# EKT 2014

Riikliku programmi

Eesti keeletehnoloogia

(2011–2017)

teine konverents


EESTI KEELETEHNOLOOGIA

NATIONAL PROGRAMME FOR ESTONIAN LANGUAGE TECHNOLOGY

**23.–24. aprill 2014**
**IT Kolledž ja EKI, Tallinn**

# RIIKLIK PROGRAMM „EESTI KEELETEHNOLOOGIA 2011-2017“

Riikliku programmi „Eesti keeletehnoloogia 2011-2017” (EKT) peaeesmärgiks on kooskõlas „Eesti keele arengukavaga aastateks 2011–2017“ saavutada Eestis keeletehnoloogiline tase, mis võimaldab eesti keelel edukalt toimida tänapäeva infotehnoloogilises maailmas.

Programm rahastab keeletehnoloogia alast teadus- ja arendustegevust alates ressursside loomisest kuni keeletehnoloogiliste rakenduste prototüüpide loomiseni.

Eesti keeletehnoloogia jätkusuutliku taseme saavutamiseks rahastatakse programmi kaudu projekte ja suunatud tegevusi viies alaeesmärgis:

1. Tarkvaraprototüüpe loovad uurimus- ja arendusprojektid;
2. Keeleressursse loovad projektid;
3. Eesti Keeleressursside Keskus;
4. Integreeritud keeletarkvara ja selle rakendused;
5. Tellitavad arendusprojektid

Programm eristub eelnenud riiklikust programmist „Eesti keele keeletehnoloogiline tugi (2006-2010)“ (EKKTT) selle poolest, et lisaks tarkvaraprototüüpide ja keeleressursside arendamisele pööratakse suurt tähelepanu keeletehnoloogia rakenduste loomisele ja olemasolevate ning loodavate ressursside ning tarkvara kättesaadavaks tegemisele.

Programmi raames loodud keeleressursid ja tarkvaraprototüübid on intellektuaalne omand, mille kasutamist erinevatel eesmärkidel (avalik kasutus, teadustöö, ärirakendus) reguleerivad eri tüüpi litsentsid. Loodud ressursse ja tarkvara hakkab haldama, kättesaadavaks tegema ning litsentsidega tegelema Eesti Keeleressursside Keskus (EKRK).

**EKT juhtkomitee**

| | | |
|---|---|---|
| **Esimees** | Jaak Vilo | Arvutiteaduse instituut, Tartu Ülikool |
| **Liikmed** | Andero Adamson | Haridus- ja teadusministeerium |
| | Tanel Alumäe | Tallinna Tehnikaülikooli Küberneetika Instituut |
| | Tiit Roosmaa | Eesti Infotehnoloogia Kolledž |
| | Hella Suvi | Haridus- ja teadusministeerium |
| | Arvi Tavast | Tübingeni Ülikool |
| | Kaarel Kaljurand | Zürichi Ülikool |
| | Uuno Vallner | Majandus- ja kommunikatsiooniministeerium |
| | Kadri Vider | Eesti Keeleressursside Keskus, Tartu Ülikool |
| | Jan Villemson | AS Cybernetica |
| | Tanel Tammet | ELIKO Tehnoloogia Arenduskeskus OÜ |

**Programmi koordinaator**

Rait Talvik – rait.talvik@ut.ee, taskutelefon 56 495 530

Programmi koduleht www.keeletehnoloogia.ee

# „EESTI KEELETEHNOLOOGIA 2011-2017“ PROJEKTID EESTI KEELETEHNOLOOGIA KONVERENTSIL 2014

| Alamprog. | Projekti nr. | Projekti nimi | Projektijuht | Asutus | Projekti kestus |
|---|---|---|---|---|---|
| Tarkvara | EKT1 | **Kõne ja teksti emotsionaalsuse statistilised mudelid** | Hille Pajupuu | Eesti Keele Instituut | 2011-2014 |
| Tarkvara | EKT5 | **Eestikeelse dialoogi pragmaatika analüsaator** | Mare Koit | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2013 |
| Tarkvara | EKT7 | **Vahendid teksti mitmekihiliseks märgendamiseks (rakendatuna Koondkorpusele)** | Kadri Muischnek | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2014 |
| Tarkvara | EKT12 | **Semantika vahendid eesti keelele** | Neeme Kahusk | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2014 |
| Tarkvara | EKT17 | **Audiovisuaalse kõnesünteesi prototüüp** | Einar Meister | Tallinna Tehnikaülikool, TTÜ Küberneetika Instituut | 2011-2014 |
| Tarkvara | EKT18 | **Kõnetuvastus** | Tanel Alumäe | Tallinna Tehnikaülikool, TTÜ Küberneetika Instituut | 2011-2014 |
| Tarkvara | EKT20 | **Kõnesünteesiliidesed** | Meelis Mihkla | Eesti Keele Instituut | 2011-2014 |
| Tarkvara | EKT22 | **Mallipõhine faktituletus tekstikorpustest** | Sven Laur | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2013 |
| Tarkvara | EKT39 | **E-keelenõu** | Arvi Tavast | Eesti Keele Instituut | 2013-2013 |
| Integr. tarkvara | EKT37 | **Subtiitrite helindamise ja tele-eetrisse edastamise tarkvaralahendus** | Meelis Mihkla | Eesti Keele Instituut | 2012-2013 |
| Integr. tarkvara | EKT38 | **Eestikeelsete dialoogsüsteemide loomise raamistik** | Margus Treumuth | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2012-2014 |
| Keele-ressursid | EKT2 | **Eesti Wordnet'i täiendamine** | Heili Orav | Tartu Ülikool, Filosoofiateaduskond | 2011-2014 |
| Keele-ressursid | EKT3 | **Kõne- ja multi-modaalsed korpused** | Einar Meister | Tallinna Tehnikaülikool, TTÜ Küberneetika Instituut | 2011-2014 |
| Keele-ressursid | EKT4 | **Eesti keele spontaanse kõne foneetilise korpuse arendused** | Pire Teras | Tartu Ülikool, Filosoofiateaduskond | 2011-2014 |
| Keele-ressursid | EKT6 | **Autentse meditsiinikeele korpuse alusel radioloogia elektroonse piltsõnastiku koostamine** | Eola Valdre | Tartu Ülikool, Filosoofiateaduskond | 2011-2014 |
| Keele-ressursid | EKT8 | **Suulise eesti keele audiovisuaalse suhtluskorpuse kogumine ja päringusüsteemi arendamine.** | Tiit Hennoste | Tartu Ülikool, Filosoofiateaduskond | 2011-2014 |
| Keele-ressursid | EKT11 | **Uued ressursid masintõlkes** | Heiki-Jaan Kaalep | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2013 |
| Keele-ressursid | EKT13 | **Võru ja seto keelekorpus** | Sulev Iva | Võru Instituut | 2011-2014 |
| Keele-ressursid | EKT35 | **Eesti avatud paralleelkorpus** | Margit Kurm | Tilde Eesti OÜ | 2012-2014 |
| EKRK | EKT10 | **Eesti Keeleressursside Keskus** | Kadri Vider | Tartu Ülikool, Matemaatika-informaatikateaduskond | 2011-2014 |

# KOGUMIKU SISUKORD

# KÕNE JA TEKSTI EMOTSIONAALSUSE STATISTILISED MUDELID

| | |
|---|---|
| **Organisatsioon** | Eesti Keele Instituut |
| **Vastutav täitja** | Hille Pajupuu |
| **Teised täitjad** | Rene Altrov, Jaan Pajupuu, Kairi Tamuri |
| **Projekti kestus** | 2011–2014 |
| **Finantseerimine** | 45 000 € (2011), 36 000 € (2012), 36 000 € (2013), 36 000 € (2014) |

## PROJEKTI EESMÄRK

Automaatselt ära tunda emotsioon kõnes ja kirjas. Realiseeritakse kahe prototüübina: 1) veebipõhine kirjaliku teksti polaarsuse määraja (emotsioonidetektor); 2) kõnelejakohane emotsionaalsuse tuvastaja.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Oleme loonud **kirjaliku teksti emotsionaalsuse leksikonipõhise ja statistilise tuvastaja**, mille analüüsiüksuseks on ortograafiline lõik. Üldhinnang terviktekstile kujuneb lõikude emotsionaalsuse ja nende pikkuse põhjal. Emotsioonidetektori versioon 0.9 vt http://peeter.eki.ee:5000/valence/. Leksikonipõhine tuvastaja kasutab ortograafilise lõigu emotsionaalsuse määramiseks 1019-sõnalist emotsioonileksikoni (413 positiivset, 606 negatiivset sõna) ja reegleid. Statistilise emotsioonituvastaja treenimiseks ja häälestamiseks oleme loonud valentsikorpuse, mis koosneb eri tüüpi tekstide 2500-st ortograafilisest lõigust, mille emotsionaalsuse (positiivne, negatiivne, vastuoluline, neutraalne) on määranud inimene, vt http://peeter.eki.ee:5000/valence/paragraphsquery/ .

