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 puhul "Poiss viskas kivi õhku" selline järeldussituatsioon ei kehti. Niisugune sõnade, lausete, eriti aga teksti maailmateadmusliku/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.

Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid

Vastutav täitja	Tiit Roosmaa		
Teised põhitäitjad	Kaarel Kaljurand, Kaili Müürisep, Helen Nigol		
Finantseerimine 2008	500 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus
Projekti eesmärgiks oli luua vajalik baas süntaksil ja semantikal põhineva keeletehnoloogilise tarkvara loomiseks. Selleks on vaja pind- ja süvasüntaktiliselt märgendatud treening- ja testkorpusi, mis sisaldavad erinevatesse tekstiliikidesse kuuluvaid tekste (ilukirjandus, ajakirjandus, juriidiline keel, teaduskeel, suuline kõne). Seejuures grammatikakorrektori arendamiseks on vaja nii grammatiliselt korrektsete tekstide korpust kui grammatiliselt vigastest lausetest koosnevat korpust (viimane peaks sisaldama esinduslikku valimit inimeste poolt tehtavatest grammatikavigadest). Samuti oli eesmärgiks luua eesti keele grammatikakorrektori ja eestikeelsete tekstide sisukokkuvõtja prototüübid.

Põhitulemused
Loodi järgmised keeletarkvarasüsteemide prototüübid:

1. Grammatikakorrektor, mis keskendub komavigade tuvastamisele, kasutades selleks kitsenduste grammatika parserit. Loodi 100 reeglist koosnev komavigade tuvastamise grammatika, mille abil on komavigade leidmise täpsus 93,8%, saagis 94,1%.

2. Eestikeelsete tekstide sisukokkuvõtja, mis töötab ekstraktori põhimõttel valides lähtetekstist välja need laused, mis tõenäoliselt sisaldavad kõige olulisemat informatsiooni. Lause olulisuse määramiseks arvestatakse lause positsiooni tekstis, tema šrifti, lauselõpumärki, samuti lauses esinevate sõnade statistikat. Projekti raames on loodud ka anafooride tuvastamise süsteem isikuliste asesõnade tarbeks.

Loodi järgmised keeleressursid:

1. Automaatseks analüüsiks vajalikud grammatikad: teisendati T. Puolakaineni loodud (Puolakainen 2001) morfoloogilise ühestaja reeglid ning pindmise süntaksi-analüsaatori reeglid uue VISL parseri formaati (1509 + 1130 reeglit). Lisaks loodi sõltuvusstruktuuri ehitav grammatika (50 reeglit), mille abil on võimalik leida enamiku lausete osaline puustruktuur. Projekti käigus tehti katseid ka fraasistruktuuri-grammatika kirjutamiseks, mis siiski ei sobi keerulisemate infiniittarinditega lihtlausete ja liitlausete analüüsiks.

2. Korpustest on olulisemad pindsüntaktiliselt märgendatud korpus ning puudepank.

Pindsüntaktiliselt märgendatud korpus koosneb eesti ilukirjandusest, tõlke-kirjandusest, verbirektsioone kirjeldavatest lihtlausete korpusest, ajakirjandus-tekstidest, seadusetekstidest, suulise keele tekstidest ja murdekeeletekstidest, kokku 450000 sõna.

Kogu käsitsi märgendatud puudepank koosneb hetkel: 388 liikumisverbiga lihtlauset Rätsepa korpusest; 732 liikumisverbiga lauset eesti frameneti testkorpusest; 175 lauset Arboresti korpusest; 20 lauset suulise keele korpusest.

Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs

Vastutav täitja	Tiit Roosmaa		
Teised põhitäitjad	Kaili Müürisep		
Finantseerimine 2008	-	**Finantseerimine 2009**	480 000

Eesmärgid ja tähtsus

Automaatne süntaktiline analüüs on vajalik paljudele keeletehnoloogilistele rakendustele, alustades automaatsest grammatikavigade tuvastajast ning lõpetades dialoogsüsteemide ja masintõlkega.

