Lisa 1
KINNITATUD
haridus- ja teadusministri 25. jaanuari 2011. a
käskkirjaga nr 71

Riiklik programm

Eesti keeletehnoloogia 2011-2017

Sissejuhatus

Keeletehnoloogia on infotehnoloogiat ja keeleteadust ühendav interdistsiplinaarne valdkond, mis tegeleb inimkeele arvutitöötluseks vajaliku keeletarkvara ning keeleressursside väljatöötamisega. **Keeletarkvara** hõlmab keelematerjali töötlemise meetodeid, algoritme ja arvutiprogramme ning on aluseks keeletehnoloogilistele rakendussüsteemidele. Otstarbekas on eristada **kõnetehnoloogiat** (nt kõnetuvastus ja -süntees) ning **kirjalike tekstide töötlemise tehnoloogiat** (nt morfoloogiline, süntaktiline ja semantiline analüüs) ehk keeletehnoloogiat kitsamas mõttes. **Keeleressursid** on elektroonilised andmekogud, mida kasutatakse keeletarkvara väljatöötamiseks: korpused (kõnesignaalide ja tekstide kogumid), elektroonilised sõnastikud ja andmebaasid.

Eesti keele keeletehnoloogilise toe arendamine on vajalik oma keelelise ja kultuurilise identiteedi säilitamiseks EL mitmekeelses keskkonnas ja suhtluseks infoühiskonnas võrdsel tasemel tehnoloogiliselt arenenud keeltega.

Käesolev programm on mõeldud jätkuprogrammina riiklikule programmile „Eesti keele keeletehnoloogiline tugi (2006-2010)“, mis andis märkimisväärse panuse keeletehnoloogia arengule Eestis, kuid selle raames loodud keeleressursid ja programmide prototüübid ei ole veel piisavad tagamaks eesti keele funktsioneerimist e-keskkonnas. Jätkuprogramm eristub eelnevast programmist selle poolest, et lisaks tarkvaraprototüüpide ja keeleressursside arendamisele pööratakse suurt tähelepanu keeletehnoloogia rakenduste loomisele ja olemasolevate ning loodavate ressursside ning tarkvara kättesaadavaks tegemisele.

Üldised suundumused keeletehnoloogilise tarkvara ja selle rakenduste ning keeleressursside alal on järgmised:

- kõnetehnoloogias: 1) kõnetuvastuse tase võimaldab seda rakendada väljaspool laboritingimusi, 2) kõnesünteesi kvaliteet läheneb loomuliku kõne omale;
- keeletehnoloogias: 1) automaatne analüüs ja süntees keele kõigil tasandeil (morfoloogia, süntaks, semantika, pragmaatika) 2) rakendussüsteemide (infootsing, masintõlge, tõlkija abivahendid, leksikograafi ja terminoloogi töökeskkond jm) olemasolu;
- integreeritud rakendustes: mitmesuguste keele- ja kõnetehnoloogiliste vahendite integreeritus lõppkasutajale vajalike rakenduste loomiseks;
- keeletehnoloogiliste ressursside alal on vajalik tarkvaraarenduse ja rakendusprojektide vajadustele vastavate korpuste ja sõnastike olemasolu.

Riikliku programmi eest vastutav ministeerium on Haridus- ja Teadusministeerium, avaliku sektori võimaliku partnerina tehakse koostööd eelkõige Majandus- ja Kommunikatsiooniministeeriumi Riigi Infosüsteemide Osakonnaga.

Programmi täitjatena on partneriteks eelkõige Tartu Ülikool, Tallinna Tehnikaülikooli Küberneetika Instituut ja Eesti Keele Instituut, aga ka teised teadus- ja arendusasutused ning infotehnoloogiaettevõtted.

Taust

Euroopa Liidu üheks põhivääruseks on keeleline ja kultuuriline mitmekesisus ning selle säilitamiseks ja kultuuridevahelise dialoogi arendamiseks on vastu võetud mitmeid EL õigusakte ja toetatakse ka projekte keeletehnoloogiate arendamiseks ning levitamiseks[1]. Kuigi kõik EL ametlikud keeled on võrdväärsed, domineerivad ärimaalimas ja tehnoloogiaarenduses suuremad keeled, sest vähese kõnelejate arvuga keelte puhul ei ole keeletehnoloogiline arendustöö majanduslikult otstarbekas. Nii tõdeb EL keeletehnoloogia vajadusi kaardistav raport[2], et keeletehnoloogia kui ärivaldkonna perspektiivist lähtudes võib eristada esma-, teise- ja kolmandajärgulisi keeli. On ilmne, et keelte võrdsuse tagamine EL asjaajamises läheb maksumaksjale kalliks maksma – 23 ametliku keele puhul on vaja tagada tõlge 506 keelepaari vahel. Hinnanguliselt kulub ainuüksi tõlketeenusteks ligikaudu 1% EL eelarvest, näiteks 2007.a oli see ligikaudu 1 miljard eurot; prognoosi kohaselt kasvavad tõlkekulud umbes 5% aastas[3].

Keeletehnoloogia arendus on kõigi keelte puhul ühtemoodi kallis, pole olulist vahet, kas seda tehakse inglise või eesti keele jaoks. Kõigi EL ametlike keelte tehnoloogilise taseme viimine inglise keelega samale tasemele on ääretult kallis ja on vältimatu, et sellesse panustavad nii Euroopa Komisjon kui ka liikmesriigid ise (kooskõlas subsidiaarsuse printsiibiga).

---

[1] **KOMISJONI TEATIS EUROOPA PARLAMENDILE, NÕUKOGULE, EUROOPA MAJANDUS- JA SOTSIAALKOMITEELE NING REGIOONIDE KOMITEELE. Mitmekeelsus: Euroopa rikkus ja ühine kohustus.** Brüssel, 18.9.2008, KOM(2008) 556.

[2] G. Lazzari. Human Language Technologies for Europe. ITC IRST/TC-Star project report, 2006. http://www.european-journalists.eu/Human%20Language%20TechHnolLogies%20for%20Europe%20-%20TC-STAR.pdf

[3] http://www.euractiv.com/en/culture/eu-translation-policy-stay/article-170516

Keeletehnoloogia seis Eestis 2010

„Eesti keele arendamise strateegia 2004−2010" ülesanded eesti keele keeletehnoloogilise toe loomisel on täidetud: eesti keel kuulub 50 kõrgelt arendatud keeletehnoloogiaga keele hulka maailmas. Selle saavutamisele aitasid kaasa riiklikud programmid „Eesti keel ja rahvuslik mälu (2004–2008)", mis rahastas keeletehnoloogiatööd aastail 2004−2005, „Eesti keele keeletehnoloogiline tugi (2006−2010)" ning ka Tartu Ülikooli juures 2005−2008 tegutsenud doktorikool „Keeleteadus ja -tehnoloogia". Enne 2004. aastat toetasid eesti keele keeletehnoloogilise toe loomist Informaatikakeskus ning riiklik programm „Eesti keel ja rahvuskultuur" (1999-2003).

Erinevate keeleressursside maht ja mitmekesisus ning loodud tarkvaraprototüüpide hulk ja kvaliteet pole siiski veel tasemel, mis võimaldaks keeletehnoloogia laialdast rakendamist e-keskkonnas.

Olukord, probleemid ja vajadused kõrghariduses

Eestis on olemas spetsialistide kriitiline mass keeleressursside, keeletarkvara ja selle rakenduste loomiseks ning võimalused uute spetsialistide koolitamiseks Tartu Ülikooli eesti ja soome-ugri keeleteaduse (spetsialiseerudes arvutilingvistikale) ja infotehnoloogia (spetsialiseerudes magistriõppes keeletehnoloogiale) erialadel. 2009. aastal on käivitatud kaks uut doktorikooli, kus osalevad arvutilingvistika ja keeletehnoloogia doktorandid: info- ja kommunikatsioonitehnoloogia doktorikool (haarab ka keeletehnolooge) ning keeleteaduse, filosoofia ja semiootika doktorikool, mis üldkeeleteaduse raames haarab ka arvutilingviste.