Emotsioonidetektor on olemas Google Chrome'i ja MS Exceli laiendusena. Installeerimisjuhis vt http://peeter.eki.ee:5000/applications/list/.

Emotsioonituvastaja kasutusvaldkonnad lisaks rakendustele inimese ja masina suhtluses:

Võib hinnata oma kirjutatud teksti (meilide, ettekannete, muu loome) emotsionaalsuse võimalikku mõju lugejale; Võib hinnata veebiteksti, nt ajaleheartikli emotsionaalsust; Võib hinnata Excelis olevate tekstide emotsionaalsust.

**Kõnelejakohase emotsionaalsuse tuvastaja** statistiliste mudelite treenimiseks oleme Eesti emotsionaalse kõne korpust laiendanud valentsi- ja aktiivsuseinfoga. Oleme läbi viinud 28 testi korpuse lausete valentsi (negatiivne, positiivne, neutraalne) määramiseks (nii lugemise kui kuulamise põhjal) ning 14 testi lause aktiivsuse-passiivsuse määramiseks (kuulamistestid), vt http://peeter.eki.ee:5000/reports/list/.

Tuvastuses kasutame vabavaralist kõnetunnusteeraldajat openSMILE http://opensmile.sourceforge.net/ (Eyben, Wüllmer, & Schuller, 2010), selle aktiivse kuulaja loomisele orienteeritud SEMAINE-projekti emotsioonianalüüsimudeleid ja koodi oleme kohandanud eesti andmete töötlemiseks. Tuvastamisel kasutame SVM-klassifitseerijat, vt http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Treeningbaasina kasutame Eesti emotsionaalse kõne korpust, vt http://peeter.eki.ee:5000/ Korpust oleme täiendanud sama naishääle spontaanse kõne emotsioonidega (need avalikult kättesaadavad 2014. a lõpus).

Olemas on *off-line* pilootversioon, mis tuvastab ühe kõneleja lausete emotsionaalse kategooria, valentsi ning aktiivsuse. Parimad tuvastustäpsused hetkel: rõõm 46% (inimtestijal keskm 76%), viha 59% (inimtestijal 74%), kurbus 51% (inimtestijal 74%), valents (positiivne-negatiivne) 71% (inimtestijal 87%), aktiivsus (aktiivne-passiivne) 68% (inimtestijal 84%).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proc. ACM Multimedia (MM), ACM. (pp. 1459-1462). Florence, Italy.

# EESTIKEELSE DIALOOGI PRAGMAATIKA ANALÜSAATOR

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Mare Koit |
| **Teised täitjad** | Sven Aller, Liina Eskor, Olga Gerassimenko, Riina Kasterpalu, Krista Mihkels, Siiri Pärkson, Andriela Rääbis, Raul Sirel |
| **Projekti kestus** | 2011-2013 |
| **Finantseerimine** | 69 000 € (2011-2013) |

## PROJEKTI EESMÄRK

Koostada tarkvara järgmiste teksti pragmaatilise analüüsi osaülesannete lahendamiseks:

- teadmuse ekstraheerimine eestikeelsest tekstist,
- dialoogiaktide tuvastamine,
- dialoogi struktuuri analüüs,
- dialoogistrateegiate analüüs.

Lisaks sellele arendatakse ühte keeleressurssi – Eesti dialoogikorpust – tarkvara loomiseks vajalikus ulatuses.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

1. On loodud n-ö vaba sisendfaili analüsaator, mis ekstraheerib KKK rubriikidest küsimused ja vastused ning moodustab dialoogsüsteemide raamistikus (projekt EKT38) kasutatava formaadi kohaselt regulaaravaldised, kusjuures kasutajalt eeldatakse ainult tulemuse järeltöötlust.
2. On loodud tarkvara olemasolevast eestikeelsete (meditsiiniliste) tekstiandmete korpusest teadmuse kaevandamiseks ja andmete visualiseerimiseks.
3. Dialoogiaktide tuvastamine vastavalt TÜ dialoogiaktide tüpoloogiale (kokku 126 akti). Analüsaator tükeldab dialoogi teksti lausungiteks ja määrab igale lausungile kuni viis tõenäolisemat aktimärgendit, kasutades Naiivse Bayesi klassifikaatorit. Seejärel saab inimene-ekspert parandada vigu ja vajadusel lasta korrata automaatset märgendamist.
4. Dialoogi struktuuri analüüs tehakse eeldusel, et sisendtekstis on märgendatud dialoogiaktid. Dialoogi struktuursed osad (sh alamdialoogid) tuvastatakse reeglipõhiselt.
5. Dialoogistrateegiate analüüs on samuti reeglipõhine ja eeldatakse, et dialoogis on eelnevalt märgendatud dialoogiaktid. Strateegiad määratakse vastavalt K. Jokineni konstruktiivses dialoogimudelis esitatud loetelule (kokku 16 strateegiat).
6. Projekti tulemusel on täiendatud Eesti dialoogikorpust. Dialoogiaktidega märgendatud korpuse kogumaht seisuga 31. detsember 2013:
    a. suulised inimestevahelised dialoogid: 1292 dialoogi (kogutud projektis EKT8),
    b. võlur Ozi meetodil kogutud (simuleeritud) dialoogid: 96,
    c. inimese ja arvuti vahelised dialoogid (dialoogsüsteemide raamistiku abil loodud Kinoagent ja Hambahaldjas): 144.

DIALOOGIKORPUSE TÖÖPINK: http://www.dialoogid.ee/dialoogid/

DIALOOGI PRAGMAATIKA ANALÜSAATOR: http://ats.cs.ut.ee/diapragma/

# VAHENDID TEKSTI MITMEKIHILISEKS MÄRGENDAMISEKS (RAKENDATUNA KOONDKORPUSELE)

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Kadri Muischnek |
| **Teised täitjad** | Kaili Müürisep, Tiina Puolakainen, Riin Kirt, Eleri Aedmaa, Dage Särg, Tarmo Vaino, Raigo Kodasmaa, Katrin Tsepelina, OÜ Filosoft alltöövõtjana |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 86 500 € (2013), 86 500 € (2014) |

## PROJEKTI EESMÄRK

Koondada eelnevalt korpuste märgendamiseks kasutatud tarkvaraprototüübid ühtseks standardiseeritud programmide koguks ning nende abil muuta eesti keele Koondkorpus mitmetasandiliselt (morfoloogiliselt, süntaktiliselt, semantiliselt) märgendatud korpuseks. Täiustada Koondkorpuse kasutajaliideseid ning esitada valmis kujul korpuse leksikaalsete ja grammatiliste kategooriate statistilise analüüsi tulemused.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

### Morfoloogiline ja süntaktiline analüüs

Põhiliselt jätkus töö kitsenduste grammatika (Constraint Grammar, CG ) morfoloogilise ühestaja ja süntaktilise (pind- ja sõltuvussüntaks) analüsaatoriga. Integreeriti kitsenduste grammatikal põhinevad morfoloogiline ühestaja ja süntaksianalüsaator. Viimane kohandati ka statistilise morfoloogilise ühestaja väljundile, st töötab eesti keele mõlema morfoloogilise ühestaja väljundiga. CG morfoloogiline ühestaja on adapteeritud EKI morfanalüsaatori väljundile, st töötab eesti keele mõlema morfoloogiaanalüsaatori väljundiga.

### Morfoloogiline ühestamine ja pindsüntaks

CG morfoloogilise ühestamise saagis on 2013. aasta lõpu seisuga 97.1%, täpsus 90.4%, vigu on 2.9%, mitmesus 6.2%. Võrdluseks: morfoloogiliselt analüüsitud, kuid ühestamata testkorpuse morfoloogilise kirjelduse saagis on 99.5%, täpsus 35.5%, vigu on 0.5%, mitmesus 57.7%. Pindsüntaktilise kirjelduse saagis testkorpusel on 92.9%, täpsus 69.3%, vigu on 7.1%.