Süntaktilise analüüs mõiste on väga lai, kuid eesti keele kontekstis hõlmab see traditsiooniliselt lauseliikmete funktsiooni kindlaksmääramist. Vähem tuntud on morfoloogilise ühestamise seostamine süntaktilise analüüsiga: sõna kontektsi sobiva morfoloogilise tõlgenduse valimine kõigi võimalike seast (nt kas sõnavorm *ilma* on nimi-, määr- või kaassõna). Inimene teeb seda kuulates või lugedes instinktiivselt, kuid algoritmiliselt on see küllaltki keeruline probleem.

Eesti keele jaoks on loodud nii pindsüntaktiline analüsaator kui ka reeglipõhine morfoloogiline ühestaja. Pindsüntaktiline analüsaator leiab 90protsendilise täpsusega iga sõna süntaktilise funktsiooni lauses, kuid ei leia sõnade omavahelisi täpseid seoseid ega lausestruktuuri. Nt. eestäiendina esinev sõna saab küll eestäiendi märgendi, kuid ei täpsustata, millist sõna ta täiendab.

Pindsüntaktilise analüüsi reeglid arvestavad nii kirjaliku kui suulise keelega, esimesed katsed on tehtud ka murdekeelsete tekstidega. Morfoloogilise ühestaja reeglid on loodud ainult kirjaliku keele automaatse analüüsi jaoks.

Projekti eesmärgiks on olemasolevale morfoloogilisele ühestajale ja pindsüntaktilisele analüsaatorile tuginedes luua:

1. Grammatikakorrektori tööversioon: kohandada grammatikareegleid mittekorrektse sisendi analüüsiks, kirjutada tüüpiliste grammatikavigade tuvastamise reegleid, luua liides mõne vabavaralise tekstiredaktoriga.
2. Suulise keele süntaksianalüsaatori arendamine: kohandada morfoloogilise ühestamise reeglid suulise keele ühestamiseks. See võimaldab poolautomaatselt analüüsida suulise keele korpust ning teha katsetusi automaatse kõnetuvastuse väljundi edasise analüüsiga.
3. Murdetekstide süntaktiline analüüs. esialgsed katsed on näidanud, et suulise keele pindsüntaktilist analüsaatorit on kerge kohandada murdekorpuse tekstide pindsüntaktiliseks märgendamiseks (mitmesus 10%, vigu 3-5%).
4. Interneti keele (uue meedia keele) süntaktiline analüüs: kombineerides suulise ja kirjaliku keele analüsaatorit ning lisades internetis kasutatava keele omapära arvestavad reeglid on võimalik internetis leiduvaid spontaanseid tekste (foorumid, kommentaarid, Skype'i vestlused, jututoad) automaatselt analüüsida.
5. Õppijakeele süntaktiline analüüs: grammatikakorrektori arendamisega samaaegselt on võimalik luua eesti keelt võõrkeelena kõnelejate tüüpvigade tuvastajat.
6. Sügavamate sõltuvusseoste tuvastamine: luua grammatika, mis püüab leida sõnadevahelised grammatilised seosed ilmutatult. See on vajalik sügavamat süntaktilist analüüsi vajavate rakenduste loomiseks ning ka semantiliseks analüüsiks.

Eesti fraseologismide elektroonilise alussõnastiku loomine

Vastutav täitja	Katre Õim		
Teised põhitäitjad	Asta Õim, Dagne Kivisild		
Finantseerimine 2008	200 000	**Finantseerimine 2009**	200 000

Eesmärgid ja tähtsus

Projekti eesmärk on toota eesti fraseologismide mõisteline alussõnastik elektroonsel kujul. Internetis üldkasutatavas sõnastikus on eri mõistetega seonduvad fraseologismid koondatud mõisteartiklitesse; selles antakse teavet fraseologismide ehituse ja kasutuse kohta loomulikus keeles. Projekt toetub Eesti kõnekäändude ja fraseologismide andmebaasile (EKFA, http://www.folklore.ee/justkui) ja Eesti kõnekäändude ja fraseologismide mõistestikule (http://www.folklore.ee/justkui/moiste.php).

