

EESTI KEELE KEELETEHNOLOOGILINE TUGI (2006-2010)

Riiklik programm

SISUKORD

1. SISSEJUHATUS.....	2
1.1. Eesti keeletehnoloogia infrastruktuuri ajakohastamine.....	5
1.2. Programm ja selle täitmisega seotud riskid.....	6
1.3. Programmi haldamine ja juhtimine.....	7
2. KEELETEHNOLOOGIA VALDKONNAD JA NENDE SEIS EESTIS.....	9
3. PROGRAMMI ALAEESMÄRGID JA OODATAVAD TULEMUSED.....	12
3.1. Keeletarkvara.....	12
3.2. Keeletehnoloogilised ressursid.....	15
3.2.1. Kirjaliku keele korpused.....	15
3.2.2. Suulise keele korpused.....	17
3.2.3. Elektroonilised sõnastikud ja andmebaasid.....	18
3.2.4. Formaalsed keelekirjeldused.....	19
4. PROGRAMMI FINANTSEERIMINE.....	20
LISA 1: Põhjamaade keeletehnoloogia doktorikooli ja Tartu Ülikooli vaheline koostööleping.....	21

1. SISSEJUHATUS

Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006–2010)" (edaspidi RP EKKTT) peaesmärgiks on eesti keele keeletehnoloogilise toe arendamine tasemele, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises keskkonnas. Keeletehnoloogia on infotehnoloogiat ja keeleteadust ühendav interdistsiplinaarne valdkond, mis tegeleb vahendite (meetodite, algoritmide, programmide) väljatöötamisega nii kirjutatud kui ka suulise keele arvutitöötluks. Kuna eesti keele kõnelejate arv on liiga väike selleks, et keeletehnoloogiat arendavatel firmadel tekiks huvi eesti keelele sobivate vahendite väljatöötamise vastu, on keeletehnoloogia arendamiseks paratamatult vaja riigi toetust.

Keeletehnoloogia kuulub Eesti teadus- ja arendustegevuse strateegia "Teadmispõhine Eesti" võtmevaldkonna "Kasutajasõbralikud infotehnoloogiad ja infoühiskonna areng" alla. Keeletehnoloogia on ka Vabariigi Valitsuse heakskiidu leidnud "Eesti keele arendamise strateegia (2004–2010)" võtmevaldkond. Keelestrateegia koostamise käigus töötati läbi ka keeletehnoloogia valdkond ning määrati kindlaks keeletehnoloogia strateegilised ülesanded ja eesmärk, mis on ka siinse programmi peaesmärk. RP EKKTT ülesehitus vastab "Eesti keele arendamise strateegias (2004–2010)" määratud jaotusele:

- keeletehnoloogia valdkonnad;
- keeletarkvara;
- keeletehnoloogilised ressursid.

RP EKKTT alaesmärgid ja oodatavad tulemused on loetletud keeletehnoloogia valdkondade, keeletarkvara ja keeletehnoloogiliste ressursside juures.

RP EKKTT on välja kasvanud riiklikust programmist "Eesti keel ja rahvuslik mälu (2004–2008)", mis sisaldab keeletehnoloogia moodulit koos alajaotustega "Keeleressursid (tekstikorpused, leksikaalsed ja grammatilised ressursid)" ning "Tarkvara rakendusala". RP EKKTT käivitumisel lõpetatakse keeletehnoloogia projektide rahastamine riikliku programmi "Eesti keel ja rahvuslik mälu (2004–2008)" raames. RP 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 tekib väga väärtuslik intellektuaalne omand. Loomise hetkest alates on see intellektuaalne omand avalik omand.

Kuigi käesolev programm rahastab teadus- ja arendustööd prototüüpide loomiseni, on programmi eesmärgiks, et tulemused jõuaksid e-ühiskonnas kasutusse. E-riik ja e-ühiskond on programmi tulemi kasutajad kõige laiemas mõttes. Programmi tulemused on mõeldud kasutamiseks e-valitsemises, e-hariduses, e-tervishoius, e-kaubanduses, e-panganduses, e-teeninduses, e-kirjastamises, e-(mobiil)sides, e-suhtluses, e-infopäringutes jne. Sellise e-ühiskonna loomiseks on vaja abivahendeid tekstitöötlusel ning kõnetehnoloogia viimist uuele tasemele. Seal, kus praegu on vaja ID kaarti, saab tulevikus isikut tuvastada hääle abil. Ei tohi unustada, et ka kodutehnikaga hakatakse lähitulevikus suhtlema loomulikus keeles. Käesolevat programmi on vaja ka selleks, et e-kodutehnikale saaks käske anda eesti keeles.

Keeletehnoloogia aktuaalseimad rakendused on (seni puuduvad) masintõlkeprogrammid automaatseks tõlkimiseks eesti keelest teistesse keeltesse ja teistest keeltest (eeskätt inglise keelest) eesti keelde, sh piiritlemata valdkonnale orienteeritud internetitõlkide analoogid ja piiritletud valdkonnale orienteeritud praktilised masintõlkeprogrammid, samuti inimtõlkijatele mõeldud abivahendid (sh nn tõlgi mälu). Tõlkeprogrammid on vajalikud ka mitmekeelses infootsingus, tõlkimaks eesti keeles sõnastatud päringuid võõrkeeltesse ja otsingutulemusi tagasi eesti keelde. Infootsingu efektiivsuse tõstmise üheks vahendiks on automaatne sisukokkuvõtja, mis teeks pikast dokumendist nt mobiiltelefoni ekraanile mahtuva lühikokkuvõtte, parandades sellega kasutajate juurdepääsu infole ja säästes aega. Seni eesti keele jaoks olemas olev õigekirjakorrektor opereerib vaid üksiksõnavormidega, seevastu kavandatav automaatne grammatikakorrekter suudaks tuvastada ja parandada kogu lauset (sh ühildumisvigu) ning leiaks lisaks kirjutaja abivahendina rakendust ka nt keeleõppes.

Automaatse kõnetuvastuse aktuaalseim realisatsioon on automaatne diktofon, mille sisendiks on eestikeelne kõne ja väljundiks ortograafiline tekst. Nii kõnetuvastus kui ka kõnesüntees leiaksid rakendust mitmesugustes telefoniteenuste süsteemides, mille analooge on maailmas (teiste keelte jaoks) juba väga palju, nt telefonikõne suunamine inimese nime järgi, liiklusinfo, sõiduplaanide info, reisi planeerimine, interaktiivsed abilauad jms, samuti erivajadustega kasutajate abistamisel.