### Sõltuvussüntaks

Kitsenduste grammatika sõltuvussüntaktilise analüüsi tulemuse saagis on 78.5%. On tehtud esialgsed katsed statistilise sõltuvussüntaktilise parseriga Maltparser (www.MaltParser.org), treeniti 130 000-sõnalisel ja testiti 33 000-sõnalisel käsitsi märgendatud korpusel, 2013. lõpu parseriversioon tuvastab 91.8% sõnadest korrektse märgendiga, 86,2% korrektse sõltuvusseosega ning 83.2% sõnadel on mõlemad korrektsed.

### Praktiline semantiline analüüs

Põhitegevus toimus 2011-2012. Loodi programm nime- ja numbriüksuste märgendamiseks tekstis ning sellega märgendati Keeleveebi (www.keeleveeb.ee) kaudu kasutatav Koondkorpuse versioon.

**Korpuste statistilise analüüsi tulemused**

Tasakaalus korpuse põhjal on koostatud sõnavormide ja lemmade sagedusloendid (http://www.cl.ut.ee/ressursid/sagedused1/), sõnaliikide ning käändsõna grammatiliste kategooriate sagedusloendid (http://www.cl.ut.ee/ressursid/gram-kat/), kollokatsioonide sagedusloendid sõnaliigikombinatsioonide kaupa (http://www.cl.ut.ee/ressursid/sagedased_kollokatsioonid/) ning lemmade ja sõnavormide mitmikute (n-grammide) sagedusloendid (http://www.cl.ut.ee/ressursid/mitmikud/).

**Koondkorpuse ja tema kasutusvõimaluste edasiarendamine**

Keeleveebi korpusepäringusse lisati metamärkide * ja ? kasutamise võimalus.

Loodi kollokatsioonide tuvastaja (https://korpused.keeleressursid.ee/clc/), mis võimaldab leida sõnaliigiliselt defineeritud lemma või sõnavormi kollokatsioone Koondkorpusest või selle allosadest.

Projekti raames on loodud järgmised **morfoloogiliselt ja süntaktiliselt märgendatud korpuseversioonid:**

Eesti keele puudepank – käsitsi sõltuvussüntaktiliselt märgendatud korpus, Koondkorpuse alamhulk, suurus 2013. aasta lõpul 343 600 sõna. Korpus pakituna koos kirjelduse ja märgendusjuhendiga on saadaval projekti kodulehel portaalis www.keeletehnoloogia.ee.

Kitsenduste grammatika morfoloogilise ühestajaga morfoloogiliselt analüüsitud Koondkorpus, (tervikuna saadaval projekti kodulehel portaalis www.keeletehnoloogia.ee), saab esitada päringuid Keeleveebis.

Pindsüntaktiliselt analüüsitud Tasakaalus korpusest on tehtud uus versioon (tervikuna saadaval projekti kodulehel portaalis www.keeletehnoloogia.ee), saab esitada päringuid Keeleveebis.

Tasakaalus korpusest (mis on samuti Koondkorpuse alamhulk) on tehtud esialgne kitsenduste grammatika (CG) analüsaatoriga sõltuvussüntaktiliselt analüüsitud versioon, mis on saadaval projekti kodulehel portaalis www.keeletehnoloogia.ee

# SEMANTIKA VAHENDID EESTI KEELELE

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Neeme Kahusk |
| **Teised täitjad** | Kadri Vare, Siiri Pärkson, Indrek Jentson, Sirli Parm, Sven Aller |
| **Projekti kestus** | 2011 - 2014 |
| **Finantseerimine** | 54 000 € (2011-2013) |

## PROJEKTI EESMÄRK

- Luua ja ühendada erinevaid semantikavahendeid eesti keele jaoks. Eelkõige keskendume sõnatähenduste ühestamisele ja freimisemantika jaoks vajalike vahendite arendamisele.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

- Eesti FrameNet
  - Freimileksikon 609 freimiga. Freimileksikon on XML kujul.
- Reeglipõhine ühestaja
  - Töötav programm, mis ühestab sõnatähendusi reeglite abil. Töötab Pythonis.
- Ühestamisreeglid
  - 90 reeglit sõnatähenduste ühestamiseks. Kasutatav reeglipõhises ühestajas.
- EKSS EuroWordNeti formaati teisendaja
  - EKSSi XML fail teisendub EuroWordNeti formaati, nii, et iga sõnatähendus saab omaette sünohulgaks.
- Pythoni moodul EuroWordNeti kasutamiseks.
  - Princetoni WordNeti moodul on integreeritud NLTK koosseisu, kuid EuroWordNeti struktuur ja failid erinevad sellest. Loodud eurown.py moodul võimaldab lugeda ja kirjutada EuroWordNeti faile, teha tehteid sünohulkadega, eksportida erinevale XML kujule, sh. ka LMF kujule.
- Projekti tulemused on saadaval aadressil http://www.keeletehnoloogia.ee/ekt-projektid/semantika-vahendid-eesti-keelele

# AUDIOVISUAALSE KÕNESÜNTEESI PROTOTÜÜP

| | |
|---|---|
| **Organisatsioon** | TTÜ Küberneetika Instituut |
| **Vastutav täitja** | Einar Meister |
| **Teised täitjad** | Rainer Metsvahi, Lya Meister |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 36 000 € (2011),29 000 € (2012), 37 000 € (2013), 37 000 € (2014) |

## PROJEKTI EESMÄRK

Projekti eesmärgiks on eestikeelse audiovisuaalse (AV) kõnesünteesi prototüübi loomine. Audiovisuaalse kõnesünteesi puhul lisatakse heliväljundile ka animeeritud inimnäo või pea kujutis. Projekti raames tegeldakse eelkõige eesti keelele omaste artikulatsioonimustrite loomisega parameetrilise peamudeli jaoks; peamudel liidestatakse Eesti Keele Instituudis loodud tekst-kõnesüntesaatori(te)ga.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

### Prototüübi loomine

Eestikeelse AV-sünteesi prototüübi loomisel on kasutatud parameetrilist peamudelit (MASSY mudel) ja eestikeelset difoonsüntesaatorit. Peamudeli juhtimisparameetrid (huulteava laius ja kõrgus, huulte torutatus, alahuule asend, keeletipu ja -keha asend) on leitud eesti viseemide artikulatsiooniandmestiku analüüsil.

### AV-sünteesi kvaliteedi hindamine

Sünteesiti VCV-ühendeid sisaldav AV-stiimulikorpus, milles konsonandid /k, p, t, l, m, n, r, s, š, f, v, j, h/ esinevad vokaalide /a, e, u/ kontekstis. Iga VCV-ühendi kohta on sünteesitud 5 helifaili, millest üks on ilma mürata ja neli erineva mürataustaga (signaal-müra suhe vastavalt +6dB, 0dB, -6dB ja -12dB), kokku 195 stiimulit. Tajukatsed viidi läbi kahes etapis: esmalt esitati katseisikule ainult audiostiimulid ja seejärel AV-stiimulid. Katse käigus kordus iga stiimul juhuslikus järjekorras kolm korda, katseisiku ülesandeks oli vastata, millist VCV-sõna ta kuulis. Testitulemused näitasid, et kui taustmürata AV- ja audiostiimuleid tajuti võrdselt hästi (tajuskoor 1.0), siis taustmüra olemasolul tuvastati AV-stiimuleid tunduvalt paremini kui audiostiimuleid (nt +6db müra korral oli AV-stiimulite tajuskoor 0.7, aga audiostiimulite tajuskoor 0.4).

### AV-sünteesi veebiliidese loomine

On loodud eestikeelse audiovisuaalse kõnesünteesi veebirakendus, vt http://massy-est.phon.ioc.ee/

Töötab praegu ainult Internet Explorer'is, eeldab Cortona 3D pleieri installeerimist.

### Uue peamudeli loomine

Loomisel on uus peamudel, selleks kasutatakse vabavaralist programmi Blender http://www.blender.org/. Realiseeritud on peamudeli staatiline osa ja erinevad tekstuurid, lõpetamisel on alalõua, huulte ja keele liigutamiseks vajalike komponentide mudeldamine.

# KÕNETUVASTUS

| | |
|---|---|
| **Organisatsioon** | TTÜ Küberneetika Instituut |
| **Vastutav täitja** | Tanel Alumäe |
| **Teised täitjad** | Kairit Sirts, Rena Nemoto |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 54 000 € (2011), 54 000 € (2012), 66 400 € (2013) |

## PROJEKTI EESMÄRK

Projekti eemärgiks on olemasoleva eestikeelse kõnetuvastustehnoloogia täiustamine, tehnoloogia kättesaadavakstegemine uute rakenduste loomiseks, juba olemasolevate rakenduste täiendamine ning uute rakenduste loomine.