Kõrvuti institutsionaalsuse, leksikogrammatilise püsivuse ja mittekompositsioonilisusega on fraseologismidele tunnuslik suur süntaktiline ja leksikaalne paindlikkus. Paraku see varieerumine eesti keeles ilmunud üld- ja erisõnastikes ei kajastu – fraseologismist jääb enamasti klišeeline mulje. Fraseologismide esitamisele formaalsete tunnuste põhjal pakub alternatiivi kognitiivsest keeleteadusest inspireeritud mõistepõhine lähenemine esitada fraseologismid erisõnastikus mõisteseoste alusel.

Eesti fraseologismide elektroonilises alussõnastikus esitatavad tänapäevaselt märgendatud (XMLis) keeleandmed ja grammatiline info moodustaksid eesti lähtekeelega tõlkesõnaraamatute fraseoloogiapoole aluse. Statistiliselt esindusliku materjali analüüs võimaldab selgitada nii ühe- kui ka mitmesõnaliste fraseoloogiliste üksuste tähenduse kujunemist, kujunditekkemehhanismi arenemist ja muutumist eelkõige diakroonilisest aspektist. Sõnastiku mõistepõhine esitusviis aitab kokku viia semantiliselt seotud fraseologismid: sünonüümid, antonüümid, osa-tervik jm leksikaalsetes suhetes fraseologismid.

2008. a põhitulemused

1. Välja on töötatud eesti fraseologismide elektroonilise alussõnastiku makrostruktuur Sõnastiku ülesehitus on mõistepõhine, selle põhiüksused on 998st põhitasandi mõistest lähtuvad artiklid.
2. Välja on töötatud eesti fraseologismide elektroonilise alussõnastiku mikrostruktuur. Mõisteartikli sees korraldatakse fraseologismid vastavalt nende ehitusele. Esmane eristus järgib seda, kas fraseologismid väljendavad lausega tähistatava situatsiooni komponente või on need laused. Lause moodustajate puhul tehakse vahet eri tüüpi fraasidel, lausete puhul erinevaid suhtluseesmärke kandvatel väljenditel. Iga fraseologismi levikuandmed esitatakse kihelkonna täpsusega. Iga fraseologismi kasutusnäited reastatakse vastavalt väljendi formaalsele varieerumisele loomulikus keeles. Kasutusnäited esitatakse lihtsustatud transkriptsioonis.
3. EKFA 35000st keelendist on eesti fraseologismide elektroonilise alussõnastiku jaoks esialgu välja valitud 20 671 fraseologismi.

VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine

Vastutav täitja	Pille Eslon		
Teised põhitäitjad	Erika Matsak, Helena Metslang, Vahur Rebas, Annekatrin Kaivapalu, Anne Kostenko		
Finantseerimine 2008	350 000	**Finantseerimine 2009**	525 000

Eesmärgid ja tähtsus

Eesti vahekeele korpus (http://evkk.tlu.ee) on eesti keele kui teise keele või võõrkeele kasutajate autentsete kirjalike tekstide kogu. Korpus on loodud Tallinna Ülikoolis arendatava vabavaralise veebitarkvara baasil; lähtekood on jagatav BSD litsentsi alusel. Korpuse veebipõhine kasutajaliides võimaldab seada eri tasandi kasutajatele erinevaid juurdepääsupiiranguid, kuid enamasti on korpuse funktsionaalsused vabalt kasutatavad. Korpusel on oma konkordantsileidja, sõna- ja vormisageduse statistika. Õppijakeele käsitsi märgendatud vigu saab näha vealiikide kaupa (leksikaalsed, leksikaalgrammatilised, morfonoloogilised, morfoloogilised, morfo-süntaktilised, süntaktilised, kommunikatiivsed), nii kitsamas kontekstis kui terviktekstis. Kasutajaliidese alusel saab metainfot teksti koostaja (sugu, emakeel, vanus jne) ja teksti kohta (teksti maht sõnedes, tekstiliik, vigade hulk tekstis vealiigiti jne). Projekti eesmärk on olemasoleva keeletarkvara alusel ja seda arendades luua Eesti vahekeele korpuse automaatseks töötlemiseks sobivad tarkvararakendused, mis võimaldavad käsitsimärgendamiselt üle minna automaatsele. Selleks on vaja luua vealeidja prototüüp, mis sisaldaks ühelt poolt morfoanalüsaatorit ja kitsenduste grammatika süntaksianalüsaatorit ja teisalt lähtuks vealiigi määramisel korpuse lingvistilisest veataksonoomiast ja veebisõnastikust. VAKO eesmärkide realiseerimine oleks samm vahekeele kui eesti kirjakeele variandi võrdsete uurimisvõimaluste suunas võrreldes olemasolevate inglise ja mitte-inglise vahekeele korpustega.