Probleemiks on kõnetehnoloogia spetsialistide ettevalmistus – TTÜ infotehnoloogia teaduskonna bakalaureuse-, magistri- ja doktoriõppekavades on üksikuid signaali- ja kõnetöötluse valikkursusi ning nende eeldusainetena vajalikku matemaatilist baasi pakkuvaid õppeaineid, kuid puudub süstemaatiline süvaõpe kõnesignaalide analüüsi, sünteesi ja tuvastuse alal.

Eesti keele arengukavas (2011-2017) püstitatud eesmärgi – eesti keele keeletehnoloogiline tugi on võrdsel tasemel arenenud keeletehnoloogiaga riikide (nt Põhjamaad) keeltega – saavutamiseks ja keeletarkvara kasutavate info- ja kommunikatsiooniteenuste loomiseks on vajalik säilitada ja tagada piisava hulga uurijate ja arendajate ettevalmistus nii kirjaliku kui suulise keele tehnoloogia alal. Otstarbekas oleks spetsialistide ettevalmistamine korraldada TÜ ja TTÜ koostöös..

Olukord, probleemid ja vajadused teaduses

Eestis on kolm peamist keeletehnoloogia uurimis- ja arendustööga tegelevat keskust:

(1) Tartu ülikooli arvutilingvistika töörühm, mille olulisemateks uurimissuundadeks on morfoloogiline, süntaktiline, semantiline ja pragmaatiline analüüs; suuline keel ja dialoogimudelid, masintõlge ning vastavate keeleressursside (kirjaliku keele korpused,

semantilised andmebaasid, dialoogi- ja spontaanse kõne korpused, paralleelkorpused) loomine. Keeletarkvara ja –ressursse on arendatud ka teistes uurimisrühmades: bioinformaatika töörühmas (hägus infootsing, tekstialgoritmid) ja foneetika uurimisrühmas (spontaanse kõne foneetiline andmebaas).

(2) Eesti Keele Instituudi keeletehnoloogia töörühma uurimis- ja arendustöö hõlmab leksikograafi töövahendeid ning eesti keele sõnavara andmebaase, tekst-kõnesünteesi meetodeid, sealhulgas ka emotsionaalse kõne analüüsi ja sünteesi. EKIs töötab keeletehnoloogia valdkonnas 16 inimest, neist 12 täis- ja 4 osalise koormusega.

(3) TTÜ Küberneetika instituudi foneetika ja kõnetehnoloogia labori peamisteks suundadeks on kõnetuvastus ja -süntees, kõnevariatsioonide eksperimentaal-foneetiline uurimine ja erinevate kõnekorpuste (loetud kõne, spontaanne ja dialoogkõne, raadiouudised ja vestlussaated, aktsendiga kõne, akadeemilised loengud jm) loomine.Aastatel 2006-2010 tegeles valdkonnaga asutuses pidevalt 2 vanemteadurit ja 1 teadur/doktorant; lühiajaliselt 1 teadur ja 2 doktoranti; käsunduslepingutega 1 magistrant, 2 üliõpilast ja 2 muud töötajat.

Need kolm keskust on andnud olulise panuse EKKTT (2006-2010) eesmärkide saavutamisse, üksikuid keeletehnoloogia projekte on teostatud ka Tallinna ülikoolis (Eesti vahekeele korpus) ja Eesti Kirjandusmuuseum (fraseologismide elektroonilise alussõnastiku loomine).

Keeletehnoloogia arendustöö jaoks oluliste keele-spetsiifiliste teoreetilise mudelite ja teadmiste arengusse annavad suure panuse sihtfinantseeritavad teadusteemad Tartu Ülikoolis, Eesti Keele Instituudis ja TTÜ Küberneetika Instituudis. Loodud on „Arvutiteaduse tippkeskus" (2008–2015), milles osalevad ka keeletehnoloogid Tartu Ülikoolist ja Tallinna Tehnikaülikooli Küberneetika Instituudist. Eesti keeletehnoloogid osalevad regulaarselt rahvusvahelistel konverentsidel ja mitmetes EL võrgustikes (CLARIN, META-NET, HEXA-NORD jm).

Teadustööks vajalik infrastruktuur on kõigis uurimisrühmades suhteliselt heal tasemel. Uurimisrühmade koostööd nii Eesti-siseselt kui rahvusvaheliselt aitab parandada Eesti teaduse infrastruktuuri teekaardi objektina käivitatav Eesti Keeleressursside Keskus, mille kaudu tehakse uurijatele kättesaadavaks partnerite loodud keeleressursid ja tehnoloogiad.

Probleemiks on uurijate ja arendajate-programmeerijate vähesus eelkõige kõnetehnoloogia alal, juurde oleks vaja vähemalt 2-3 teadurit ja 3-4 inseneri/programmeerijat. Puudu on programmeerijatest ka elektroonilise leksikograafia ülesannete täitmiseks, näiteks EKIs vajatakse lisaks vähemalt 2 täiskohaga programmeerijat, üks programmeerimisoskusega teadustöötaja aga läheb peagi pensionile. Teadustöö taseme säilitamiseks on vaja jätkuvalt investeerida uurimisrühmade infrastruktuuri kaasajastamisse ja uurijate ettevalmistusse.

Olukord, probleemid ja vajadused ettevõtluses

Eestis on üksikud keeletehnoloogia valdkonnas tegutsevad firmad:

- Filosoft (http://www.filosoft.ee, asutatud 1993) – põhilised tooted on eesti keele speller, poolitaja ja tesaurus erinevate operatsioonisüsteemide (MS Windows, Unix, Mac OS) ja tarkvarapakettide (MS Office, OpenOffice, Lotus Notes) jaoks. Filosoft haldab ka internetiportaali Keeleveeb (http://www.keeleveeb.ee), mille kaudu on tehtud vabalt kasutatavaks mitmed sõnastikud ja tekstikorpused.
- Keelevara (http://www.keelevara.ee, asutatud 2004) veebiportaali kaudu on kasutatavad mitmed eesti keele sõnaraamatud, tõlke- ja oskussõnastikud ning nimede andmebaasid.
- Tilde Eesti (http://www.tilde.ee) on 1991.a asutatud Läti firma Tilde haru Eestis. Tilde on Baltikumi suurim ja tuntuim keeletehnoloogia firma, mille peamisteks toodeteks on Ida-Euroopa keeltele kohandatud tekstitöötlus-vahendid (lokaliseeritud fondid, spellerid, grammatika kontrollijad), elektroonsed sõnaraamatud, tõlkija abivahendid jm. Tilde Eesti pakub peamiselt tarkvara lokaliseerimis- ja tõlketeenuseid.
- Tarkvara Tehnoloogiate ja Rakenduste Arenduskeskus OÜ (http://www.stacc.ee) on Ettevõtluse Arendamise Sihtasutuse tehnoloogia arenduskeskuste programmi (2009-2015) raames rahastatav üksus, mille üheks töösuunaks on keeletehnoloogia meetodite arendus meditsiinivaldkonna tekstide analüüsiks.