Kõnetuvastustehnoloogiat täiustamisel pööratakse põhitähelepanu sellistele aspektidele, mille puhul on hetkel kvaliteet suhteliselt madal. Eesmärgid on:

- parem tuvastuskvaliteet madalama kvaliteediga kõnesalvestuste puhul (eelkõige läbi telefonikanali salvestatud kõne puhul);
- parem kvaliteet spontaanse kõne puhul;
- aktsendiga kõnelejate parem käsitlemine;
- nimega üksuste mainimiste parem tuvastus;
- kõne indekseerimine terminite ja nimega üksuste otsimiseks.

Pikkade kõnesalvestuste transkribeerimise osas on eesmärgiks on vähendada vigade arvu suhteliselt 25% võrra võrreldes 2010. a tasemega, ehk näiteks u 20%-ni raadio ja televisiooni vestlussaadete puhul.

Lisaks eelnevale on kavas tegeleda kõnetuvastuse väljundi struktureerimise meetoditega, mis võimaldaksid kõnetuvatuse väljundi “kirjavahemärgistamist” ning näiteks tele- ja raadiosalvestuste puhul ka ka kõneleja nime identifitseerimist.

Loodav tehnoloogia avaldatakse tasuta koos lähtekoodiga sellises vormis, mis võimaldab teda võimalikult lihtsalt integreerida kolmandate isikute loodavatesse rakendustesse.

Programmi raames on kavas luua uusi kõnetuvastusrakendusi. Uute rakenduste osas on plaanis tähelepanu pöörata järjest populaarsemaks saavate nutitelefonide rakendustele.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

### Mobiilirakendused

Valminud on mitmed tuvastustehnoloogiat kasutavad rakendused Android mobiiliplatvormile. Rakendus “Kõnele” lubab eestikeelse kõne abil sisestada teksti kõikides Androidi rakendustes. Rakendus “Arvutaja” võimaldab kõne abil teostada matemaatilisi tehteid, teha ühikuteisendusi, otsida Eesti kohanimesid. Rakenduse “Diktofon” abil saab salvestada pikki kõnelõike (näiteks intervjuud) ning neid automaatselt tekstiks teisendada.

**Kõnetuvastus veebibrauseris**

Valminud on veebirakendus "Dikteeri" (http://bark.phon.ioc.ee/dikteeri/), mille abil saab kõnetuvastuse abil dikteerida pikemaid ja lühemaid tekste otse veebibrauseris ilma igasuguse lisatarkvarata. Dikteerimine toimub reaalajaliselt.

Mitte-reaalajalist kuid kõrgekvaliteetset kõnetuvastust saab katsetada teise veebirakenduse abil (http://bark.phon.ioc.ee/webtrans/).

**Pikkade kõnesalvestuste transkribeerimise süsteem**

2013. a jooksul tehti suuri edusamme mitte-reaaljalise kõnetuvastuse vallas, mida kasutatakse põhiliselt pikkade kõnesalvestuste automaatseks transkribeerimiseks. Tänu närvivõrkudel põhinevate akustiliste mudelite kasutamisele, uute kõnekorpuste lisandumisele ja uute tekstikorpuste rakendamisele õnnestus tunduvalt parandada kõnetuvastuse kvaliteeti. Alljärgnevalt on toodud kõnetuvastuse sõnavigade osakaalu ('word error rate', WER) vähenemine aastate lõikes mitmes eri valdkonnas.

| Kõne tüüp | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|
| Raadio vestlussaated | 28,6 | 27,1 | 25,6 | 20,3 |
| Konverentsikõned | 37,1 | 33,9 | 33,0 | 26,4 |
| Raadio telefoniintervjuud | | 29,1 | 26,6 | 22,8 |

Pikkade kõnesalvestuste transkribeerimise süsteem on avaldatud koos lähekoodiga (http://github.com/alumae/kaldi-offline-transcriber). Selle on juba kasutusele võtnud kolm Eesti juhtivat meediamonitooringufirmat, kes kasutavad seda tele- ja raadiosaadete sisu analüüsiks.

**Lähtekood**

Kõik projekti raames implementeeritud rakendused on tasuta ja avaldatud koos lähtekoodiga. Lingid erinevatele komponentidele on leitavad projekti kodulehel.

# KÕNESÜNTEESILIIDESED

| | |
|---|---|
| **Organisatsioon** | Eesti Keele Instituut |
| **Vastutav täitja** | Meelis Mihkla |
| **Teised täitjad** | Indrek Hein, Indrek Kiissel, Elgar Kudritski, Liisi Piits |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 73 000 € (2011), 60 000 € (2012), 65 000 € (2013), 66 000 € (2014) |

## PROJEKTI EESMÄRK

- luua nutikaid liideseid (SAPI – Speech Application Programming Interface), mis võimaldaksid juhtida eestikeelset kõnesünteesi, jälgida tekst-kõne teisendusprotsessi, arvestada edastatava dokumendi struktuuri ja muuta sünteeshääle parameetreid (hääletugevus, kõnetempo, häälekõrgus) erinevates häälrakendustes;
- luua kõneliideseid erinevatele kõnesünteesi rakendustele: veebisõnastike helindamisliidesed, subtiitrite helindamine, heliraamatute genereerimine, nutirakendused jms

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

1. Salvestati uusi kõnekorpusi ja loodi uusi sünteeshääli (kuula ja võrdle http://heli.eki.ee/syntees):

- kõnekorpuse Eva põhjal valmisid nii HTS-meetodil eval_hts kui üksuste valikul põhinev hääl eval_clunits. Ainult sõnavormide põhjal loodi üksiksõnade ettelugemiseks sobiv evas_hts, mis ortograafilise teksti lugemise kõrval suudab arvestada ka tekstis esinevaid vältemärke.
- kõnekorpuse Luukas põhjal valmisid lapse sünteeshääl luukas_hts
- eestikeelne formantsüntees realiseeriti avatud koodiga kõnesünteesi arendussüsteemis eSpeak. Töö tulemusena valmis kaks sünteeshäält: formantsünteesil põhinev espeak-et ja difoonhääl espeak-mb-ee1

2. Eestikeelse kõnesünteesi kasutusvõimaluste laiendamiseks nutitelefonides

- arendati multiplatvormne HTS-sünteeshääl Androidi TTS API-ks, mille abil saab eestikeelset kõnesünteesi integreerida erinevatesse rakendusprogrammidesse. Vt http://heli.eki.ee/koduleht/index.php/rakendused
- loodi Androidi operatsioonisüsteemiga mobiiltelefonidele rakendus, mis loeb sünteeshäälega ette uudiseid. Rakendust on võimalik tasuta alla laadida leheküljelt http://heli.eki.ee/uudistelugeja

3. Loodi veebisõnastike helindamisliides, mida saab kasutada põhisõnavara sõnastiku märksõnade, muutevormide ja näitelausete helindamiseks (koostöös projektiga EKKM204). Kuula helindamisnäiteid http://www.eki.ee/dict/psh/

4. Valmis HTS-mootori ja eestikeelsete sünteeshäälte sobitusliides Sapi 5-le, mis võimaldab kasutada HTS-et sünteeshääli Windowsi platvormil.

5. Et eestikeelsed Markovi peitmudelitel põhinevad sünteeshääled oleksid kasutatavad lisaks Windowsile ka teistel platvormidel, teisendati HTS-sünteeshäälte moodulid C#-koodist C++-koodi. Optimeeriti label'ite struktuuri ja täiustati kõnemudelit. Iseseisva HTS-kõnesüntesaatori Linuxi versioon on kättesaadav kodulehel (http://heli.eki.ee/koduleht/index.php/konesuentees).

6. Loodi vaegnägijatele mõeldud veebileht, kust on võimalik alla laadida ja paigaldada sünteeshääli, mis ühilduvad SAPI-4 või SAPI-5-liidesega ja on sel moel kasutatavad erinevates Windowsi operatsioonisüsteemi rakendustes. Vt http://heli.eki.ee/vaegnagijale/

7. Adaptiivse sünteesi edasiarenduseks on läbi viidud hulk eksperimente erinevate sisendikombinatsioonide ja parameetrite valikuga. HTSi küsimustefaili on teisendatud eesti keelele omaste tunnuste ja reeglitega.