2008. a põhitulemused

1. Eesti vahekeele korpuse keeletarkvara arendamisel keskenduti tööle, mis on suunatud automaatse vealeidja järkjärgulisele loomisele ja korpuse lemmatiseerimisele. EVKK vormisageduse veebisõnastik, mille põhjal on alustatud korpuse lemmatiseerimisega. Selleks on välja töötatud programm, mis võimaldab EKI lemmatiseerijat kasutades määrata nii õigesti moodustatud vorme kui siduda lemmaga ebareeglipärased ja raskesti mõistetavad vormid. Tulemus: EKI lemmatiseerija tuvastas ligikaudu 70000 sõnest üheselt ära pooled. Uuendatud on korpuse tarkvara.
2. Vealeidja loomisel alustati sagedasemast vealiigist – sõnajärjest. Tundsime huvi öeldise, aluse, sihitise, öeldistäite, määruse vastu seotud laiendina või väljendverbi liikmena, vaatlesime nende järjestust vastavalt Eesti keele käsiraamatu sõnajärje reeglitele. Vaatluse alt on välja jäetud nt kõik täienditüübid, omadussõna laiendavad määrsõnad, rõhumäärsõnad ja sidesõnad. Analüsaatori loomise vahenditena kasutasime Eesti vahekeele korpust, TÜ süntaktiliselt ühestatud ilukirjanduskorpust (vt http://lepo.it.da.ut.ee/~heli_u/SA/), Kaili Müürisepa kitsenduste grammatika analüsaatorit. Sõnajärje analüüsiks töötasime välja abivahendeid, mis on suunatud õigete/valede morfoloogiliste koosluste väljaselgitamisele (makrosid ja filtreid sisaldavat märgendamise vahendit Excelis). Sõnajärge uuriti ka tajutesti abil.

Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid

Vastutav täitja	Kristiina Praakli		
Teised põhitäitjad	Neeme Kahusk, Kadri Sõrmus, Tiina Kikerpill		
Finantseerimine 2008	200 000	**Finantseerimine 2009**	320 000

Eesmärgid ja tähtsus
Tartu Ülikooli eesti keele (võõrkeelena) osakonna õppijakeele korpus on õppijakeele elektrooniline kogu, mis sisaldab Tartu Ülikoolis eesti keelt teise keelena või võõrkeelena õppivate üliõpilaste loodud autentseid eri liiki kirjalikke tekste. Õppijakeel on keelevariant, mida õppija keeleõppeprotsessis loob.

Õppijakeele korpus on loodud kahel eesmärgil:
1) luua andmebaas, mis pakub autentset keelematerjali õppijakeele kirjaliku keelekasutuse uurimiseks;
2) arendada õppijakeele korpuse baasil välja eesti keele (võõrkeelena) õppimist toetav veebikeskkond.

Põhitulemused
Tartu Ülikooli eesti keele (võõrkeelena) osakonna elektoonilised kogud:
1) Paralleelkorpus (2006–2007)
Sisaldab vigaseid üksiklauseid (kokku 9000), sõnadena 128 000 sõna. Iga vigase lause juurde on lisatud parandatud lause/laused (kokku 9100). Iga vealause juurde kuulub keelekasutaja profiil järgmiste andmetega: sugu, emakeel, elukoht, keeleoskuse tase.

2) Tekstikorpus (alates 2008)
2008. a alustati uue töörühmaga (Kristiina Praakli, Neeme Kahusk, Kadri Sõrmus) täiendavat keeleainestiku kogumist uutest kogumisprintsiipidest lähtuvalt. Eesmärgiks on koguda terviktekste, mis võimaldavad näha ja analüüsida vea konteksti tervikuna. Kogutava materjali aluseks on mitte-eestlastest üliõpilaste (emakeel vene, soome või hispaania keel) kirjalikud tööd.