Eesti IKT firmade vähene huvi arendada keeletehnoloogilisi rakendusi on tingitud eelkõige sellest, et keeletehnoloogiline arendustöö baseerub pikaajalistel teadusuuringutel ning nõuab suuri investeeringuid, mis eestikeelse turu suurust arvestades pole majanduslikult otstarbekad. Samas on keeletehnoloogiliste rakenduste loomine ja seeläbi eesti keelt toetava e-keskkonna arendamine eelkõige just firmade, mitte uurimisgruppide ülesanne. Riikliku programmi raames loodavate keeleressursside ja tarkvara prototüüpide kättesaadavus on oluliseks eelduseks eesti keelt toetavate lõpptarbijale orienteeritud rakenduste loomisele.

Eesti keeletehnoloogia võrdluses maailma ja EL suundumustega

Euroopal Liidu seisukohad keelelise ja kulutuurilise mitmekesisuse väärtustamiseks ja arendamiseks on esitatud Euroopa Ühenduste Komisjoni mitmetes dokumentides, nt "Uus mitmekeelsuse raamstrateegia"[4] (2005) ja "Mitmekeelsus: Euroopa rikkus ja ühine kohustus"[5] (2008). Muude meetmete (keeleõppe tõhustamine, meediatoodete subtiitritega varustamine, mitmekeelse multimeedia infosisu loomine jne) kõrval pööratakse neis suurt tähelepanu ka keeletehnoloogia arendamise vajadusele. Lisaks komisjoni kavandatavatele meetmetele, kutsutakse liikmesriike üles toetama keeletehnoloogia arendust ja käivitama riiklikke kavasid, mis aitaksid mitmekeelsuse edendamise meetmeid struktureerida, sihipärastada ja kooskõlastada.

EL infoühiskonna tehnoloogiaprogrammi raames toetatakse teadusuuringuid, mis käsitlevad keelebarjääride ületamist uute info- ja kommunikatsioonitehnoloogiate abil:

- abivahendid tõlkijate töö tõhususe suurendamiseks (tõlkemälud, võrgusõnaraamatud ja tesaurused), pool- ja täisautomaatsed tõlkesüsteemid;
- kõnetuvastus ja -süntees, dialoogsüsteemid.

Komisjoni teatis (2008) sedastab, et "seoses üleilmastuva Interneti-põhise majanduse ja kasvava infotulvaga kõigis võimalikes keeltes on tähtis, et kodanikel oleks teabele juurdepääs ja võimalus kasutada teenuseid riigipiiridest ja keeletõketest olenemata, abiks Internet ja mobiilseadmed. Info- ja kommunikatsioonitehnoloogias tuleb võtta arvesse keeleküsimusi ja edendada lahenduste sisu loomisel paljude keelte kasutamist."

Euroopa infoühiskonna raamdokumendi "i2010 – Euroopa infoühiskond majanduskasvu ja tööhõive eest" (2005)[6] ja "Euroopa digitaalne tegevuskava"[7] (2010) seavad eesmärgiks Euroopa ühtse kõiki kodanikke kaasava inforuumi loomise, milleks kavandatakse tegevusi info- ja kommunikatsioonitehnoloogia teenuste arendamiseks ja selleks vajalike investeeringute suurendamiseks. Ühtse inforuumi loomiseks on vajalik ka keeletehnoloogia arendus, mis aitab ületada

[4] KOMISJONI TEATIS EUROOPA PARLAMENDILE, NÕUKOGULE, EUROOPA MAJANDUS- JA SOTSIAALKOMITEELE NING REGIOONIDE KOMITEELE: Uus mitmekeelsuse raamstrateegia, 22.11.2005, KOM(2005) 596.

[5] KOMISJONI TEATIS EUROOPA PARLAMENDILE, NÕUKOGULE, EUROOPA MAJANDUS- JA SOTSIAALKOMITEELE NING REGIOONIDE KOMITEELE: Mitmekeelsus: Euroopa rikkus ja ühine kohustus, Brüssel, 18.9.2008, KOM(2008) 556.

[6] KOMISJONI TEATIS EUROOPA PARLAMENDILE, NÕUKOGULE, EUROOPA MAJANDUS- JA SOTSIAALKOMITEELE NING REGIOONIDE KOMITEELE: i2010 -- Euroopa infoühiskond majanduskasvu ja tööhõive eest, Brüssel, 01.06.2005, KOM(2005) 229.

[7] KOMISJONI TEATIS EUROOPA PARLAMENDILE, NÕUKOGULE, EUROOPA MAJANDUS- JA SOTSIAALKOMITEELE NING REGIOONIDE KOMITEELE: Euroopa digitaalne tegevuskava, Brüssel, 26.8.2010, KOM(2010) 245.

keelebarjääre ja tagada ligipääsu mitmekeelsele informatsioonile ning teenustele ja soodustab kõigi kodalike osalust e-keskkonnas.

„Euroopa digitaalse tegevuskava" kohaselt esitab komisjon 2010. aastal põhjaliku teadus- ja innovatsioonistrateegia, mis moodustab suurprojekti "Innovaatiline liit", et rakendada Euroopa 2020. aasta strateegia.[8] Euroopa peab oma IKT-alase juhtkoha saavutamise strateegia[9] alusel suurendama, suunama ja koondama oma investeeringuid, et säilitada konkurentsivõimet selles valdkonnas, ja jätkama investeeringute tegemist kõrge riskiga teadustegevusse, sealhulgas mitut valdkonda hõlmavatesse alusuuringutesse. Komisjon kavandab selleks muuhulgas järgmisi meetmeid:

- tagab piisava finantstoetuse ühistele IKT teadusinfrastruktuuridele ja innovatsioonikeskustele, arendab edasi e-infrastruktuuri ja kehtestab Eli strateegia pilvandmetöötluse jaoks eelkõige valitsuse ja teaduse tarbeks;
- teeb koostööd sidusrühmadega, et töötada välja veebipõhiste rakenduste ja teenuste, sealhulgas mitmekeelse infosisu ja võrguteenuste uus põlvkond, toetades standardeid ja avatud platvorme ELi rahastatavate programmide kaudu.

Liikmesriigid peaksid:

- kahekordistama 2020. aastaks IKT teadus- ja arendustegevuseks ettenähtud riiklikke kogukulutusi 5,5 miljardilt eurolt 11 miljardile eurole (mis sisaldab ELi programme) selliselt, et ka erasektori kulutused suureneksid 35 miljardilt eurolt 70 miljardile eurole;
- osalema laiaulatuslikes katseprojektides, et katsetada ja arendada innovaatilisi ja koostalitlusvõimelisi lahendusi avalikku huvi pakkuvates valdkondades, mida rahastab CIP (konkurentsivõime ja uuendustegevuse raamprogramm).

Euroopa Liidus on käivitatud mitmed üle-euroopalised võrgustikud, nt CLARIN, META-NET jt, mille eesmärgiks on luua infrastruktuur keeleressursside ja -tehnoloogiate kättesaadavuse tagamiseks ning keeletehnoloogia arendamiseks kõigi EL keelte jaoks. EL 7.raamprogrammis on mitmeid alavaldkondi, mille raames finantseeritakse ka keeletehnoloogiat kaasavaid projekte, nt kognitiivsed süsteemid, suhtlus ja robootika; digitaalne sisu ja raamatukogud; jt. Samas ei ole viimased EL raamprogrammid (RP6 ja RP7) otseselt sisaldanud keeletehnoloogia ressursside ja baastarkvara arenduse temaatikat ega ka vähem arendatud tehnoloogilise toega keeltele suunatud tegevusi.

[8] The 2009 Report on R&D in ICT in the European Union. http://ftp.jrc.es/EURdoc/JRC49951.pdf

[9] Info- ja sidetehnoloogia teadus- ja arendustegevuse ning innovatsiooni strateegia Euroopas: suurendame panuseid, KOM(2009) 116.