8. Valmis kõnesünteesikorpuste analüüsi ja täiendamise programm, mis võimaldab analüüsida korpuse struktuuri lause-, fraasi-, sõna-, silbi- ja foneemitasandil ning otsida baaskorpusest lauseid, mis sisaldaksid foneeme sobivas ümbruses. Programm on kättesadav http://heli.eki.ee/koduleht/index.php/korpused.

# MALLIPÕHINE FAKTITULETUS TEKSTIKORPUSTEST

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Sven Laur |
| **Teised täitjad** | Timo Petmanson, Aleksandr Tkatšenko, Fanny-Dhelia Pajuste, Jaak Vilo |
| **Projekti kestus** | 1.01.2011 - 31.12.2013 |
| **Finantseerimine** | 19 000 € (2012), 23 500 € (2013) |

## PROJEKTI EESMÄRK

- Faktituletusmeetodite loomine ja arendamine eestikeelsete vabatekstide analüüsimiseks.
- Tarkvarakomponentide loomine eestikeelsetest tekstidest struktuurse info eraldamiseks.
- Faktituletusmeetodite rakendamine praktikas faktibaaside loomiseks.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Lausetest eraldamiskõlbliku infomatsiooni võib tüübi järgi jagada kaheks peamiseks grupiks: üht suurust või omadust käsitlevad lihtfaktid ning mitut lauseosa omavahel ühendavad seosed. Eraldatavateks lihtfaktideks on näiteks patsiendi pikkus ning kaal. Tüüpiliseks binaarse seose näiteks on toimumine, mis ühendab endas nii sündmust kui ka asukohta. Lihtfaktide eraldamine on üldiselt oluliselt lihtsam, sest nende tuvastamiseks vajalik kontekst on kitsam ning masinõppe meetodite treenimiseks vajalikke näiteid on lihtsam leida ning märgendada. Seetõttu oleme praktikas tegelenud eelkõige diagnostiliste mõõtmisväärtuste (lihtfaktide) eraldamisega meditsiinitekstidest ja aidanud luua vastavaid suurusi sisaldavat meditsiinilist andmebaasi.

Projekti käigus loodud meetodeid kirjeldab kõige paremini Timo Petmansoni magistritöö [7], mis annab teoreetilise ülevaate näitepõhisestest mallikaeve meetoditest ja nende rakendamisest eestikeelsetele tekstidele. Töös kirjeldatakse otsitavate mustrite struktuuri, peamisi mõõdikuid mustrite headuse hindamiseks ning kombinatoorset otsialgoritmi sobivate mustrite leidmiseks. Lisaks seletatakse, kuidas kasutada mustrite sarnasust ning aktiivõpet uute näidete poolautomaatseks märgendamiseks.

Projekti ühe alameesmärgina uurisime isikute, organisatsioonide ja asukohtade tuvastamist tekstist. Tegemist on keeletöötluse standardülesandega, mida on teistes keeltes palju uuritud. Ka on sellist tuvastusalgoritmi vaja keerukamate faktituletusalgoritmide loomiseks. Töö tulemusena valmis kõrge saagise ja täpsusega NER-tööriist (F-skoor 87%) ning avalik käsitsi annoteeritud nimeolemite korpus [6]. Tööriista täpsem kirjeldus on avaldatud artiklis [1]. Kasutatud mustrite valikut on täpsemalt käsitletud raportites [10] ja [13]. Lisaks eelnevale loodi tööriista arenduse käigus vabavaraline lemmatiseerija [5].

Teise mallikaeve rakendusena uurisime statistiliselt oluliste süntaktiliste ja sõnavaramustrite kasutamist autorituvastuses. Tulemusi kirjeldav artikkel on avaldamisel ERÜ aastaraamatus [2]. Sama meetodit kasutasime ka autorituvastuse ülesande lahendamisel inglise-, hispaania- ning kreekakeelsetel korpustel [14].

Loodud tarkvara oleme praktikas rakendanud Elioni kliendikaebuste ning E-tervise meditsiinitekstide analüüsimisel. Mõlemal juhul on tagasiside olnud postiivne. E-tervise meditsiiniandmete analüüsimisel kasutatud anonümiseerija on loodud meie NER-tööriista

kohandades. Ühe suure praktilise eesmärgina oleme tegelenud erinevate numbriliste ja kategooriliste tunnuste nagu vererõhk, kaal, pikkus, kolesterool, suitsetamine eraldamiseks mõeldud meetodite arendamisega. Töö käigus valminud keeletehnoloogilised vahendid on kättesaadavad veebilehelt http://ats.cs.ut.ee/keeletehnoloogia/.

Töö käigus ilmnes, et kõige suurem takistus lihtfaktide eraldamisel on suure heterogeense näitebaasi loomine. Reeglina on otsitavate lihtfaktide esinemissagedus üsna madal ja seega on kõikide näidete leidmiseks kuluv töömaht väga suur. See raskendab oluliselt algoritmide saagise hindamist. Probleemi saab lahenda vaid poolautomatiseeritud märgendusmeetoditega, mis kasutavad ülisuuri märgendamata tekstikorpusi ebakõlade tuvastamiseks algoritmide väljundis.

Kogu projekti käigus loodud tarkvara on mõeldud avalikuks kasutamiseks ning on litsenseeritud GNU GPL v3 alusel (https://www.gnu.org/copyleft/gpl.html). Mallieraldustarkvara on saadaval GIT repositooriumis [3] ning Wiki lehtedel on näited teegi kasutamisest: lausestamine, morfoloogiline analüüs, mustrikaeve, nimeolemite eraldamine, Wikipedia korpuse kasutamine. Repositoorium sisaldab ka koodi autorituvastuse ülesande lahendamiseks. Teek on optimeeritud eesti keelele (kasutab t3mesta programmi morfoanalüüsiks), kuid võimalik on kasutada ka TreeTagger programmi ning töödelda ka teistes keeltes olevaid tekste. Tööd loomuliku keele töötlemisvahendite arendamisel keeles Python on plaanis jätkata.

**EKT22 projektiga seotud artiklid**


[1] A. Tkachenko, T. Petmanson, S. Laur. Named Entity Recognition in Estonian. Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8-9 August 2013. The Association for Computational Linguistics. ACL: Omnipress, 2013, (2013), 78 – 83.

[2] Petmanson, Timo (2014). Authorship verification in Estonian opinion pieces. Eesti Rakenduslingvistika Ühingu aastaraamat. Eesti Rakenduslingvistika Ühing. Ilmumas.


**Tarkvara ja keeleressursid**


[3] Vabavaraline liittunnuste eraldamise teek. https://code.google.com/p/patnlp/

[4] Eesti keele nimeolemite tuvastaja. http://patnlp.googlecode.com/files/estner.zip

[5] Vabavaraline eesti keele lemmatiseerija. https://github.com/brainscauseminds/suffixlemmatizer

[6] Käsitsi märgendatud NER treeningkorpus. https://metashare.ut.ee/repository/browse/estonian-ner-corpus/88d030c0acde11e2a6e4005056b40024f1def472ed254e77a8952e1003d9f81e/


**Projekti kodulehelt kättesaadavad raportid**


[ 7] T. Petmanson. Pattern-Based Fact Extraction from Estonian Free-Texts. Master thesis.

[8] T. Petmanson, S. Laur. EKT22 projekti esimese aasta (2011-2012) raport.

[9] F.-D. Pajuste, T. Petmanson, S. Laur. Korpuse eelklasterdamine.

[10] T. Petmanson. Reeglipõhiste tunnuste klastrid NER ülesande lahendamisel.

[11] T. Petmanson. Meditsiiniandmetest vererõhkude ekstrahheerimine.

[12] T. Petmanson. Mustripõhine kaebuste tuvastamine meditsiinilistest andmetest.

[13] T. Petmanson. Kasulikud tunnused nimeüksuste tuvastamiseks eesti keeles.

[14] T. Petmanson. Authorship identification using correlation of frequent features.