Keeletehnoloogiline tugi on vajalik infoühiskonna arendamiseks. See annab oma panuse meid ümbritseva tehisintellekti (*Ambient Intelligence*) e-keskkonna loomisesse aastaks 2020. Tänu keeletehnoloogilisele toele toimib see keskkond eestikeelsena. Arvuti ja väliskeskkonna vahetu interaktsioon ja ühtesulamine toimub eesti keele ja masintõlke

vahendusel. Meie visioon eeldab, et mitmesuunaline masintõlge toimib nii suuliste kui ka kirjalike tekstide puhul. Tänu kõneleja automaattuvastusele saab kõiki e-toiminguid teha suulist kõnet kasutades. Tõlkeprogrammid võimaldavad meil kasutada elektroonilisi ressursse paljudes keeltes.

Paljud maailma riigid on mõistnud, et oma keelelise ja kultuurilise identiteedi säilitamiseks infoühiskonnas on vaja välja töötada rahvuskeelte keeletehnoloogiline tugi ning seda arendada. Keeletehnoloogia on üks Euroopa Liidu prioriteete, mida toetatakse mitme programmi kaudu. Eesti jaoks on samasugune riiklik programm eriti vajalik nüüd, mil kuulumine Euroopa Liitu nõuab meilt nüüdisaegsete keeletehnoloogiliste vahendite olemasolu samal tasemel teiste riikidega. Keeletehnoloogia arendamine on viimastel aastatel olnud eriti intensiivne Põhjamaades, kus on käivitatud kõiki Põhjamaid ühendav ja NorFA rahastatav "Nordic Language Technology Research Program" ning selle allprojektina loodud "Nordic Network of Documentation Centres for Language Technology", mis koondab ja levitab infot keeletehnoloogia-alaste tööde ja tulemuste kohta. Selle võrgustikuga on Eesti keeletehnoloogidel kujunenud head sidemed ja lähitulevikus on ette näha sellega liitumine. 28. oktoobril 2004 sõlmiti koostööleping (lisa 1) Tartu Ülikooli ja eelnimetatud programmi raames toimiva keeletehnoloogiaalast doktoriõpet korraldava "Nordic Graduate School of Language Technology" vahel.

Eestis muutus keele- ja kõnetehnoloogia areng intensiivseks 1990. aastate keskpaigast alates, mil käivitus Euroopa Liidu COPERNICUS-programm, samuti on Eesti osalenud nii 4. kui ka 5. raamprogrammi keeletehnoloogia projektides. Vajalike spetsialistide ettevalmistuse tagab Tartu Ülikooli arvutilingvistika ja keeletehnoloogia alane kraadiõpe ning Tartu Ülikooli poolt koos Eesti Keele Instituudi ja Tallinna Tehnikaülikooli Küberneetika Instituudiga riikliku arengukava meetme 1.1 raames käivitav doktorikool "Keeleteadus ja -tehnoloogia". Spetsialistide ettevalmistusele annab uue mõõtme liitumine Põhjamaade eespool nimetatud keeletehnoloogia doktorikoolide võrgustikuga.

Keeletehnoloogiline tugi koosneb keeleressurssidest ja keeletarkvarast ning viimase rakendustest. Keeleressursid on elektroonilised andmekogud, mida kasutatakse ka keeletarkvara väljatöötamiseks. Keeletarkvara koondab endas meetodeid, algoritme ja programme keelematerjali töötlemiseks ja rakendussüsteemide väljatöötamiseks. Keeletehnoloogia ülesandeks on luua ressursid ja vahendid, mis tagavad eesti keelele võrdsed võimalused infotehnoloogilises keskkonnas suhtlemiseks.

Keeleressursid luuakse nii, et oleks tagatud nende omavaheline semantiline koostoime ning koostoime kõigi teiste infosüsteemidega. Suutlikkus teiste süsteemidega regulaarselt andmeid vahetada on muutumas süsteemide üheks peamiseks oskuseks/komponendiks üldse, ning semantiline koosvõime on selle oskuse põhituum. Semantiline koostoime ei ole absoluutne, sest ei ole võimalik kasutada ainult ühte andmebaaside struktuuri ega universaalset keelt (nt XML). Seepärast pööratakse suurt tähelepanu kasutajaliideste loomisele, mis lubab välja arendada semantiliselt koostoimivad võrgustikud ning luuakse vastavad standardid.

PR EKKTT prioriteetideks kirjaliku keele töötamise alal on masintõlge ning süntaksi- ja semantikapõhine, eelkõige mitmekeelne infootsing. Need eeldavad eesti keelele kohandatud keeletöötamise tarkvara, nagu morfoloogiline, süntaktiline ja semantiline analüüs ja süntees. Viimaste eelduseks on keeleressursside olemasolu, mis on vajalikud programmide väljatöötamiseks ja treenimiseks.

Kuna inimene-masin-inimene suhtluses muutub üha olulisemaks suulise kõne roll (klaviatuurita arvutid ja kodumasinad, mida saab juhtida suulises keeles), siis käib selles valdkonnas kogu maailmas intensiivne uurimis- ja arendustöö. Et eesti keelel oleks keeletehnoloogiline tugi suulise keele valdkonnas, on riikliku programmi põhisuundadeks kõnesüntees, kõnetuvastus ja inimene-masin dialoogsüsteemid.

1.1. EESTI KEELETEHNOLOOGIA INFRASTRUKTUURI AJAKOHASTAMINE

Programmi edukaks täitmiseks on vaja ajakohastada Eesti keeletehnoloogia infrastruktuur (vt ka 3. ptk, tabel 2). Selleks on vaja pöörata tähelepanu:

- riistvarale ja
- tarkvarale.

RP EKKTT täitjate riist- ja tarkvara tuleb ajakohastada ning ühtlustada. Samuti on vaja välja arendada ühilduv infotehnoloogiline platvorm, mis võimaldab erinevate uurimiskeskuste töö ühtlustamist ja vastastikust info ning ressursi jagamist.

Eesti keeletehnoloogia infrastruktuuri ajakohastamise rahastamiseks planeeritud vahendid on alates 2007. a kavas taotleda tõukefondidest (vt 3. ptk, tabel 1).

Riistvara puhul on esmatähtis spetsialiseeritud tööjaamade (*work station*) hankimine ning mahukaks andmetöötamiseks vajaliku platvormi väljaarendamine. Programmi täitjatel peab olema kasutusel ühilduv tarkvarakeskkond.