Mitmetes EL riikides on viimasel paaril kümnendil käivitatud riiklike programme keeletehnoloogia arendamiseks, näiteks:

- Prantsusmaal 1994. a *Francil*-võrgustik ja 2002 a. *Techno-Langue French national program (http://www.technolangue.net)*;
- Hollandis alustati keeletehnoloogia-alast koordineeritud arendustööd 1999. a ja 2005. a käivitati Hollandi-Belgia ühisprogramm STEVIN (2005-2011) (http://taalunieversum.org/taal/technologie/stevin/) hollandi keele tehnoloogilise toe arendamiseks;
- 2000 a. loodi Põhjamaade keeletehnoloogia võrgustik (http://cst.dk/nordoknet) ja keeletehnoloogia uurimisprogramm (2000-2004) taani, islandi, norra, rootsi ja soome keele jaoks;
- Soome tehnoloogia-agentuuri TEKES keeletehnoloogiat sisaldavad programmid: USIX (1999-2002) ja FENIX (2003-2007) (selle raames otseselt keeletehnoloogiale suunatud alamprogramm PUMS);
- Rootsi innovatsiooni-agentuur VINNOVA finantseeris keeletehnoloogia-programmi aastatel 2001-2004.

Keeletehnoloogiat on maailmas arendatud enim inglise keele jaoks, tugeva tehnoloogilise toega on ka prantsuse, saksa, hispaania, itaalia, jaapani ja hiina keel. Suhteliselt heal tasemel on ka hollandi, taani, rootsi, soome, norra, tšehhi, ungari, poola, portugali, kreeka ja sloveenia keele tehnoloogiline tugi. Tehnoloogiliselt nõrgemad on bulgaaria, iiri, läti, leedu, malta, rumeenia ja slovaki keel.

Eestis vastuvõetud tegevuskavad eesti keele ja keeletehnoloogilise toe arendamiseks (Eesti keele arendamise strateegia (2004-2010), RP Eesti keele keeletehnoloogiline tugi (2006-2010)) on igati kooskõlas eelkirjeldatud EL suundumustega. EKKTT on andnud märkimisväärse panuse keeletehnoloogia arengusse – olulisel määral on kasvanud keeleressursside maht ja mitmekesisus ning prototüüpide hulk ja kvaliteet, mis kokkuvõttes lubab eesti keele paigutada hea tehnoloogilise toega keelte hulka.

Keeletehnoloogiat riiklikul tasemel koordineeritult arendavaid riike on teadaolevalt vähe ja seetõttu on Eesti keeletehnoloogia riiklik programm pälvinud tähelepanu mitmetel rahvusvahelistel konverentsidel (LREC 2008, NODALIDA 2009, BalticHLT 2010, EFNIL 2010).

Riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006-2010)“ eesmärgid ja tulemused

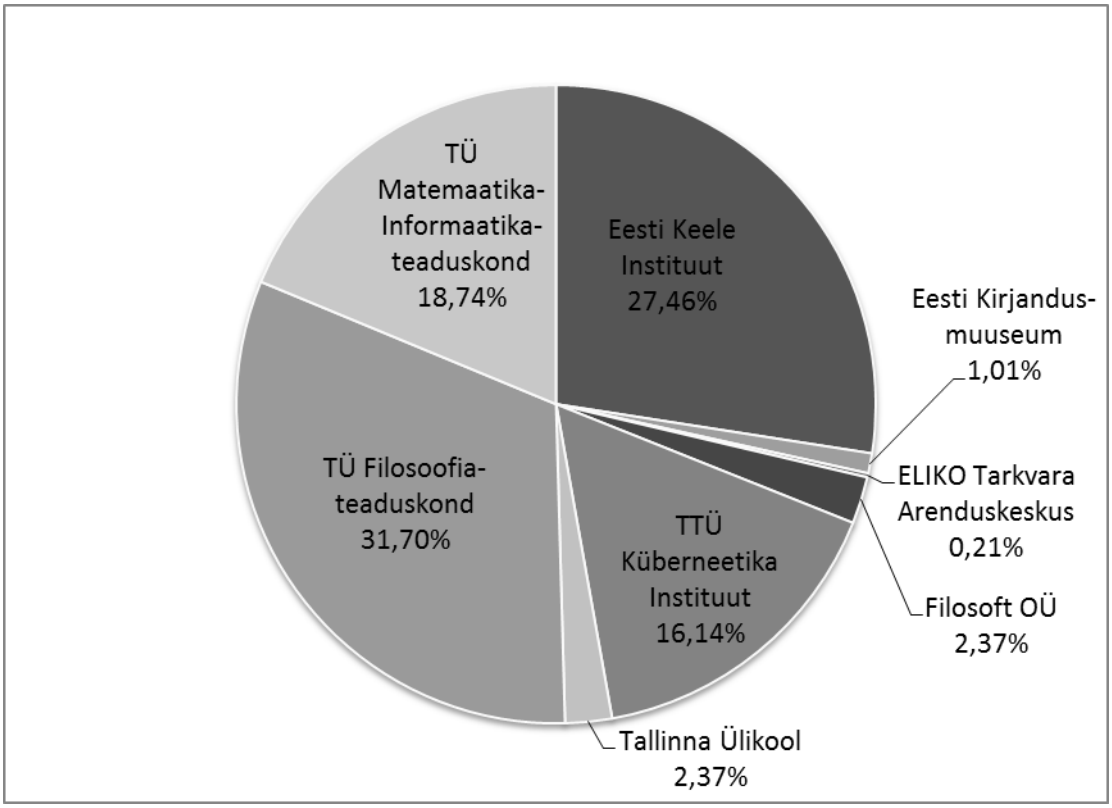
Riikliku programmi “Eesti keele keeletehnoloogiline tugi (2006–2010)” (EKKTT) peamiseks eesmärgiks oli eesti keele keeletehnoloogilise toe arendamine tasemele, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises keskkonnas. EKKTT rahastas keeletehnoloogiaalast teadus- ja arendustegevust alates ressursside loomisest kuni keeletehnoloogiliste rakenduste prototüüpide loomiseni.

Kõik EKKTT raames finantseeritud projektid on suunatud programmi peaeesmärgi saavutamisele ja programm tervikuna on olnud edukas. Aastatel 2006-2010 finantseeriti kokku 33 projekti, millest 18 projekti tegeles keeletehnoloogia meetodite uurimise ja tarkvaraprototüüpide loomisega, 14 projekti erinevate keeleressursside kogumise ja nende kasutajaliideste arendusega ning üks projekt keeleressursside ja –tarkvara halduseks vajaliku infrastruktuuri kavandamisega.

Tabel 1. Projektide arv ja finantseerimine aastate lõikes, sulgudes eraldi jätkuprojektide ja uute projektide hulk ja programmi rahastus ümberarvestatuna tuhandetesse eurodesse.

	2006	**2007**	**2008**	**2009**	**2010**
Finantseeritud projektide arv	**18**	**20** (18+2)	**23** (20+3)	**23** (15+8)	**24** (22+2)
Programmi rahastus, miljonit krooni (tuhandet eurot)	**7,3** (466,5)	**7,1** (455,7)	**13,4** (856,4)	**12,9** (842)	**11,8** (765,3)

Finantseerimise jaotus asutuste kaupa:



Juhtkomitee tõstab esile mõned projektid, mis iseloomustavad kõige eredamalt EKKTT 2006-2010 saavutusi.

EKKTT 2006-2010 edulood

Eestikeelne dialoog arvutiga (TÜ, projektijuht M.Koit)

Projektide *Eestikeelne infodialoog arvutiga ja Intelligentne kasutajaliides andmebaasidele* käigus on välja töötatud veebipõhine tarkvara, mis võimaldab adapteerumist erinevatele ainevaldkondadele 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, tõest infot. Intelligentses liideses on lõimitud eesti keele automaattöötluse vahendid: morfoloogiline analüüs ja süntees, õigekirjakontroll ja vigaste vormide korrigeerimine, ajaväljendite ja pärisnimede tuvastamine, tekst-kõnesüntees. Liidest on testitud kinoinfo ja hambaraviinfo andmebaasidega. Vt www.dialoogid.ee.

Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid (TÜ, projektijuht T.Roosmaa)

Eestikeelsete tekstide sisukokkuvõtja EstSum on orienteeritud veebi uudisartiklitest sisukokkuvõtete tegemisele ning on hetkel prototüüpversioon. Automaatne sisukokkuvõtete tegemine tekstist on protsess, mille käigus luuakse tekstist olemasoleva põhjal uus, lühendatud versioon, mis sisaldab ainult kasutajale vajalikku informatsiooni.

Loomulikult ei mõista arvuti nii töödeldava teksti sisu kui ka kasutaja ootusi sisukokkuvõttele. Seepärast kasutatakse erinevaid statistilisi ja lingvistilisi meetodeid, et leida tekstist lauseid, mis peaksid olema kõige ülevaatlikumad ja informatsioonirikkamad ning kasutajale esitatakse nende kogum kui sisukokkuvõte.

Kuigi üldiselt on vahendid ja meetodid sisukokkuvõtete automaatseks genereerimiseks keelest sõltumatud, ei saa siiski ühe keele jaoks loodud vahendeid kasutada teise keele jaoks sisukokkuvõtte genereerimiseks. Arvestama peab nii keele morfoloogilise kui ka süntaktilise eripäraga, kokkuvõttesse sobivate lausete standardväljenditega, aga ka lihtsalt leksikaalsete erinevustega.

Programmi saab katsetada veebilehelt: http://lepo.it.da.ut.ee/~kaili/estsum/

Korpusepäring keeleveebis (Filosoft, projektijuht H.-J.Kaalep)

Keeleveebis (http://www.keeleveeb.ee/) on tehtud tasuta kasutatavaks 30 erialasõnastikku kogumahuga 200 000 mõistet. Mõned olulisemad aspektid:

1. Kõik erialasõnastikud on kasutatavad ühispäringus, millesse on hõlmatud ka 30 keeleveebi-välist sõnastikku. See tähendab, et saab otsida sõna või terminit kuni 60 sõnastikust korraga.
2. Muuhulgas on välja pandud 14 põhikoolile mõeldud ainesõnastikku, mis koostati Haridus- ja Teadusministeeriumi tellimusel Tartu Ülikoolis ja anti välja aastal 2005. Sõnastikes on eestikeelne termin ja termini seletus ning venekeelne vaste. Mitme aine puhul on tegemist antud valdkonna esimese eestikeelse sõnastikuga.
3. Koos sõnastikupäringuga saab teha päringu ka TÜ Eesti keele koondkorpusest (http://www.cl.ut.ee/korpused/segakorpus/, 250 miljonit sõna), mis on keeleveebi panemise käigus morfoloogiliselt analüüsitud ja ühestatud, indekseeritud sõnavormi, lemma ja grammatilise info järgi. Sel moel saab lisaks sõnastikudefinitsioonile ka hulga kasutusnäiteid, mis võivad sõnastikuinfot täpsustada.

Leksikograafi töökeskkond (EKI, projektijuht Ü.Viks)

Projekti tulemusena on loodud veebipõhine leksikograafi töökeskkond (EELex), mis ühendab leksikograafile vajaliku tarkvara ja keeleressursid, toetab rühmatööd ja pakub eesti keele tuge. Professionaalse leksikograafi töökeskkonna baasil on selle kõrvale loodud EELexi avalik

laiatarbeversioon (http://exsa.eki.ee/), mille abil saab oma sõnastiku teha veebis ka tavakasutaja. EELexi töökeskkond muudab sõnastiku koostamise ja toimetamise töö lihtsamaks, kiiremaks ja kvaliteetsemaks.

Maailmas olemasolevatest elektroonilistest sõnastikusüsteemidest eristab EELexi eesti keele toe olemasolu (integreeritud automaatne morfoloogia, Eesti-X sõnastiku andmebaas), suur paindlikkus sõnastiku struktuuri suhtes, rikkalik valik toimetamise tööriistu ja vaba kasutus. Eesti Keele Instituudi sõnaraamatud valmivad kõik EELex süsteemis. Alanud on koostöö kirjastustega ja teiste asutustega, kus sõnastikke koostatakse või uuteks rakendusteks (nt telefonisõnastikud) ette valmistatakse – nii Eestis kui ka väljaspool.

Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine (Kübl, projektijuht T.Alumäe)

Projekti eesmärgiks oli eesti keele keelele sobivate kõnetuvastuse meetodite arendamine ja erinevate tuvastussüsteemi prototüüpide loomine, selle käigus uuriti järgmisi alateemasid: (1) millised on eestikeelse kõne tuvastuseks optimaalsed tuvastusühikud (difoonid, silbid, pseudo-morfeemid, jms), (2) erinevatel tuvastusüksustel baseeruvaid morfo-süntaktilisi keelemudelid ja statistiliste keelemudelite adapteerimise probleeme, (3) semantiliste seoste modelleerimist statistilises keelemudelis, (4) eesti keele väldete modelleerimist silbikestuste suhete alusel, (5) tehnoloogilisi lahendusi piiratud ja piiramatu sõnavaraga tuvastussüsteemide loomiseks.

Projekti tulemusteks on mitmed prototüübid:

- autosegmenteerija – tarkavara, mis automaatselt segmenteerib eestikeelset kõnet sõnadeks ja häälikuteks http://www.phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:est-align.et
- kõnesalvestuste (näit. raadiosaadete, konverentsiettekannete või loengu-salvestuste) täisautomaatne segmenteerimis- ja transkribeerimissüsteem,
- kõnesalvestuste transkriptsioonide sirvimise, indekseerimise ning otsingu veebirakendus http://bark.phon.ioc.ee/tsab/
- radioloogia valdkonna kõnetuvastussüsteem (koostöös Cybernetica AS-ga).

Eesti keeletehnoloogia teekaart (2011-2017)

2005.a koostati Eesti keeletehnoloogia teekaart (2004-2011), mis kirjeldas keeletehnoloogia seisu aastal 2004 ja prognoosis keeleressursside ja prototüüpide arendust kuni aastani 2011. Enamus prognoositud tulemustest on saavutatud või vastav uurimis- ja arendustöö on käimas, mõnede eesmärkide saavutamine on ülesannete keerukuse tõttu osutunud prognoositust tunduvalt aeganõudvamaks, näiteks dialoogiaktide automaatse tuvastuse, semantika-pragmaatika liidese, audio-visuaalse kõnesünteesi ja suulise kõne kasutust võimaldava dialoogsüsteemi prototüüpide loomine.

Keeletehnoloogia teekaart (2011-2017) fikseerib Eesti keeletehnoloogilised ressursid ja prototüübid 2010. a lõpu seisuga ja prognoosib arenguid järgnevaks seitsmeks aastaks. Olemasolevate keeleressursside ja prototüüpide ning vajaduste kaardistamiseks on läbi viidud keeletehnoloogia uurijate-arendajate ning potentsiaalsete kasutajate ankeetküsitlus, mille tulemused on aluseks uue teekaardi koostamisel. Teekaart kinnitatakse programmi lisana 2011. a jooksul.