# E-KEELENÕU

| | |
|---|---|
| **Organisatsioon** | EKI |
| **Vastutav täitja** | Arvi Tavast |
| **Teised täitjad** | Elgar Kudritski, Kaur Männiko, Tõnis Nurk, Ülle Viks, Kati Sein, Kristian Kankainen |
| **Projekti kestus** | 2013 |
| **Finantseerimine** | 50 000 € (2013) |

## PROJEKTI EESMÄRK

Projekti sisu oli luua e-sõnastike tehnoloogia arendamist, keelenõuannet ja keelekorraldust toetav süsteem, millel on kolm peamist eesmärki:

1. Pakkuda lõpptarbijale intuitiivselt lihtsal viisil vastuseid keelealastele küsimustele (nt normingukohasus, vasted teistes keeltes, selgitused, etümoloogia, kasutusinfo jms).
2. Varustada keelekorraldajaid regulaarselt empiiriliste andmetega süsteemi kasutajate otsingukäitumise kohta.
3. Leida ja katsetada uusi tehnilisi lahendusi keeletehnoloogia viimiseks lõpptarbijale lähemale, esmajärjekorras e-sõnastike kasutajapoole funktsionaalsuse arendamine tõeliste e-sõnastike suunas (vastandina pabersõnastike veebiversioonidele).

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

E-keelenõu portaalil (kn.eki.ee) on kolme sorti väljundid: esimesena näeb kasutaja lihtsat päringuvälja, kus saab otsida eestikeelset sõna või fraasi ja saada otsingutulemusi sõnastikest, kohanimede andmebaasist, keelenõuvakast ja "Eesti keele käsiraamatust". Teine, vähem silmatorkavalt paigutatud päringutüüp lubab sisestada pikema teksti ja sellele rakendada olemasolevaid keeletehnoloogia vahendeid. Kolmas on tavakasutajasse mittepuutuv tagasiside keelekorraldajatele.

Võimalik on esitada keele- või termininõu päring, mis edastatakse vastamiseks EKI keele- või termininõule.

Viisime läbi süvaintervjuud 7 professionaalse keeletoimetajaga, kes olid vaimustunud, et samal ajal sõnastikega tehakse otsing ka keelenõuannete seas ja käsiraamatus ning et keeletööriistad on ühes kohas, koos keelenõu päringuga. Soovivad näha portaali edasiarendamist allikate (terminoloogia, võõrsõnade leksikon, korpus, keeletoimkonna otsused jne) ja keeletööriistade lisamise näol. Kogu aasta jooksul tehtud küsitlustega saadi ettepanekuid portaali käitumise ja välimuse parandamise kohta. Kasutajad esitasid mitu ettepanekut sõnastike (sh terminoloogiasõnastike) koostajatele (näit viide allikale, kui termin on võetud teosest või standardist; lisaks normile võiksid olla toodud ka laialt levinud ent lubamatu kasutus jm).

Päringustatistika analüüsi tulemusi kasutame peale toote enda täiustamise ka täiendava sisu lisamiseks (EKI keelenõuandjate kirjutatav õigekirjareegleid ja nende põhjendusi sisaldav õigekeelsuskäsiraamat, valmimine 2016-2018).

Projekti käigus õppisime oluliselt paremini tundma kasutajate ootusi ja olemasolevate ressursside tehnilisi võimalusi, mis nii mõneski kohas tingis algse plaani muutmise.

# SUBTIITRITE HELINDAMISE JA TELE-EETRISSE EDASTAMISE TARKVARALAHENDUS

| | |
|---|---|
| **Organisatsioon** | Eesti Keele Instituut |
| **Vastutav täitja** | Meelis Mihkla |
| **Teised täitjad** | Indrek Hein, Indrek Kiissel, Elgar Kudritski, Artur Räpp, Risto Sirts, Tanel Valdna |
| **Projekti kestus** | 2012-2013 |
| **Finantseerimine** | 28 560 € (2012), 29 000 € (2013) |

## PROJEKTI EESMÄRK

- televisioonis kasutatavate subtiitrifailide alusel kõnesüntesaatoriga helifailide genereerimine ning eraldi helikanalis digitelevisiooni eetrisse edastamine;
- toimetajaliidese loomine koos võõrnimede ja lühendite häälduSbaasiga;
- võimaluste loomine erivajadustega televisioonivaatajatele (vaegnägijad ja düsleksikud ning lapsed ja eakad inimesed) keelebarjääri ületamiseks, kui nad vaatavad televisioonis subtiitritega võõrkeelseid saateid.

## PROJEKTI TULEMUSED (2013. aasta lõpuks)

Eesti Keele Instituudi (EKI), Eesti Rahvusringhäälingu (ERR) ja Eesti Pimedate Liidu (EPL) ühisprojekti raames on loodud subtiitrite helindamise ja tele-eetrisse edastamise tarkvaralahendus (vt http://heli.eki.ee/koduleht/index.php/rakendused):

- subtiitrite toimetajale on helindamiseks loodud interaktiivne keskkond (toimetajaliides), milles valitud binaarse subtiitrifaili alusel genereeritakse kõnesüntesaatoriga ajakooditäpne helifail (EKI);
- toimetajaliides tagab subtiitrite redigeerimise vastavalt vajadusele ning toimetajale tuuakse teksti analüüsi põhjal esile tekstis leiduvad suurtähelised sõnad (võõrnimed, lühendid), kui neil puudub vaste häälduSbaasis (EKI);
- võõrnimede ja lühendite andmebaas http://heli.eki.ee/koduleht/programmid/mate/sonastik.php sisaldab ligi 50000 kirjet, üle 10000 võõrnimel on eestikeelne häälduSvaste (EKI);
- ERR on loonud subtiitri- ja helifailide haldamise, eetrisse planeerimise ning eetrisse edastamise tarkvara, millega muuhulgas miksitakse kokku programmi heli ja subtiitrite kõnefailid lisakanalis edastamiseks;
- EPL on testinud proovifilmide ja -saadete subtiitrite helindamist ja hinnanud erinevate sünteeshäälte meeldivust, tekstist arusaamist ning programmi- ja subtiitrite heli balanssi;
- alates 2013. aasta juunist on võimalus Eesti Televisiooni telekanalite (ETV ja ETV2) võõrkeelsete saadete ja filmide jälgimisel kuulata lisahelikanali kaudu helindatud subtiitreid.

# EESTIKEELSETE DIALOOGSÜSTEEMIDE LOOMISE RAAMISTIK

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool, Matemaatikainformaatikateaduskond |
| **Vastutav täitja** | Margus Treumuth |
| **Teised täitjad** | |
| **Projekti kestus** | 2012-2014 |
| **Finantseerimine** | 22 024 € (2012-2013) |

## PROJEKTI EESMÄRK

Projekt katab riikliku programmi alameesmärgi "seotud teksti (sh dialoogi) analüüs kõnes ja kirjas", sh täpsemalt "dialoogsüsteemide ja kasutajaliideste prototüübid teatud valdkondades".

Projekt on jätkuks senistele projektidele:

- Suhtlusstrateegiad suhtlusmudelis: eestikeelse dialoogi modelleerimine,
- Eestikeelne infodialoog arvutiga,
- Intelligentne kasutajaliides andmebaasidele.

Projekti käigus on varasemalt loodud eestikeelsete dialoogsüsteemide raamistik, mille abil saab luua dialoogsüsteeme kitsas ainevaldkonnas. Kasutades eelmainitud raamistikku, on loodud kaks veebipõhist dialoogsüsteemi, mis on avalikus kasutuses.

Raamistiku kasutamine dialoogsüsteemide loomiseks on seni olnud võimalik vaid raamistiku autori vahendusel.

Käesoleva projekti eesmärk on pakkuda avalik juurdepääs dialoogsüsteemide loomisele. Selleks luuakse raamistikule administreerimisliides, mille abil saab luua dialoogsüsteemi, seadistada selle parameetreid ning kohaldada teadmusbaasi kindla ainevaldkonna jaoks.

Projekti tulemusel tekib mugav võimalus kasutada raamistiku põhifunktsionaalsust, mis on ennast õigustanud mitmes ainevaldkonnas.

Kasutaja saab kiiresti seadistada omanäolise dialoogsüsteemi, kasutada inimabiliidest teadmusbaasi kogumiseks, kasutada teadmusbaasi täiendamiseks mugavat administreerimisliidest, vaadata vestluslogisid ja seadistada kujunduslikke elemente veebikihis. Valminud dialoogsüsteemi saab kasutaja integreerida oma asutuse veebilehega.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Aasta 2013 jooksul keskenduti projekti tulemuste kasutatavuse laiendamisele. Kui varem oli ADS raamistiku dialoogsüsteemi liideseks veebipõhine graafiline kasutajaliides, siis nüüdsest on loodud liides, mida saab kasutada käsurealt või oma rakendusest. Sellisel juhul on arendajal võimalus ise kujundada lõplik liides endale sobivaks.