Kõnetehnoloogia arendamiseks on vaja hankida kõnetuvastussüsteemide arenduskeskkond ja inimene-masin dialoogi arenduskeskkond, samuti multimeediakeskkond. Masintõlke ja infootsisüsteemide arendamiseks on vaja muretseda vastav platvorm. Andmekogude, sh sõnaraamatute haldamiseks on vajalik nt Oracle andmebaas; nende loomiseks Interneti-keskkonnas aga ka mitmete igapäevaseimate programmide litsentsid.

1.2. PROGRAMM JA SELLE TÄITMISEGA SEOTUD RISKID

Programmile on suureks riskiks see, et rahvusvahelised suurkontsernid võivad mõned eesti keelele suunatud lahendused ise ära teha, need patenteerida ja muuta need kättesaadavaks vähestele jõukamatele elanikkonnakihtidele.

Riskiks on ka võimalikud keelehoiakute muutused ning keeleseaduste muudatused, millega kaasneks laialdane ingliskeelse tarkvara kasutamine ning muutuks tarbetuks eestikeelse infotehnoloogilise keskkonna loomine. Selle riskiga on seotud ingliskeelse tarkvara ulatuslik kommertsiaalne pealetung. Seda riski aitab maandada tarkvara eestikeelsete versioonide loomine.

Programmi täitmise riskiks tuleb pidada ka seda, et keeletehnoloogiat hakatakse fetiseerima ning programmilt oodatakse rohkem kui see pakkuda suudab, jättes unarusse igapäevase keelekorralduse ja keelehoolded.

RP EKKTT edukas täitmine sõltub programmi optimaalsest rahastamisest. Programmi alarahastamise korral tuleb teha valik ning jätta osa ülesandeid täitmata. Kuna keeletehnoloogilised lahendused eeldavad järjestikuste ülesannete täitmist, siis võib alarahastamine tähendada seda, et luuakse küll keeletehnoloogilised ressursid, aga keeletarkvara ja keeletehnoloogiliste lahenduste prototüübid jäävad loomata. Samal ajal ei taga programmi optimaalsest suurem rahastamine selle kiiremat täitmist, sest keeletehnoloogia vallas töötavate inimeste arv on Eestis piiratud. Oluliseks riskifaktoriks programmi täitmisele on vajalike, sh doktorikraadiga spetsialistide koolitusprogrammi mittetäielik rakendumine ülikoolides. Kui erasektoris tõusevad palgad Euroopa Liidu tasemele ja programm on alarahastatud, siis lahkuvad inimesed akadeemilisest sfäärist erasektoris ja programmi eesmärgid jäävad saavutamata. Programmi rahastamine tuleb viia vastavusse erasektori palkade üldise tõusuga.

Programmi eesmärged ei ole võimalik saavutada ilma Eesti keeletehnoloogia infrastruktuuri ajakohastamiseta, st ilma RP EKKTT täitjate riist- ja tarkvara tänapäevastamise ning ühtlustamiseta ja ilma ühtse infotehnoloogilise platvormita. Riskiks tuleb pidada ka programmi

täitjate võimalikku omavahelist konkurentsi ja koostöö takerdumist programmi teiste täitjate ning erasektoriga. Riski maandamiseks on vaja laialdast koostööd nii kodu- kui ka välismaiste partneritega.

1.3. PROGRAMMI JUHTIMINE JA HALDAMINE

Programmi sisuliseks juhtimiseks moodustatakse programmi juhtkomitee. Programmi juhtkomitee moodustatakse keeletehnoloogia ekspertidest, Haridus- ja Teadusministeeriumi ning teadusüldsuse esindajatest.

Juhtkomitee ülesandeks on analüüsida keeletehnoloogia arengut Eestis ning kogu maailmas. Juhtkomitee jaotab käesoleva programmi alaosadele "Keeletarkvara" ja "Keeletehnoloogilised ressursid" eraldatud vahendid konkursi korras. Keeletehnoloogia infrastruktuuri kaasajastamiseks vajalikud tööd tellitakse juhtkomitee ettepanekul vastavalt kehtivatele õigusaktidele. Juhtkomitee töötab välja programmi taotlusvormid ning vastutab programmi eesmärkide täitmise eest ja jälgib, et programmi vahendeid kasutataks sihipäraselt. Selleks kasutab juhtkomitee laiapõhjaliselt ekspertide abi.

Programmi haldamiseks leiab juhtkomitee programmi kureeriva asutuse ning konkursi korras programmi koordinaatori ning esitab need heakskiitmiseks Haridus- ja Teadusministeeriumile. Ministeerium sõlmib programmi haldamiseks lepingu programmi kureeriva asutusega, kes omakorda sõlmib töölepingu programmi koordinaatoriga. Programmi koordinaatori ja programmi haldamise kulud kaetakse programmi vahenditest.

Programmi koordinaator võtab osa juhtkomitee koosolekutest, kuid ei oma hääleõigust. Programmi koordinaator vastutab programmi konkursi korraldamise, vahendite sihipärase kasutamise ja aruandluse läbiviimise eest. Programmi koordinaator valmistab ette lepingud programmi täitjatega. Programmi koordinaator valmistab ette Eesti keeletehnoloogia infrastruktuuri ajakohastamiseks eraldatud vahendite jaotuskava ning programmi juhtkomitee esitab selle kinnitamiseks haridus- ja teadusministrile. Programmi koordinaator valmistab ette keeletehnoloogia infrastruktuuri kaasajastamise taotlused, korraldab programmi avalikkusele tutvustamise eesmärgil info jagamist ning täidab muid ülesandeid, mis on vajalikud programmi eesmärkide saavutamiseks. Programmi koordinaatoril on õigus teha juhtkomiteele ettepanekuid abitööjõu palkamiseks eelarve piires. Koordinaator esitab programmi täitmise sisulised aruanded kinnitamiseks juhtkomiteele ning rahalised aruanded Haridus- ja Teadusministeeriumile.

Programmi juhtkomitee otsustab programmi käivitamisel, milliste riikliku programmi "Eesti keel ja rahvuslik mälu (2004-2008)" keeletehnoloogia projektide rahastamist jätkatakse ja millised projektid lõpetatakse.