Aluseks olevad õigusaktid ja programmiga seotud pikaajalised arengukavad

Eesti keele arengukava 2011-2017

Eesti keele arengukava 2011-2017 (EKA, VV korraldus 26.11.2010 nr 451) esitab peatükis „Eesti keele keeletehnoloogiline tugi" olukorra analüüsi ning esitab keeletehnoloogia arendamise eesmärgid, mõju ühiskonnale ning indikaatorid. EKA seab eesmärgiks:

- eesti keele keeletehnoloogiline tugi on võrdsel tasemel arenenud keeletehnoloogiaga riikide (nt Põhjamaad) keeltega suundades, mida nõuavad eesti keelele orienteeritud tarkvara arendused ja rakendused

Arengukava täitmise mõju ühiskonnale seisneb kahes peamises punktis:

- keelematerjali adekvaatselt töötlevad programmid on ühiskonnas laialdaselt kasutusel
- eesti keelt tunnustatakse arenenud keeletehnoloogiaga keelena, Eesti kuulub enim arenenud keeletehnoloogiaga riikide hulka

Eesti keele keeletehnoloogilise toe arendamiseks tuleb:

- toetada Eesti keeletehnoloogide osalemist rahvusvahelises tööjaotuses, avatud koodiga rakenduste loomist ning oma ressursside ja lahenduste protokollimist ja standardimist
- luua keskne depositoorium keeleressursside ja korduvkasutustarkvara haldamiseks

EKA täitmise indikaatorid on seotud võimalusega kasutada loomulikku kõnesünteesi, piiratud valdkondadele loodud kõnetuvastust, inimene-masin-dialoogisüsteemi ja masintõlke prototüüpi.

Infoühiskonna arengukava 2013

Arengukava sätestab, et infoühiskonna arendamisel tagatakse eesti keele ja kultuuri järjepidevus. Arengukava väljakutseteks on ligipääs arvutitele ja internetile, infoühiskonna infrastruktuur, interneti kasutamine majapidamistes ja ettevõtetes ning IKT sektori konkurentsivõime.
Infoühiskonna arenguvisiooni kohaselt on eesmärgiks kõiki ühiskonnaliikmeid kaasav ning nende elukvaliteeti tõstev arenev infoühiskond ja konkurentsivõimeline Eesti majandus, kus kasutades ratsionaalselt loodud ja loodavaid IKT lahendusi saavutamaks suuremat tootlikkust ning tööhõive määra.
Käesolev programm on heas kooskõlas arengukava ja selle eesmärkidega.

Eesti teadus- ja arendustegevuse strateegia „Teadmistepõhine Eesti 2007-2013"

Eesti teadus- ja arendustegevuse ning innovatsiooni strateegias on ette nähtud käivitada riiklikud teadus- ja arendusprogrammid sotsiaalmajanduslike probleemide lahendamiseks ja eesmärkide saavutamiseks iga Eesti elaniku jaoks tähtsust omavates sotsiaalmajanduslikes valdkondades, nagu näiteks infoühiskond ja selle seos eesti keele järjepidevuse tagamise ja edendamisega.

Seos teiste riiklike ja rahvusvaheliste programmidega

Eesti teaduse infrastruktuuride teekaart

Vabariigi Valitsuses kinnitatud Eesti teaduse infrastruktuuride teekaardi objektide loetelus on riiklikult olulise objektina kirjas ka Eesti Keeleressursside Keskus (EKRK), mis hakkab toimima hajusinfrastruktuurina TÜ, TTÜ ja EKI vahel sõlmitud konsortsiumlepingu alusel. EKRK saab CLARIN-ERICu (European Research Infrastructure Consortium) keskuseks Eestis.

Keeletehnoloogilise toe arendamisel võib oluliseks osutuda vajadus teha koostööd ka teise teaduse infrastruktuuri teekaardi objektiga "Eesti e-varamu ja kogude säilitamine".

Riiklikud programmid

EKKTT (2006-2010) on välja kasvanud riiklikust programmist "Eesti keel ja rahvuslik mälu (2004-2008)", mille jätkuprogramm "Eesti keel ja kultuurimälu" (2009-2013)" (EKKM) hõlmab kolme tegevussuunda: 1. Eesti keel, 2. Kultuurimälu, 3. Humanitaarväljaannete taseme tõstmine. 1. tegevussuunas „Eesti keel" on alategevusena 1.2 kirjas „Keeleteaduslike andmebaaside korrastamine, digiteerimine ja publitseerimine", kui ei toetata keeletehnoloogiliste lahenduste jaoks loodud andmebaase ja toetatavad tegevused ei tohi dubleerida keeletehnoloogia riikliku programmi tegevusi – selle viimase

kindlustamiseks tehakse koostööd EKKM juhtkomiteega ja ka tulevikus kontrollitakse, et EKKM taotluste osas ei oleks kattuvusi käesoleva programmi 2. alaeesmärgi taotlustega.

Eestikeelse terminoloogia toetamise riiklik programm (2008-2012) haarab muude tegevuste seas ka terminiloomet toetava IT-keskkonna ja üldkasutatava terminibaasi loomise Eesti Keele Instituudis. Programmi tekstis on soovitatud standardid kooskõlastada EKKTT juhtkomiteega, hetkel kujunenud olukord võiks soosida standardite ühtlustamist Eesti Keeleressursside Keskusega, mille konsortsiumi partnerite hulka hakkab kuuluma ka EKI.

CLARIN

CLARIN (www.clarin.eu) on ESFRI teekaardi objekt, mille eesmärgiks on luua üle-euroopaline juba eksisteerivate keeleressursside ja keeletehnoloogiate võrgustik, eesmärgiga teha need ressursid ja tehnoloogiad kättesaadavaks ja kasutatavaks kõigile teadlastele, eriti aga humanitaar- ja sotsiaalvaldkonna teadlastele. CLARINi eesmärgiks on pakkuda kõrgetasemelist teenust, mis ületaks keele- ja valdkondade piirid ning aitaks kaasa nii Euroopa mitmekeelse ja mitmekultuurilise rikkuse säilimisele kui ka selle aktiivsele kasutamisele. CLARIN püüab ületada praegust seisu, mida iseloomustab olemasolevate keeleandmete ja ressursside killustatus ja koordineerimatus, pakkudes välja riikide ja valdkondade ülesed veebiteenused. CLARIN tugineb erinevate riikide rahvuslikele infrastruktuuridele/keskustele mis suudavad hallata rahvuslikke keeleressursse ja tehnoloogilisi lahendusi.

CLARINi kaudu töötatakse välja ühised standardid, kvaliteedinõuded jms ning lepitakse kokku ressurssidele juurdepääsuõiguste põhimõtetes. Seega muutuvad CLARINi keskuste kaudu kõik partnerkeskustes olevad erinevate keelte ressursid Euroopa teadlastele kergesti kättesaadavaks. Selleks, et eesti teadlased saaksid osa sellest üleeuroopalisest keeleressursside ja tehnoloogiate rikkusest ning selleks, et lisada sellesse võrgustikku ka eesti keele ressursid luuakse Eesti keeleressursside keskus.

Programmi alaeesmärgid ja oodatavad tulemused

Programm on jaotunud 5 alaeesmärgi vahel.