Aastal 2013 integreeriti uuesti ka kõnesünteesiliides, sest EKI oli vahepeal oma liidest palju täiendanud ning vana liidese kasutusvõimaluse peatanud. Lülitusime EKI uue liidese kasutamisele. Selle tulemusel on ADS raamistiku dialoogsüsteemidel nüüdsest olemas ka naishäälne kõnesüntees ning üks ADS raamistikul loodud dialoogsüsteem naishäälse kõnesünteesiga on juba ka avalikus kasutuses.

# EESTI WORDNETI TÄIENDAMINE

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Heili Orav |
| **Teised täitjad** | Sirli Parm, Kadri Vare, Helen Türk, Eleri Aedma, Katrin Alekand, Ingmar Jaska, Lauri Eesmaa, Maria Reile, Piia Taremaa, Ahti Lohk |
| **Projekti kestus** | 1. jaanuar 2011 - 31. detsember 2014 |
| **Finantseerimine** | 159 200 € |

## PROJEKTI EESMÄRK

Üks eesti keeleressursse paljude teiste hulgas on Eesti Wordnet. Uuema põlvkonna mõistelises arvutisõnastikus on peale sõnade tähenduse eristamise fikseeritud ka tähendustevahelised seosed (sünonüümid, antonüümid, ülem- ja alammõisted, osa ja terviku suhted, põhjussuhted, osalussuhted jms). Mõistetele on lisatud ka nende ingliskeelsed vasted.

Eesti Wordneti projekti eesmärgiks on arvutitesauruse suurendamine (vähemalt 70 tuhande mõisteni), täiendamine ja olemasoleva kontrollimine ning parandamine. Projekt kestab neli aastat – 2011–2014.a.

## PROJEKTI TEGEVUSED

Töö Eesti Wordneti täiendamisel on toimunud järgmiselt:

- Läbi on töötatud erinevate valdkondade sõnavara (praegu töötatakse nt religiooni ja käsitöö valdkondade sõnavaraga).
- Tegime sõnaloendid neist sõnadest, mis on olemas eesti kirjakeele korpustes ja EKSS-is, kuid puuduvad Eesti Wordnetis ja mille järgi mõistete moodustamisega me juba oleme alustanud. Selliseid nimisõnu on loendis u 12500 ja tegusõnu u 5000. Nende seas on ka nt palju tuletisi (nt vähendava ehk deminutiivse tähendusega liidet -ke(ne) (nt tuleke(ne)) või tegevust, nähtust või omadust väljendav -us liide (nt võõrdumus)), mida saab vajadusel (pool)automaatselt tesaurusesse lisada.
- Jätkame uute omadus- ja määrsõnaliste sünohulkade lisamist koos täiendavate semantiliste suhetega
- Parandati ära mõistete ingliskeelsed vasted juhul, kui oli pandud sama equal_synonym mitmele mõistele korraga (u 1000 juhtu), kuid ingliskeelsete vastete kontrollimist tuleb jätkata. Hetkel on Eesti Wordnetist puudu umbes 45 000 ILI vastet, mida on kasutatud PWN-i 1.5 versioonis. See võib osutada kas puuduolevatele mõistetele või vigadele ingliskeelse vaste leidmisel.
- Jätkame olemasolevate mõistete vaheliste seoste korrigeerimist, nt saame kasutada Ahti Lohki visuaalsete graafide tulemusi, mis on välja töötatud wordneti hierarhilise struktuuri jaoks.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Projekti kahe aasta jooksul on lisandunud tesaurusesse u 24 800 uut mõistet – Eesti Wordnetis ver 69-s on (seisuga jaanuar 2014) üle 65 500 mõiste (projekti alguses 2011. a. oli u 40 700 mõistet). Sõnu, mis mõisteid moodustavad, on ligi 90 000, ingliskeelseid suhteid on üle 96 000 ja semantilisi suhteid üle 203 000, mis teeb keskmiselt 3 semantilist suhet mõiste kohta.

Eesti Wordnetti saab lehitseda läbi Keeleveebi või TÜ arvutilingvistika uurimisrühma kodulehelt http://www.cl.ut.ee/ressursid/teksaurus/.

# KÕNE- JA MULTIMODAALSED KORPUSED

| | |
|---|---|
| **Organisatsioon** | TTÜ Küberneetika Instituut |
| **Vastutav täitja** | Einar Meister |
| **Teised täitjad** | Lya Meister, Rainer Metsvahi, Martin Külvik |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 36 000 € (2011), 36 000 € (2012), 39 000 € (2013), 39 000 € (2014) |

## PROJEKTI EESMÄRK

Eestikeelsete kõnekorpuste salvestamine ja märgendamine kõnetuvastuse statistiliste mudelite treenimiseks ja kõne eksperimentaalfoneetiliseks uurimiseks.

Projekti tegevused:

1. olemasolevate kõnekorpuste mahu suurendamine ja salvestuste märgendamine,
2. uute korpuste kavandamine, salvestamine ja märgendus,
3. korpuste salvestusteks, töötluseks ja haldamiseks vajaliku infrastruktuuri arendus.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

**Loengukõne korpus**: eesmärk 30 tundi uusi märgendatud salvestusi.

2013 seis: on kogutud 45 tundi ja märgendatud 40 tundi konverentsiettekannete ja akadeemiliste loengute salvestusi.

**Aktsendikorpus**: eesmärk salvestada 40 uut eri keeletaustaga keelejuhti, märgendada 50 keelejuhi salvestused.

2013 seis: on salvestatud 28 uut eri keeletaustaga (19 läti, 5 rootsi, 3 jaapani, 1 saksa) keelejuhti, automaatselt on segmenteeritud loetud laused kogu korpuses, käsitsi on märgendati 100 keelejuhi spontaanset kõnet sisaldavad laused.

**Uudistekorpus**: eesmärk 20 tundi värskeid uudistesalvestusi.

2013 seis: on kogutud ja märgendatud 40 tundi eri raadiojaamade ja ETV uudiste salvestusi.

**Intervjuude korpus**: eesmärk 80 tundi märgendatud salvestusi.

2013 seis: on kogutud ja märgendatud 60 tundi eri raadiojaamade vestlussaadete ja intervjuude salvestusi.

**Noorukite kõnekorpus**: eesmärk salvestada 200 keelejuhti vanuses 8-16 ja märgendada kogu korpus.

2013 seis: on salvestatud 190 keelejuhti, automaatselt segmenteeritud kõigi keelejuhtide loetud laused, käsitsi märgendatud 40 keelejuhi spontaanset kõnet sisaldavad laused.

**Nimega üksuste korpus**: eesmärk salvestada ja märgendada kuni 50000 nimega üksust kuni 200 keelejuhiga

2013 seis: loobuti nimega üksuste häälduslvariatsioonide salvestamisest esialgselt kavandatud viisil (tavatelefoni, mobiili ja Skype kaudu), selle asemel märgendatakse nimega üksusi (isiku-, koha-, organisatsiooni- ja objektinimed) uudistekorpuses.

**Multimodaalsed korpused**: eesmärgiks on eestikeelse kõne artikulatsiooni kirjeldava andmebaasi salvestamine 2 keelejuhiga (maht ca 4 tundi) kasutades erinevaid mõõtesüsteeme – larüngograaf, palatograaf (EPG), EMA (elektro-magnetiline artikulograafia).

2013 seis: EPG-salvestused (VCV, CVCV üksused ja sagedasemaid konsonantklastreid sisaldavad sõnad) on tehtud kahe keelejuhiga, kokku ca 4 tundi; EMA-salvestused (VCV ja CVCV üksused) on tehtud ühe keelejuhiga.

# EESTI KEELE SPONTAANSE KÕNE FONEETILISE KORPUSE ARENDUSED

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikooli eesti ja üldkeeleteaduse instituut |
| **Vastutav täitja** | Pire Teras |
| **Teised täitjad** | põhitäitjad: Pärtel Lippus, Karl Pajusalu, Nele Salveste, Tuuli Tuisk, täitjad: Kätlin Aare, Anton Malmi, Ann Metslang, Margot Möller, Sander Pajusalu, Anette Ross, Helen Türk |
| **Projekti kestus** | 2011–2014 |
| **Finantseerimine** | 41 000 € (2011), 25 000 € (2012), 32 000 € (2013), 32 000 € (2014) |

## PROJEKTI EESMÄRK

Projekti eesmärgiks on arendada eesti keele spontaanse kõne foneetilist korpust. Korpust saab kasutada keeletarkvara väljatöötamiseks, kõnetuvastuse ja -sünteesi arendamiseks. Projekt on jätkuks riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006–2010)" projektile „Eesti keele spontaanse kõne foneetiline korpus", mille käigus loodud ressursid ei olnud veel piisavad ning vajasid arendamist.