2. KEELETEHNOLOOGIA VALDKONNAD JA NENDE SEIS EESTIS

Selles jaotises anname ülevaate eesti keele keeletehnoloogilise toe loomiseks relevantsetest valdkondadest ja nende hetkeseisust. Põhjalikum käsitus keeletehnoloogia seisundist leidub "Eesti keele arendamise strateegia (2004–2010)" ametliku lisana koostatud eesti keele seisundi uuringus "Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara" (koostanud Kadri Muischnek, Heili Orav, Heiki-Jaan Kaalep ja Haldur Õim, toimetanud Urve Talvik; Tallinn: Eesti Keele Sihtasutus, 2003).

Lõppkasutajale orienteeritud keeletehnoloogia valdkonnad

2.1. Kõnesüntees on ortograafilise teksti teisendamine tehiskõneks. Kõnesüntesaatori sisendiks võib olla mis tahes eestikeelne tekst ja väljundiks samasisuline tehiskõne. Kõnesüntesaatorit kasutatakse inimene-masin dialoogsüsteemides, puuetega inimestele mõeldud abivahendites jm. Kasutaja seisukohalt on oluline väljundkõne kvaliteet, st arusaadavus ja loomulikkus. Aastatel 1997–2002 loodud kõnesüntesaatori prototüübi laialdast rakendamist piirab väljundkõne monotoonsus ja kõne halb sidusus, samuti puuduvad liidesed kõnesünteesi integreerimiseks erinevatesse arvutikeskkondadesse.

2.2. Kõnetuvastus on inimekõne teisendamine kõne sisule vastavaks tekstiks. Kõnetuvastussüsteeme kasutatakse inimene-masin dialoogsüsteemides, teksti sisestamiseks arvutisse, seadmete ja süsteemide hääljuhtimiseks jm. Eestikeelse kõnetuvastuse aktuaalseim realisatsioon on nn automaatne diktofon, mille sisendiks on sidus eestikeelne kõne ja väljundiks ortograafiline tekst. Kõnetuvastuse uuringuteks on kohandatud mitmed keelest sõltumatud tuvastusmootorid, samuti on loodud ja testitud eesti keele eripära arvestavaid statistilisi kõnemudeleid. Sellegipoolest on uurimis- ja arendustöö eestikeelse kõnetuvastuse väljatöötamiseks alles algusjärgus.

2.3. Inimene-masin dialoogsüsteemide arengus on märgatav jäikade, rangete reeglite ja küsimustike alusel toimivate süsteemide asendamine paindlike ja kasutajasõbralike süsteemidega. Infotelefoni dialoogiaktide tuvastaja peaks olema võimeline aru saama kindlat ainevalda puudutavatest erinevas vormis esitatud küsimustest ja oskama vajaduse korral

esitada täpsustavaid küsimusi, et selgitada kasutaja vajadusi. Maailmas on olemas telefoniteenuste süsteeme, mis lõimivad paljusid keele automaattöötuse mooduleid (sh kõnetuvastuse ja -sünteesi) ning annavad kasutajale nt liiklusinfot, ostuabi jne, ent eesti keele jaoks sellised süsteemid seni puuduvad.

2.4. Grammatikakorrektor osutab grammatiliselt vigastele eesti keele lausetele ja pakub välja õiged. Eesti keele jaoks on olemas speller ehk õigekirjakorrektor, kuid see opereerib vaid üksiksõnavormidega. Eesti keele jaoks pole veel olemas automaatset grammatikakorrektorit, mis võimaldaks tekstist avastada grammatikavigu, nagu näiteks ühildumisvead, komavead, suure ja väikese algustähe vead, kokku-lahkukirjutamise vead, öeldise puudumine lauses jne.

2.5. Infootsingute (Information retrieval) lahendused on üha kasvavate, valdavalt tekstipõhiste teoste, sõnastike, korpuste, Interneti keskkondade ja teiste andmebaaside efektiivse kasutamise oluline eeldus, et võimaldada kiiresti ja täpselt leida ning kokkuvõtvalt ja arusaadavalt esitada olulist informatsiooni erinevates tehnoloogilistes keskkondades (Internet, CD, mobiil, jne). Eesti keele omapära eeldab morfoloogilise analüüsi, sünteesi ja ühestamise moodulite kasutamist, võimalusel ka tesauruste ja teiste sõnastike integreerimist. Välja tuleb arendada modulaarsed lahendused ning tarkvara, mis oleks ühelt poolt võimalikult intelligentne ning teiselt poolt piisavalt efektiivne, et seda saaks skaleerida väga suurte andmebaaside kasutamiseks. Lõppkasutajale suunatud lahendustes on lisaks oluline kasutajaliideste mugavus ning kasutajate erinevate soovide arvestamine.

2.6. Masintõlge on tekstide tõlkimine ühest loomulikust keelest teise. Mitmekeelse infootsingu puhul on vaja päring tõlkida muukeelseks ja seejärel vastused tagasi päringukeelseks. Kuigi täielik masintõlge on keeruline ülesanne, on ka osatõlgetest sageli palju kasu. Saadud tulemusi saab infootsingu kõrval rakendada nt tõlkijatele mõeldud abivahendites jm. Esimene samm on eesti-inglise sõnastiku kasutamine päringuväljendite tõlkimiseks inglise keelde, järgmise sammuna tuleb ingliskeelsed tulemused tõlkida eesti keelde.

2.7. Leksikograafi töökeskkond on keeletehnoloogiliste vahendite süsteem, mis lõimib keeleressursid (elektroonilised sõnastikud, andmebaasid, tekstikorpused) ja tarkvara ning võimaldab automatiseerida sõnastike koostamist. Süsteem on vajalik nii traditsiooniliste sõnaraamatute kui ka keeletehnoloogia teistes moodulites kasutatavate leksikaalsete ressursside loomiseks. Seni on igaks üksikjuhuks koostatud ühekordsed vahendid, mis ei haaku omavahel ega võimalda taaskasutada olemasolevaid ressursse. Töövahendid on vaja luua kahes versioonis: (a) laiatarbesüsteem, mida saavad veebi vahendusel kasutada kõik soovijad, ning (b) professionaalne süsteem, mida kasutavad leksikograafiaspetsialistid. Lisaks

on tarvis liidest, mis teisendab traditsioonilise sõnastikuinfo keeletehnoloogia jaoks sobivasse vormingusse.

Keeletehnoloogia valdkonnad, mis on vajalikud lõppkasutajale suunatud rakenduste loomiseks

Järgnevalt loetletud analüüsi- ja sünteesiprogramme kasutatakse kõikides eespool loetletud rakendustes, kuid keeletehnoloogia praeguse arengutaseme juures moodustavad nad ühtaegu omaette uurimisvaldkonnad.