Tekstiliigid:

a) kodukirjandid	133 483 sõna
b) eksamiesseed	6110 sõna
c) üliõpilaste e-kirjad	2523 sõna
d) referaadid, lõputööde sissejuhatused ja kokkuvõtted jms	97488 sõna
e) praktikapäevikud	11 666 sõna
Kokku	u 260 000 sõna

Mitmesõnaliste verbide ja nende kokku-lahku kirjutamise vigade äratundmine eestikeelsetes tekstides

Vastutav täitja	Heiki-Jaan Kaalep		
Teised põhitäitjad	Kadri Muischnek		
Finantseerimine 2008	500 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus

Mitmesõnalised verbid on üks püsiühendite alaliike. Püsiühendi all mõeldakse kahe või enama sõna(vormi) ühendit, mida mingi tähenduse väljendamiseks on tavaks koos kasutada; selle definitsiooni alla mahuvad nii idiomaatilised kui ka kollokatiivsed ühendid. Keeletehnoloogias on püsiühendid probleemiks, sest nad komplitseerivad teksti analüüsimudelit, mille järgi lause struktuuri ja tähenduse ehituskiviks on üksiksõna. Püsiühendite automaattöötlus nõuab nende määratlemist-piiritlemist, mis omakorda tähendab ka püsiühendite loendite ja leksikonide koostamist ning püsiühendite tekstis esinemise vormide kirjeldamist ja tekstikorpuses märgendamist.

Põhitulemused

On loodud kolm omavahel seotud keeleressurssi:

1. Verbikesksete püsiühendite andmebaas
2. Kolmest erinevast tekstiklassist (ilukirjanduse, ajakirjanduse ja Horisondi tekstid, igaühes 100 000 sõna) koosnev morfoloogiliselt ühestatud tekstikorpus, milles verbikesksed väljendid on märgendatud.

Andmebaasis on kirjas ka väljendi kasutussagedus ülalnimetatud tekstikorpuses; korpuses pole ühtegi väljendit, mida poleks ka andmebaasis.

3. Programm, mis ülalnimetatud andmebaasi ja morfoloogiliselt ühestatud korpuse alusel märgendab verbikeskseid väljendeid. Programmi sisendiks on morf. ühestatud korpus. Nii programmi täpsus kui ka saak on 90%.

Reeglipõhine keeletarkvara

Vastutav täitja	Jan Villemson		
Teised põhitäitjad	Jaak Pruulmann-Vengerfeldt		
Finantseerimine 2008	150 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus

Projekti üldine eesmärk oli Eesti Keele Instituudis välja töötatud morfoloogiatarkvara ja -reeglite taaskasutatavaks muutmine ning värskendamine. EKIs kirjeldatud reeglite rakendamiseks ja kasutamiseks loodud programmid on moraalselt vananenud ning uutele (sh vabavaralistele) süsteemidele raskesti kohandatavad.

Konkreetselt oli plaanis luua vahendid, mille abil reegleid täiendada ning hiljem morfoloogilist analüüsi ja sünteesi vajavates praktilistes süsteemides otse kasutada või kasutamiseks teisendada saaks. Senisest suuremat rõhku kavatseme panna ka liitsõnadega seotud reeglite kirjeldamisele ja töötlemisele.

Esimese praktilise süsteemina on plaanis luua vabavaraline spellerimootor, mis töötaks kasutades morfoloogilist analüüsi (ja mitte sõnastikuvõrdlust nagu teised vabad spellerimootorid tavaliselt teevad). Plaanis on saadavate tulemuste kasutamine ka teistes EKKTT projektides.

Usume, et reeglikomplekti ja selle töötlemiseks vajalike vahendite sõltumatutele arendajatele kättesaadavaks tegemine võimaldab luua uusi ja huvitavaid keeletehnoloogilisi rakendusi, mille loomisel on seni olnud takistuseks eesti keele keeruline morfoloogia.

Põhitulemused

Peamine projekti käigus tekkinud uus asi on universaalne reegliformaat ning tarkvara selle kasutamiseks. Reegliformaadi mõte on esitada sõnade muutumise kõiki olulisi aspekte selleks, et kasutades keelekirjeldust kompileerida konkreetseid praktilisi rakendusi.