1. Tarkvaraprototüüpe loovad uurimis- ja arendusprojektid

Uurimis- ja arendusprojektid tarkvaraprototüüpide loomiseks järgnevates valdkondades (loetelu ei ole lõplik):

- kõnetuvastus:
    - keele- ja rakenduse-spetsiifiliste tuvastusmoodulite (akustilised mudelid, keelemudelid) uurimine ja arendus
    - rakendused eri valdkondades (spontaanse ja dialoogkõne, raadio- ja telesaadete automaatne transkribeerimine, piiratud valdkonna dialoogsüsteemid, kõnetuvastus piiratud sagedusribaga sidekanalites)
    - kõneldava keele automaatne identifitseerimine
- kõnesüntees:
    - kõnesünteesi liidesed ja rakendused eri valdkondades (digiraamatute genereerimine, audiosüsteemid nägemispuudega inimestele, subtiitrite helindamine digitelevisioonivõrgus, liidesed tekst-kõne sünteesi kasutamiseks erinevates süsteemides ja rakendusprogrammides, suvalistest kõnekorpustest uute sünteeshäälte automaatne genereerimine)
    - kõnesünteesi prosoodiamudelite seostamine teiste keeletasanditega (süntaks, semantika, pragmaatika) ja keeleväliste teguritega (emotsioonid)
    - audiovisuaalse kõnesünteesi (nn rääkiv pea) mudelid
- süntaktiline/semantiline/pragmaatiline analüüs-süntees:
    - suulise keele eripära arvestav süntaksianalüsaator
    - süntaktiline süntees semantilise esituse põhjal
    - sidustekstide süntaktiline analüüs-süntees
    - liitlausete ja sidustekstide semantiline analüüs (teatud valdkondades) ja vahendid semantiliseks sünteesiks
    - pragmaatiline analüüs-süntees (teatud valdkondades) ja seosed teiste keeletasanditega (süntaks, semantika)

- seotud teksti (sh dialoogi) analüüs-süntees kõnes ja kirjas:
    - seotud teksti struktuuri automaatne tuvastamine/süntees, koherentsuse, kategoriseerimise vahendid tekstis, dialoogi struktuur (eraldi suuline ja kirjalik, nt Interneti-dialoog)
    - dialoogiaktide automaatne tuvastamine/süntees
    - dialoogistrateegiate automaatne tuvastamine/süntees
    - dialoogsüsteemide ja kasutajaliideste prototüübid (teatud valdkondades)
    - suulise keele (pool)automaatse transkribeerimise süsteem
    - keeleväliste suhtlussignaalide (häälekvaliteet, emotsioonid, naer, jms) automaatne tuvastamine
- tekstitöötluse abivahendid:
    - tekstiliigi automaatne tuvastaja
    - tõlkija töökeskkond
    - terminoloogi töökeskkond
    - leksikograafi töökeskkonna modifitseerimise moodulid
    - eestikeelsete veebidokumentide indekseerimise süsteem, mis võimaldab autorsuse määramist ja plagiaadi väljaselgitamist
- masintõlge:
    - eesti keele töötlemise vahendite rakendamine masintõlkesüsteemis

Projekte taotletakse avaliku konkursi korras, projekti maksimaalne kestus kuni 4 aastat; konkurss toimub igal aastal.

2. Keeleressursse loovad projektid

Projektid korduvkasutatavate keeleressursside loomiseks ja arenduseks järgnevates valdkondades (loetelu ei ole lõplik):

- tekstikorpused:
    - internetikeele korpus
    - õppijakeele korpus
    - ilukirjanduse tasakaalustatud korpus
    - eesti teaduskeele korpus

- süntaksi/semantika ressursid:
    - süntaktiliste puude pank
    - semantiline andmebaas
    - freimileksikon
    - valdkonnaspetsiifilised ontoloogilised andmebaasid
- leksikograafia ressursid:
    - leksikaalgrammatiline andmebaas
    - sõnastike andmebaasid
- masintõlke ressursid:
    - paralleelkorpused statistilise masintõlke vajadusteks (Euroopa Liidu dokumendid, Eesti teadlaste tähtsaimad välissuhtluse valdkonnad)
- kõnekeele ressursid:
    - eri liiki loomuliku kõne korpused (spontaanse, emotsionaalse, võõrkeelse aktsendiga kõne, valdkonna- ja tekstiliigispetsiifilise kõne korpused)
    - dialoogikorpused
- multimodaalsed ressursid:
    - audio-video korpused
    - viipekeele korpus
    - kõneproduktsiooni (glotograafia, palatograafia) andmebaasid
- elektroonsete sõnastike ja ontoloogiliste andmebaaside arendus, standardiseerimine ja avaliku kasutuse võimaldamine
- korpuste märgendus-, otsingu- ja haldussüsteemide arendus

Projekte taotletakse avaliku konkursi korras, projekti maksimaalne kestus kuni 4 aastat; konkurss toimub igal aastal.

3. Eesti Keeleressursside Keskus

Luua keskne depositoorium keeleressursside ja -tarkvara haldamiseks – Eesti Keeleressursside Keskus (EKRK)

Eesti Keeleressursside Keskus on partnerite konsortsiumlepingu alusel loodav infrastruktuur, mille abil tehakse eestikeelsed keeleressursid huvilistele kättesaadavaks. Lisaks olemasolevatele ja EKKTT 2006-2010 tulemusel saadud keeleressursside kogumisele ja arhiveerimisele käivitab keskus süsteemi kogutud keeleressursside tutvustamiseks ja potentsiaalsete kasutajate koolitamiseks.

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. Keskuse tegevusvaldkond oleks alates keeletehnoloogiliste standardite väljatöötamisest/fikseerimisest kuni keeleressursside kasutamiseks vajalike juriidiliste lepingute/litsentside koostamiseni.

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.

Käesolevast programmist võidakse finantseerida mh EKRK tegevusi nagu standardite fikseerimine, litsentsilepingute väljatöötamine ja sõlmimine, juurdepääsuõiguste kontroll, arhiveeritavate ressursside kvaliteedi kontroll ja dokumenteerimine, kasutajate keskkonna loomine, PR ja koolitused jms.

Projekte taotletakse vastavalt keskuse tegevusplaanile.

4. Integreeritud keeletarkvara ja selle rakendused

Integreeritud keeletarkvara ja selle rakendused:

- inimene-masin-dialoogisüsteemid piiratud valdkondades, nt teabetelefonides
- abivahendid erivajadustega inimestele
- arvuti- ja/või internetipõhine keeleõpe
- avalike teenuste kasutajaliidesed

Projekte taotletakse avaliku konkursi korras, projekti maksimaalne kestus kuni 2 aastat. Konkurss ei toimu igal aastal, konkursi väljakuulutamise aja otsustab programmi juhtkomitee. Projektitaotluse eelduseks on avaliku või erasektori partneri(te) olemasolu, vajalik on partneri(te) reaalselt mõõdetav panus (rahaline või intellektuaalne).

5. Tellitavad arendusprojektid

Juhtkomitee või avaliku sektori asutuse ettepanekul tellitavad arendusprojektid.

Projekti ülesande ja tehnilised tingimused koostab projekti tellija; projekti taotletakse avaliku konkursi korras, projekti maksimaalne kestus kuni 2 aastat; konkurss ei toimu igal aastal, konkursi väljakuulutamise aja otsustab programmi juhtkomitee.

Arendustegevused

Arendustegevused toimuvad kõigi alaeesmärkide, kuid eelkõige 4. ja 5. alaeesmärgi raames. Arendusprojektide avaliku sektori peamise partnerina nähakse Majandus- ja Kommunikatsiooniministeeriumi, aga ka teisi ministeeriume ja riigiasutusi. Programmi raames loodavate ressursside ja tarkvara prototüüpide kättesaadavaks tegemise kaudu stimuleeritakse äriettevõtteid kasutama keeletehnoloogia tarkvara ja seeläbi arendama uusi innovatiivseid e-teenuseid nii avaliku sektori huvides kui ka ärirakendustes.

Loodavate ressursside ja tarkvara kasutamise litsentseerimine

Põhimõtted:

- programmi raames loodud keeleressursid ja tarkvaraprototüübid on intellektuaalne omand,
    - mille kaitsmise viisi valikul lähtutakse eesmärgist soodustada loodud omandi kasutuselevõttu, eelistatakse avatud juurdepääsuga (Open Access) litsentseerimist, lähtutakse mittekasumitaotlusest ning välditakse eksklusiivsete varaliste õiguste võõrandamist;
    - mille kasutamist erinevatel eesmärkidel (avalik kasutus, teadustöö, ärirakendus) reguleerivad eri tüüpi litsentsid, mis võivad olla nii tasuta kui tasulised ja sõltuda ka teistest õiguslikest regulatsioonidest nagu näiteks isikuandmete kaitse.
- loodud ressursse ja tarkvara haldab, teeb kättesaadavaks ning litsentsidega tegeleb Eesti Keeleressursside Keskus.