2.8. Morfoloogiline analüüs ja süntees. Morfoloogiline analüsaator on programm, mille sisendiks on sõna muutevorm ning väljundiks sama sõna algvorm ja grammatilised tunnused. Kuna eestikeelses tekstis saab poolt sõnavormidest analüüsida mitmel moel, tuleb õige variandi valimiseks (morfoloogiliseks ühestamiseks) arvestada ka konteksti. Morfoloogilise süntesaatori sisendiks on sõna algvorm ja grammatilised tunnused ning väljundiks sõna muutevorm. Eesti keele jaoks on olemas mitu morfoloogilist analüsaatorit, ühestajat ja süntesaatorit, kuid neid on vaja kohandada konkreetsete rakenduste jaoks.

2.9. Sõnamoodustuslik analüüs ja süntees. Sõnamoodustuse tarkvara on vajalik morfoloogilise analüsaatori ja süntesaatori töö kvaliteedi parandamiseks. Sõnamoodustusega seotud tarkvarasse kuuluvad (a) sõnamoodustuse programmid, mis võimaldavad analüüsida ja sünteesida eesti keele tuletisi ja liitsõnu, ning (b) sõnamoodustuse analüüsi programmidega ühilduv morfoloogiline ühestaja. Eesti keele jaoks puudub seni spetsiaalne, süntaktilisi ja semantilisi kitsendusi arvestav sõnamoodustuse tarkvara.

2.10. Süntaktiline analüüs ja süntees. Süntaktiline analüsaator on programm, mis määrab tema sisendile saabuva morfoloogiliselt analüüsitud ja ühestatud lause grammatilise ehituse. Eesti keele jaoks on olemas süntaktilise analüsaatori esmane versioon, kuid rakenduste jaoks on vaja seda tunduvalt täiustada. Süntaktiline süntesaator on programm, mis saab sisendiks lause grammatilise struktuuri koos sõnade algvormidega ning varustab iga sõna grammatiliste tunnustega, mida seejärel saab kasutada morfoloogiline süntesaator konkreetsete sõnavormide moodustamiseks. Tööd eesti keele süntaktilise süntesaatoriga on alles algusjärgus.

2.11. Semantiline analüüs ja süntees. Semantiline analüsaator on programm, mis leiab sisendile saabunud morfoloogiliselt ja süntaktiliselt analüüsitud lause tähenduse. Allülesanneteks on sõnatähenduste ühestamine, nendevaheliste semantiliste seoste määramine, sõnade asendamine nende semantilise esitusega ning semantiliste järelduste tuletamine. Semantiline süntesaator on programm, mis teisendab muus vormis (ka pildina) esitatud info keele semantilise esituse kujule. Eesti keele jaoks on loodud sõnatähenduste automaatse ühestaja testversioon, kuid see vajab täiendamist. Tööd semantilise analüsaatori loomisega on alles algusjärgus, semantilise sünteesiga pole eesti keele osas veel tegeldud.

2.12. Kõneaktide tuvastaja on programm, mis määrab ära, millist kõneakti väljendab sisendiks olev semantiliselt analüüsitud lause. Kõneakte (nt küsimused, vastused, käsud jne) kasutatakse inimestevahelises ning inimese ja arvuti vahelises suhtluses. Kõneaktide tuvastaja on vajalik eelkõige inimene-arvuti dialoogsüsteemide loomises. Eesti keele jaoks selline programm seni puudub.

3. PROGRAMMI ALAEESMÄRGID JA OODATAVAD TULEMUSED

3.1. keeletarkvara

Keeletarkvara all mõistetakse siin meetodeid, algoritme ja arvutiprogramme keelematerjali töötlemiseks, mis siinse programmi raames on kavandatud eesti keele keeletehnoloogilise toe loomiseks ja vajalike rakendussüsteemide väljatöötamiseks. Selleks kohandatakse eesti keelele mitmeid keelest sõltumatuid keeletötlusmeetodeid ja -algoritme ning luuakse uusi ainult eesti keele spetsiifikast lähtuvaid lahendusi.

Lõppkasutajale orienteeritud alaeesmärgid ja oodatavad tulemused

3.1.1. Kõnesüntees. Eesmärgiks on kõrgekvaliteedilise eestikeelse tekst-kõne sünteesi prototüübi loomine, mida saab koos kõnetuvastusega rakendada inimene-masin-inimene suhtluses. Sünteeskõne kvaliteedi tagab nn prosodia generaator, mis juhib kõneüksuste

kestuste ja häälekõrguse muutumist ajas. Väljundkõne ajaline struktuur on vaja modelleerida erinevatele kõnestiilidele ja tekstitüüpidele. Kõnesüntesaatori mitmekülgse rakendamise huvides on vaja luua meeshääle sünteesi kõrvale ka naishääle süntees.

Oodatavad tulemused: täiustatud kõneprosoodia mudelid, mees- ja naishäälega eestikeelse kõnesünteesi loomine.

3.1.2. Kõnetuvastus. Eesmärgiks on kõnetuvastuse arendamiseks vajaliku töökeskkonna ja eesti keele spetsiifiliste kõnetuvastusmoodulite väljatöötamine. Uurimis- ja arendustöö on suunatud eesti keelele sobivate akustiliste mudelite loomisele, erinevate keelemudelite väljatöötamisele ning signaalitöötlusmeetodite arendamisele. Kõnelejast sõltumatu piiratud sõnastikuga (kuni 10 000 sõna) kõnetuvastuse prototüübi väljatöötamise kõrval on tarvis arendada meetodeid konkreetsele kõneleajale kohandatava suure sõnastikuga (üle 60 000 sõna) kõnetuvastuse prototüübi loomiseks.

Oodatavad tulemused: piiratud sõnastikuga (kuni 10 000 sõna) kõnetuvastus, mis võimaldab automaatselt moodustada suulisest kõnest kirjalikku teksti, meetodid ja programmid tuvastussüsteemide treenimiseks, suure sõnastikuga kõnetuvastussüsteemi erinevad moodulid.

3.1.3. Inimene-masin dialoogsüsteemid. Eesmärgiks on inimene-masin dialoogsüsteemide modelleerimiseks ja arendamiseks ning dialoogimudelite testimiseks vajaliku tehnoloogilise keskkonna väljatöötamine ning erinevatele ainevaldkondadele kohandatava dialoogsüsteemi prototüübi loomine.

Oodatavad tulemused: dialoogsüsteemide arenduskeskkond, ainevaldkonnale kohandatavad dialoogsüsteemid.