Kahjuks ei kata loodud tarkvara veel kogu ahelat keelereeglite esitusest kuni praktilise rakenduseni spellerimootori või vastava sõnastikuna, kuid tarkvara on vabalt laiendatav ning projektimeeskonnal on huvi ja kavatsus seda ka pärast projekti lõppu arendada, et spellerimootorini siiski jõuda.

Kõrvaltulemusena tekkis komplekt programme, mis suudavad EKI olemasolevat keelekirjeldust universaalsele kujule teisendada.

Projekti käigus loodud programmid on saadaval avalikust koodihoidlast github:
git://github.com/jjpp/reeglitega-keel.git

Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid

Vastutav täitja	Jaak Vilo		
Teised põhitäitjad	Siim Orasmaa, Kristo Tammeoja, Reina Käärik, Deniss Sudak BIIT rühm: http://biit.cs.ut.ee/		
Finantseerimine 2008	400 000	**Finantseerimine 2009**	-

Eesmärgid ja tähtsus

Struktureerimata ja struktureeritud tekst ehk vaba- ja märgendatud (XML) tekst moodustab tänapäeval ühe mahukama ja olulisema infokoguse. Otsimine Internetist või spetsiifilistest tekstiandmebaasidest nagu tekstikorpused, teatmeteosed jne on meie igapäevase töö osa. Projekti eesmärk oli välja töötada metoodikaid sõnastike ja teatmeteoste märgendamiseks ja tekstiandmebaaside jaoks ettevalmistamiseks ning lihtsate ja kasutajasõbralike päringusüsteemide välja töötamine tekstiandmebaaside ja teatmeteoste ning sõnastike jaoks.

Põhitulemused

Projekti käigus töötati välja üks struktureeritud sõnastike andmebaasi genereerimise vahend, milles vastavalt kasutaja sisestatud märgendustele kujunes automaatselt sõnastiku märgenduse struktuur (keel millele peavad vastama märgendused). Deniss Sudaki loodud tööriista prototüüp võimaldas veebis redigeerida sõnastikku ning tuletada sõnastikule XML märgenduse struktuuri. Sõnastike keskkonna arendamine toimub nüüd Eesti Keele Instituudi Ülle Viksi juhitud töörühmas. Kuna keerulisema struktuuriga sõnastike XML ja vastav visuaalne küljendus on küllalt keerulised, siis jõudsime projekti käigus ka tulemusele, et küljendus tuleks genereerida otse sõnastike koostamise keskkonnas, kui võimalik, ning päringusüsteemide ja avalike veebiliideste jaoks saab kasutada valmis küljendatud kirjeid koos kirje aluseks oleva XML lõiguga.

Teine oluline algoritmiline väljakutse projektile oli arendada ligikaudset sobitamist võimaldavaid meetodeid. Ehk siis selliseid, kus kasutaja sisestatud sõna ei pruugi täpselt vastata andmebaasis sisestatud sõnadele. Lahendasime probleemi kahel viisil: 1) realiseerisime kiire ligikaudse regulaaravaldiste sobitamise programmi (Kristo Tammeoja) ning 2) realiseerisime üldistatud teisenduskauguse arvutamise algoritmi ja veebiliidesed (Reina Käärik, Siim Orasmaa). Üldistatud teisenduskaugus lubab sõnastada täheühendite muutumistele „hinna“: näiteks sh->š, zh->ž, hv->ff, w->v, v->w, mis võimaldab leida sõnu nagu dušš või Dušanbe, vaffa jt. Või vana kirjapildi sõnu nagu walletellenut, wallitzetuttelle, jne. Eksperimenteerisime mitme viisiga kuidas automaatselt tuletada vastavaid teisendusi ja anda nendele sobivaid kaale. Avalikud prototüübid kus saab katsetada vastavat otsimist, asuvad praegu:

Kirjapildi ja häälduse põhjal eesti keeles inglise tekstist päringute tegemiseks: (n. bjuutiful -> beautiful [ˈbju:təf(ə)l, -if-])
https://biit-dev.cs.ut.ee/~orasmaa/ing_ligikaudne/
Venekeelsete sõnade otsimine ladina tähtede abil (n. utsitelnitsa -> учительница).
https://biit-dev.cs.ut.ee/~orasmaa/gen_ed_test/