Kõik programmi projektide tulemused deponeeritakse projekti lõppversioonina (koos dokumentatsiooniga) EKRK juures ning sõlmitakse projekti tulemuste kasutamise raamleping. Väga erandlikel juhtudel, kui projekti tulemust ei ole võimalik EKRK juures deponeerida, tagab projekti täitja projekti tulemuse lõppversiooni säilimise ja kasutatavuse enda serverites ja andmekandjatel.

Programmi juhtimine

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

Programmi juhtkomitee

Juhtkomitee ülesandeks on analüüsida keeletehnoloogia arengut Eestis ning kogu maailmas. Juhtkomitee töötab välja programmi taotlusvormid ning otsustab igal aastal eraldi, milliste alaeesmärkide taotlusvoorud avada. Juhtkomitee jaotab käesoleva programmi alaeesmärkidele eraldatud vahendid konkursi korras või sihtotstarbeliselt Eesti Keeleressursside Keskuse tööks tegevuskava alusel, kusjuures alaeesmärkide vahelised proportsioonid kinnitatakse igal aastal eraldi, lähtudes laekunud projektitaotluste hulgast ja kvaliteedist.

Juhtkomitee vastutab programmi eesmärkide täitmise eest ja jälgib, et programmi vahendeid kasutataks sihipäraselt. Selleks kasutab juhtkomitee laiapõhjaliselt ekspertide abi. Programmi juhtkomitee moodustab konkreetsete ülesannete lahendamiseks töörühmasid.

Programmi haldamine ja koordineerimine

Programmi haldamise delegeerib juhtkomitee esimesel võimalusel Eesti Keeleressursside Keskusele, kes leiab konkursi korras programmi koordineerija. Haridus- ja Teadusministeerium sõlmib programmi haldamiseks lepingu Eesti Keeleressursside Keskust kureeriva asutusega, kes omakorda sõlmib töölepingu programmi koordineerijaga. Programmi koordineerija ja programmi haldamise kulud kaetakse programmi vahenditest.

Programmi koordineerija võtab osa juhtkomitee koosolekutest, kuid ei oma hääleõigust. Programmi koordineerija vastutab programmi konkursi korraldamise, vahendite sihipärase kasutamise ja aruandluse läbiviimise eest. Programmi koordineerija valmistab ette lepingud programmi täitjatega. Programmi koordineerijal on õigus teha juhtkomiteele ettepanekuid abitööjõu palkamiseks eelarve piires. Koordineerija esitab programmi täitmise sisulised aruanded kinnitamiseks juhtkomiteele ning rahalised aruanded Haridus- ja Teadusministeeriumile.

Teavitustegevus ja avalikud suhted

Programmi teavitustegevuse ja avalike suhetega tegeleb programmi koordineerija.

Programmi rakendamine

Programm on kavandatud seitsme aasta pikkusena. Esimestel aastatel on prioriteetsed alaeesmärgid 1 ja 2, viimasel 3 aastal kasvab alaeesmärkide 4 ja 5 olulisus.

Programmi rahastamisvajadus

Riikliku programmi rahastamisvajadus eurodes.

	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**Kogu-maksumus**
1. Tarkvara-prototüübid	443866	499150	499150	499150	499150	499150	499150	**3438766**
2. Keeleressursid	229635	226567	226567	226567	226567	226567	226567	**1589037**
3. Eesti Keeleressursside Keskus	76502	76502	99702	108650	117597	127823	134214	**740990**
4. Integreeritud keeletarkvara	0	72731	69664	110886	161697	209950	225768	**850696**
5. Tellitavad arendusprojektid	0	14380	68705	105454	161697	209950	225768	**785954**
Programmi haldamine	15339	18215	25245	27802	35151	38027	40264	**200043**
Kokku	**765342**	**907545**	**989033**	**1078509**	**1201859**	**1311467**	**1351731**	**7605486**

Tabelis toodud summad on indikatiivsed, tegelik rahastamine sõltub riigieelarvest. Tabeli koostamisel on arvestatud dünaamikat, et esimestel aastatel on prioriteetsed alaeesmärgid 1 ja 2, viimasel 3 aastal kasvab alaeesmärkide 4 ja 5 olulisus.

Programmi rakendamisel soovitud olukord ning tulemuslikkuse seisukohalt kriitilised tegurid

Programmi eesmärgiks on tagada olukord, kus eesti keele keeletehnoloogiline tugi on võrdsel tasemel arenenud keeletehnoloogiaga riikide (nt Põhjamaad) keeltega suundades, mida nõuavad eesti keelele orienteeritud tarkvara arendused ja rakendused. Ka enamarendatud keeletehnoloogilise toega keelte puhul on tõdetud, et paljud KT ülesanded ei ole nii lihtsalt lahendatavad kui eelnevalt prognoositi, eriti masintõlke ja suulise suhtluse dialoogsüsteemide loomine on osutunud oodatust tunduvalt keerukamaks ülesandeks. Seetõttu on programmi eesmärkide saavutamise tõenäosuseks 2017. aastal hinnatud 70-80%.

Programmi mõjud:

- keelematerjali adekvaatselt töötlevad programmid on ühiskonnas laialdaselt kasutusel; võimalus kasutada loomulikku kõnesünteesi, piiratud valdkondadele loodud kõnetuvastust, üldlevinud tekstitoimetaja abivahendeid, inimene-masin-dialoogisüsteeme ja masintõlke prototüüpe;
- eesti keelt tunnustatakse arenenud keeletehnoloogiaga keelena, Eesti kuulub enim arenenud keeletehnoloogiaga riikide hulka.

Kriitilised tegurid

Käesoleva riikliku programmi edukas täitmine sõltub programmi juhtimisest, osapoolte (nii põhitäitjad kui ka ettevõtlussektor) vastutustundlikust tööst, spetsialistide ettevalmistusest ja olemasolust ning 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 erasektorisse ning programmi eesmärgid jäävad saavutamata. Programmi rahastamine tuleb viia vastavusse erasektori võrreldavate töökohtade palkade üldise tõusuga.

Programmi eesmärke ei ole võimalik saavutada ilma Eesti keeletehnoloogia infrastruktuuri ajakohastamiseta, st ilma RP EKKTT täitjate riist- ja tarkvara tänapäevastamise ning ühtlustamiseta ja ilma ühtse infotehnoloogilise platvormi ja ühtsete standarditeta. Selleks aitab oluliselt kaasa Eesti keeleressursside keskuse loomine. 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.

Programmile on teataval määral riskiks olemasolevad eesti keelele rakendatud suletud lähtekoodiga või patenteeritud IT-lahendused, mis pole taskukohased kogu avalikule kasutajaskonnale, kuid on samal ajal argument keeldumaks rahastada dubleerivaid keeletehnoloogilisi lahendusi. Programmi riskide maandamiseks on oluline arendada tänapäevast litsentsipoliitikat ning tagada eesti keele arengu jaoks võtmetähtsusega keeletehnoloogiliste lahenduste vabavaralisus.

Riskiks on ka võimalikud keelehoiakute muutused, 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 fetišeerima ning programmilt oodatakse rohkem kui see pakkuda suudab, jättes unarusse igapäevase keelekorralduse ja keelehoolde.

Indrek Reimand
teadusosakonna juhataja