3.1.4. Grammatikakorrektor. Eesmärgiks on luua programm, mis tuvastab eestikeelsetes lihtlausetes grammatilised vead (ühildumisvead, komavead, öeldise puudumine jm).

Oodatavad tulemused: morfoloogilise analüsaatori ja ühestajaga ühilduv grammatikakorrektor, mida saab kasutada samuti nagu õigekirjakontrolli.

3.1.5. Infootsingute lahendused. Eesmärgiks on arendada välja eestikeelsete tekstiandmebaaside jaoks mugandatud päringusüsteemid, mis kasutavad morfoloogilise analüüsi, sünteesi ja ühestamise mooduleid, sünonüümisõnastikku ja tesauruseid ning teisi päringu täpsust ja efektiivsust tõstvaid lahendusi. Lõppkasutajale suunatud lahendustes on lisaks oluline kasutajaliideste mugavus ning kasutajate erinevate soovide arvestamine.

Oodatavad tulemused: eesti keelele kohandatud infootsingu tarkvaralahendused ning nende praktiline juurutamine lõppkasutajale suunatud infokeskkondadesse.

3.1.6. Masintõlge. Eesmärgiks on käivitada inglise-eesti ja eesti-inglise masintõlke alased teadus- ja arendustööd, lähtudes paralleelkorpustest.

Oodatavad tulemused: paralleelkorpusega kaetud kitsa ainevaldkonna inglise-eesti ja eesti-inglise tõlkeabi süsteemi (masintõlke) prototüüp.

3.1.7. Leksikograafi töökeskkond. Eesmärgiks on luua interaktiivne töökeskkond, mille vahendid võimaldavad kasutada olemasolevaid keeleressursse (elektroonilisi sõnastikke, andmebaase, tekstikorpusi) ja olemasolevat keeletarkvara (morfoloogia, sõnamoodustuse, süntaksi, semantika programme) ning luua uusi elektroonilisi sõnastikke mitmeks eri otstarbeks. Samuti luuakse eesti keele leksikaalgrammatiline andmebaas, mis ühendab erinevate andmebaaside info.

Oodatavad tulemused: töötlusliides traditsioonilise sõnastikuinfo teisendamiseks keeletehnoloogia jaoks sobivasse vormingusse, leksikograafi töökeskkonna prototüüp. Sõnaraamatute koostamise programmid – Eesti–X-keele sõnaraamatu põhjad erinevas mahus (20 000, 40 000 ja 90 000 märksõna).

Lõppkasutajale orienteeritud tulemuste saavutamiseks vajalikud alaeesmärgid ja oodatavad tulemused

3.1.8. Morfoloogiline analüüs ja süntees. Eesmärgiks on olemasoleva morfoloogiatarkvara edasiarendamine kahes suunas: (a) morfoloogia funktsioonide täiendamine: lisaks senisele sõna muutevormide töötlusele ka sõnade (tuletiste ja liitsõnade) moodustuslik e morfoloogiline analüüs ja süntees; (b) allkeelte valiku laiendamine: lisaks senisele kirjalikule tänapäeva kirjakeelele ka suuline keel ja muud allkeeled (nt dialoog, teaduskeel, Internetikeel jne).

Oodatavad tulemused: (a) tuletiste ning liitsõnade analüüsi ja sünteesi programmide prototüübid; (b) eri allkeelte orienteeritud morfoloogiatarkvara.

3.1.9. Süntaktiline analüüs ja süntees. Eesmärgiks on olemasoleva pindsüntaktilise analüsaatori edasiarendamine rakendusi võimaldava tasemeni, samuti formaalsete

keelekirjelduste loomine eesti keele süvasüntaktilise analüüsi ja süntaktilise sünteesi programmide väljatöötamiseks.

Oodatavad tulemused: süntaksipõhiste refereerimis- ja sisukokkuvõtteprogrammide prototüübid, grammatikakorrektori prototüüp, süntaksi- ja semantikapõhise infootsüsteemi prototüüp, interaktiivsed eesti keele süntaksi õpiprogrammid.

3.1.10. Semantiline analüüs ja süntees. Eesmärk on valida sõnade ja lausete semantilise esituse vorm eesti keele lihtlausete jaoks; valida ja töötada läbi valdkonnad, mille alaseid tekste hakatakse töötlemas; koostada semantiline leksikon nende valdkondade sõnadest; töötada välja semantiliste järeldusreeglite vorm ja konkreetset järeldusreeglid valitud valdkondade jaoks.

Oodatavad tulemused: väljatöötatud semantilise esituse vorm, semantiliste seoste loetelu, semantiline leksikon ja semantilised järeldusreeglid ning eesti keele lihtlause semantilise analüüsi programmi prototüüp.

3.1.11. Foneetiline analüüs ja süntees. Eesmärgiks on eestikeelse kõne modelleerimine, akustiliste ja tajumudelite väljatöötamine.

Oodatavad tulemused: kõnesünteesi ja -tuvastuse eesti keele spetsiifiline tarkvara.

3.1.12. Kõneaktide tuvastaja. Eesmärgiks on välja töötada eestikeelsetes infodialoogides kasutatavate kõneaktide tüpologia ning meetodid kõneaktide automaatseks määramiseks (tehisearvutõrgud või otsustuspuud, mis kasutavad morfoloogilisi, süntaktilisi ja semantilisi tunnuseid).

Oodatavad tulemused: kõneaktide tüpologia ning kõneakte tuvastava programmi prototüüp.

3.2. Keeletehnoloogilised ressursid

Keeleressursside all mõistame siin elektroonilisi andmekoguseid, mida kasutatakse keeletarkvara väljatöötamiseks: korpused (kõnesignaali ja tekstide kogumid), elektroonilised sõnastikud ja andmebaasid, formaalsed keelekirjeldused (grammatikad). Keeleressursid on keelespetsiifilised ning neid ei saa teistest keeltest üle võtta.

3.2.1. KIRJALIKU KEELE KORPUSED

3.2.1.1. Eesti kirjakeele koondkorpus on formaalsete keelekirjelduste alusmaterjal ning elektrooniliste sõnastike ja andmebaaside koostamise abivahend. Koondkorpust eeldavad kõik

kirjaliku keele töötlemisega seotud tööd, lisaks on suurel keelekorpusel oluline roll keeleteaduslikus uurimistöös.

Oodatavad tulemused: (1) eesti keele koondkorpuse arendamine 200 miljoni sõnani, ajakirjandustekstide kõrval tuleb tähelepanu pöörata ka kirjaliku keele teistele allkeeltele (ilukirjandus ja teadustekstid) ning nn uut tüüpi kirjalikele tekstidele (jututubade, uudisgruppide jms keel); (2) koondkorpuse automaatne morfoloogiline märgendamine, mis on vajalik seetõttu, et ainult morfoloogiliselt märgendatud korpusele on võimalik esitada päringuid sõna algvormi järgi.

3.2.1.2. Mitmekeelseid paralleelkorpusi on vaja eelkõige tõlkija abivahendite ja masintõlkeprogrammide arendamiseks, kuid nad leiavad kasutamist ka lingvistikas.

Oodatavad tulemused: võimalikult erinevate paralleelkeeltega (mitte ainult inglise-eesti, vaid ka saksa-eesti, prantsuse-eesti, soome-eesti ja vene-eesti) suurte paralleelkorpuste loomine.

3.2.1.3. Süntaktiliselt analüüsitud korpus on vajalik kõigi süntaktilist analüüsi eeldavate rakenduste arendamiseks. Praegu on olemas 200 000 sõna suurune pindsüntaktiliselt analüüsitud korpus ja 350 lause suurune märgendatud süvastruktuuriga korpus (puude pank).

Oodatavad tulemused: märgendatud süvastruktuuriga korpuse (puude panga) arendamine 100 000 sõnani.

3.2.1.4. Semantiliselt ühestatud ja märgendatud korpus on vajalik semantilise analüsaatori ja automaatse ühestaja loomiseks ning treenimiseks, samuti leksikaalsemantilise andmebaasi täiendamiseks. Praegu on olemas 100 000-sõnaline korpus.

Oodatavad tulemused: korpuse täiendamine 300 000 sõnani.

Lisaks nimetatutele läheb mitmete keeletöötlemise etappide automatiseerimiseks vaja veel spetsiaalselt märgendatud korpusi (diskursus, anafoorid jms).

3.2.1.5. Vigade korpus on paralleelkorpus, kus on paralleelselt esitatud grammatilise veaga lause ja sama lause õige vaste. Vigade korpus on vajalik grammatikakorrektori ja keeleõppeprogrammide loomiseks. Olemas on 50 000 sõnast koosnev korpus

Oodatavad tulemused: vigade korpuse täiendamine 200 000 sõnani.

3.2.1.6. Korpuste kasutajaliides(ed) on vajalikud selleks, et muuta korpused Interneti kaudu kättesaadavaks kõigile soovijatele. Samuti tuleb olemasolevad korpused ühtlustada ja standardida.

Oodatavad tulemused: kirjaliku keele korpuste jaoks selliste kasutajaliideste loomine, mis võimaldaksid esitada korpustele päringuid ja teha lihtsamaid statistilisi analüüse.

3.2.2. SUULISE KEELE KORPUSED

3.2.2.1. Suulise eesti keele korpus on koondkorpus suulise keelekasutuse ja kommunikatsiooni uurimiseks. Suulise eesti keele korpuse põhjal luuakse mitmeid erinevateks uurimisülesanneteks vajalikke alamkorpuseid. Praegu sisaldab korpus 1100 transkribeeritud teksti, kokku 700 000 sõnavormi.

Oodatavad tulemused: suulise eesti keele korpuse mahu suurendamine 2 miljoni sõnavormini.

Suulise eesti keele korpuse alamkorpused:

3.2.2.2. Dialoogikorpus, mida on vaja nt loomulikkult keelt võimaldavate suhtlusprogrammide väljatöötamiseks. Dialoogikorpuses on laused märgendatud kõneaktide terminites ja see on vajalik nt kõneaktide tuvastaja väljatöötamiseks ja treenimiseks. Praegune korpus sisaldab 100 000 sõnavormi.

Oodatavad tulemused: dialoogikorpuse mahu suurendamine 500 000 sõnavormini.

3.2.2.3. Kõnepuudega inimeste kõne erikorpus, mille abil saaks uurida kõne- ja suhtluspuude vahelisi seoseid ning töötada välja suhtlusprobleeme leevendavaid abivahendeid.

Oodatavad tulemused: kõnepuudega inimeste kõne erikorpuse loomine mahuga 10 000 sõnavormi.

3.2.2.4. Segmenteeritud sidusa kõne (dialoogid, ettelõetav tekst, spontaanne kõne) korpus, mille abil saaks modelleerida kõneprosoodiat ja mis on vajalik kõnesünteesi ja kõnetuvastuse moodulite väljatöötamiseks.

Oodatavad tulemused: ühtsel kodeerimis-, transkribeerimis- ja segmentimisel koostatud kõnekorpus, mis koondaks vähemalt 300 inimese (150 meest ja 150 naist) märgendatud salvestisi.

3.2.2.5. Kõnetehnoloogia andmebaasid:

1) Difoonide andmebaasid on vajalikud eestikeelse kõnesünteesi arendamiseks. Praegu on olemas ühe meeshääle salvestustest loodud andmebaas.

Oodatavad tulemused: naishääle sünteesiks vajaliku difoonide andmebaasi koostamine.

2) Kõnetuvastuse andmebaas on vajalik eestikeelse kõnetuvastuse uuringuteks ning tuvastussüsteemide treenimiseks ja testimiseks. Andmebaas peab sisaldama paljude inimeste erinevaid kõnenäiteid (teksti ettelugemine, spontaanne kõne, dialoog), mis on salvestatud ühesugustes akustilistes tingimustes (kajavaba ruum, kvaliteetne mikrofoni).

Oodatavad tulemused: salvestada ja märgendada vähemalt 200 inimese (100 meest ja 100 naist) kõnenäiteid (igalt inimeselt kuni 30 minutit kõnet).

Erinevate kõnetuvastussüsteemide arendamiseks on vaja eri tüüpi korpusi:

3) Uudiste korpus on vajalik kõne automaatse transkribeerimis- ja dikteerimissüsteemi väljaarendamiseks. Uudiste korpus sisaldab raadiodiktorite loetud uudiste salvestisi.

Oodatavad tulemused: salvestada ja märgendada mitme erineva diktori loetud uudiseid 30 tunni ulatuses.

3.2.2.6. Aktsendikorpus on vajalik aktsendinähtuste akustiliseks uurimiseks ja modelleerimiseks. Samuti on seda vaja aktsendiga kõne akustiliste mudelite loomiseks kõnetuvastussüsteemide tarvis ning eesti keele kui võõrkeele õpetamisel. Aktsendikorpus sisaldab kõnesalvestisi eri emakeelega isikute eesti keele hääldusnäidetest.

Oodatavad tulemused: salvestada ühtsel printsiibil valitud keelenäiteid vene, saksa, prantsuse, rootsi, soome, inglise ja muu emakeelega eesti keele kõnelejatelt (igalt kõnelejalt vähemalt 15 minutit kõnet).

3.2.3. ELEKTROONILISED SÕNASTIKUD JA ANDMEBAASID

3.2.3.1. Elektroonilisi sõnastikke kasutatakse nii veebis kui ka keeletehnoloogia rakendustes. Elektrooniliste sõnastike standardsüsteemil on kaks põhikomponenti: (a) traditsiooniliste sõnastike formaliseeritud versioonid ning (b) eesti lähtekeelega kakskeelsete sõnastike andmebaas, mis on aluseks uute tõlkesõnastike koostamisel. Eestis praegu koostatavad sõnaraamatud on küll elektroonilisel kujul, kuid sageli ainult küljendusformaadis, ilma struktuurimärgendusega, mis teeb raskeks nende kasutamise veebis ja keeletehnoloogias.

Oodatavad tulemused: XML-põhised märgendusstandardid eri tüüpi sõnastike jaoks ning vastavalt nendele märgendatud eesti ükskeelsed ja kakskeelsed baassõnastikud.

3.2.3.2. Eesti keele leksikaalsemantiline andmebaas on vajalik nt infootsingus, aga ka masintõlkes ning refereerimis- ja sisukokkuvõtte süsteemides. Eesti leksikaalsemantiline andmebaas sisaldab praegu umbes 15 000 sõna.

Oodatavad tulemused: andmebaasi täiendamine 100 000 sõnani, arendades seejuures eriti nende valdkondade sõnavara, mille jaoks luuakse infootsi- ja muid süsteeme.

3.2.3.3. Püsiühendite andmebaas on vajalik süntaktilise ja semantilise analüsaatori töö parandamiseks. Praegu sisaldab andmebaas, mis on loodud peamiselt sõnaraamatute andmete põhjal, umbes 20 000 kirjet.

Oodatavad tulemused: täiustada püsiühendite andmebaasi ja luua programm, mis märgendaks teksti püsiühendid.

3.2.4. FORMAALSED KEELEKIRJELDUSED

3.2.4.1. Formaalset keelekirjeldused. Eesmärgiks on morfoloogilise, sõnamoodustusliku, süntaktilise, semantilise ja pragmaatilise analüüsi ning sünteesi aluseks olevate kirjeldusmeetodite ja formaalsete grammatikate väljatöötamine. Sellised formaalsed keelekirjeldused moodustavad keeletehnoloogias ühe osa nn korduvkasutatavatest keeleressurssidest: vastava keeletasandi kirjeldus on aluseks kõigile rakendustele, mis kasutavad selle tasandi analüüsi- või sünteesiprogramme.

Oodatavad tulemused: olemasolevate formalismide edasiarendamine; eesti keele sõnamoodustuse formaalse grammatika väljatöötamine; eesti keele süvasüntaktilise (sh puude panga) esituse aluseks oleva formalismi väljatöötamine; sõnade ja lausete tähenduste esitusformalismide väljatöötamine; dialoogi formaalse mudeli väljatöötamine rakendusteks valitud valdkondade jaoks.

4. PROGRAMMI FINANTSEERIMINE

Tabel 1. Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006-2010)" finantseerimisvajadus (tuhandetes kroonides).

Alaeesmärk / Aasta	2006	2007	2008	2009	2010
Keeletarkvara	4850	4850	7160	8480	9800
Keeletehnoloogilised ressursid	2450	2450	3640	4320	5000
Infrastruktuuri ajakohastamine*	0	8000*	6000*	4000*	2000*
Programmi halduskulud	150	150	200	200	200
Kokku	7 450	15 450	17 000	17 000	17 000

* Infrastruktuuri ajakohastamise vahendid on planeeritud taotleda tõukefondidest.

Märkus. Programmi juhtkomitee võib igal aastal vastavalt vajadusele moodulite mahtu muuta.

Tabel 2.. Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006-2010)" põhjal Eesti keeletehnoloogia infrastruktuuri kaasajastamise finantseerimisvajaduse (tuhandetes kroonides, mis on kajastatud tabeli 1 Infrastruktuuri ajakohastamise real) jagunemine riist- ja tarkvara vahel (prognoos).

	Protsent	Fin. vajadus
Riistvara	38 %	7 600
Tarkvara	62 %	12 400
Kokku	100 %	20 000

Indrek Reimand
Haridus- ja Teadusministeeriumi teadusosakond

TV leping 05.11.2004 nr 1-10/P2-9724

KOOPIA



Agreement concerning collaboration with the Nordic Graduate School of Language Technology
(NGSLT)

between

NGSLT and University of Tartu

The Nordic Graduate School of Language Technology is devoted to the promotion advanced training in Language Technology in the Nordic countries, the Baltic States and NW Russia. It has the following goals

- to provide a forum for graduate students in language technology in which they can obtain advanced research training of a standard and breadth which is not available at any one of the individual participating institutions, not even in any single Nordic country alone to raise the general standard of language technology education in Scandinavia in order to meet the increasing need in industry and academic research of researchers and developers with competence in language technology
- to create a broad interdisciplinary platform for graduate education in language technology. This platform should provide a multidisciplinary basis on which the student can build further
- to create an international profile by inviting instructors from foreign universities and research institutes and by encouraging interaction between students from the Nordic countries and their peers in other countries (including nonNordic countries)
- to exploit similarities between many of the Nordic languages by encouraging students to reuse linguistic resources and technologies designed for one Nordic language as a base for creating similar tools for the other Nordic languages

We agree to collaboration on working towards these goals, giving eligible students the opportunity to apply for financial support to take part in NGSLT activities and for university faculty to collaborate with NGSLT in the design and execution of the school's events.

Robin Cooper

Professor of Computational
Linguistics
Göteborg University
co-director NGSLT

Kimmo Koskeniemi

Professor of Computational
Linguistics
Helsinki University
Co-director NGSLT

Jaak Kangilaski

Vice Rector
University of Tartu

28 October 2004

Postadress	Besöksadress	Telefon
Box 200, S-405 30 GÖTEBORG, Sweden	Humanisten Renströmsgatan 6	Nat 031-773 1000 växel 031-773 5916 direktval Int +46 31 773 1000 Fax: 031-773 4853