Projekti üks eesmärke on olnud kasvatada korpuse salvestuste maht kuni 80 tunnini, mis tähendab salvestusi umbes 50 tunni ulatuses. Uusi lindistusi märgendatakse esmalt sõna- ja häälikutasandil. Lisaks käsitsi märgendamisele katsetatakse sõnatasandil poolautomaatset märgendamist, kasutades kõnetuvastuse abi. Jätkatakse ka nii varem tehtud kui uute lindistuste märgendamist muudel lingvistilistel kihtidel. Silbikihist alates arendatakse poolautomaatset märgendust skriptide abil, aga arendatakse ka märgendamise kontrollsüsteemi.

Arendatakse ka korpuse veebipõhist otsingumootorit (http://www.murre.ut.ee/otsing/ekskfk.php), mis võimaldaks teha korpusest keerulisemaid kombineeritud päringuid, aga automaatse morfoloogilise märgenduse järel saada infot ka spontaankõne morfoloogia kohta. Arendatav korpus on kõigile kättesaadav Internetis: http://www.keel.ut.ee/et/foneetikakorpus.

Projekti raames tehakse koostööd teiste kõnekeele korpustega, kõnetuvastuse ja kõnesünteesiga seotud projektidega.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

2013. aasta lõpu seisuga oli tehtud korpuse tarvis helisalvestusi 36 tundi ning korpuse kogumaht oli 63 tundi. Salvestuste tegemist on hõlbustanud 2012. a suvel soetatud salvestuskabiin. Lindistatud on peamiselt stuudios dialooge (19), aga ka monolooge (10, millest 8 on Einar Meistri salvestatud konverentsiettekanded, mis saadi Küberneetika Instituudist) ja vähemal määral dialooge välitöödel (7). Korpusesse on lisandunud projekti käigus 55 uut keelejuhti (24 naist ja 21 meest). Keelejuhte on korpuses kokku 90 (50 naist ja 40 meest). Keelejuhid on vanuses 21–85 aastat (keskmine vanus 39 aastat).

Segmentimisel ja märgendamisel on kahel viimasel aastal keskendutud sõna- ja häälikutasandile. Nendel tasanditel oli eelmise aasta lõpu seisuga märgendatud kokku 21 tundi kõnet (sõnatasandil 160 818 segmenti ja häälikutasandil 463 807 segmenti).

Muudel tasanditel (silbi-, takti-, lausungitasand), mille märgendamine oli keskmes eelkõige projekti esimesel aastal, on korpusesse lisandunud 461 295 segmenti.

Silbi- ja taktitasandi märgendamisel on kontrollitud ja ühtlustatud kahe esimese tasandi märgendamist. Nende tasandite eelmärgendamisel kasutatavaid skripte on täiustatud. Lisaks on arendatud märgendust kontrollivaid skripte, parandatud märgendamisel tehtud näpuvigu,

ühtlustatud varem märgendatut (nt häälikute lisakvaliteeti märkivate sümbolite järjekorda, liitsõnade märkimist jms).

Sõnatasandil on eelmärgendatud kõik seni tehtud lindistused, kasutades TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia labori automaatse kõnetuvastuse abi. Eelmärgendusele järgneb käsitsi märgendamine ning eelpool on toodud vaid see sõnade arv, mille puhul eelmärgendus on kontrollitud ja märgendatud on ka häälikutasand. Faile, kus sõnatasand on kontrollitud ja häälikutasand märgendatud, kasutatakse omakorda eestikeelse kõnetuvastuse arendamiseks.

Eesti keele spontaanse kõne foneetilise korpuse kogumaht oli 2013. a lõpu seisuga 1 944 380 segmenti (sõnatasandil 340 826 segmenti, häälikutasandil 993 373 segmenti ja muudel tasanditel 610 181 segmenti; kokku on märgendatud ligi 47 tundi kõnet). Hääliku- ja sõnatasandil on märgendatud 75% kogu korpusest, lisaks ka silbitasandil on märgendatud 41% ja kõigil lingvistilistel tasanditel 30% kogu korpusest.

Sõnatasandile märgitud lisainfot on hakatud üle kandma muudele tasanditele. Alustatud on häälelaadi tasandist, kus lisaks käsitsi märgendatud häälelaadile (eelkõige kärinale) on kasutanud automaatset kärinatuvastust (kasutatakse John Kane'i (Trinity College Dublin) häälekvaliteedi analüüsi tööriistu, vt https://github.com/jckane/Voice_Analysis_Toolkit).

Koostöös Heiki-Jaan Kaalepiga on kogu korpus morfoloogiliselt märgendatud, kasutades Filosofti morfanalüsaatorit. Tehnilistel põhjustel on märgendus ühestamata, sest ühestamiseks peaks sisend olema lause kaupa, aga spontaanse kõne korpuse märgendamisel pole sõnatasandil infot lausepiiride jms kohta. Tegu on niisiis täisautomaatse märgendusega. Ühestamatusest hoolimata on lisatud morfoloogilise info kiht ka veebiotsingusse ja arvame, et sellest võib siiski juba ka praegu olla kasu kasutajatel, kes tahavad saada infot spontaankõne morfoloogia kohta.

Projekti jooksul on koolitatud välja kuus uut märgendajat, kes kõik on osalenud korpuse töös helifailide segmentimisel ja märgendamisel ning kellest hetkel teeb neid töid viis.

Uurijaid, kellele on antud väljastpoolt foneetikalaborit ligipääs korpuse tervikfailidele, on kokku 8 (asutused: TTÜ Küberneetika Instituut, Eesti Keele Instituut, Göteborgi Ülikool, Helsingi Ülikool).

Projekti jooksul on valminud Pärtel Lippuse doktoritöö "The acoustic features and perception of the Estonian quantity system", milles on muuhulgas ka spontaankõne vältehäälduse analüüsi tulemused, *Kätlin Aare bakalaureusetöö* „Kärin eesti keele spontaanses kõnes" (juhendaja Pärtel Lippus), ilmunud on Pire Terase artikkel „Eesti diftongid spontaankõnes" (2012), Pärtel Lippuse, Eva Liina Asu, Tuuli Tuisu ja Pire Terase artikkel „Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian" (2013) jt.

# AUTENTSE MEDITSIINIKEELE KORPUSE ALUSEL RADIOLOOGIA ELEKTROONSE PILTSÕNASTIKU KOOSTAMINE

| | | | |
|---|---|---|---|
| **Organisatsioon** | TÜ Eesti ja üldkeeleteaduse instituut | | |
| **Vastutav täitja** | Eola Valdre | | |
| **Teised täitjad** | Peeter Ross, Katrin Tsepelina, Heiki-Jaan Kaalep, Kaarel Veskis, Tarmo Vaino | | |
| **Projekti kestus** | 2011—2014 | | |
| **Finantseerimine** | 22 000 € (2011) | 17 000 € (2012) | 23 320 € (2013) |

## PROJEKTI EESMÄRK

Projekti eesmärk on koostada radioloogiauuringute vastuste tekstidel, s.o tegelikul keelekasutusel, põhinev radioloogiasõnastik, mis hõlbustaks terviseandmete analüüsi, radioloogiaõpet, vabatekstipõhiste päringute tegemist, asjakohaste tõlkerakenduste loomist ja/või haiglainfosüsteemide kaasajastamist.

Eesmärgi saavutamiseks on töös kaks eri suunda: sõnastiku ja pildikogu koostamine. Radioloogiauuringute vastuste korpuse põhjal radioloogide töist kirjakeelt kajastava erialasõnastiku koostamiseks töötatakse välja metoodika erialaselt oluliste leksikaalsete ühendite korpusest leidmiseks ning määratakse nende esinemissagedus. Pildikogu jaoks valitakse ning kirjeldatakse radioloogilised kujutised, millel nähtavaid anatoomilisi struktuure ja patoloogilisi protsesse on võimalik siduda vastuste tekstide põhiste sõnastikukannetega. Projektil on Tallinna Meditsiiniuuringute Eetikakomitee luba nr 2169.

## PROJEKTI TULEMUSED

**2011. a** koostati AS Ida-Tallinna Keskhaiglas aastatel 2009—2011 tehtud radioloogiauuringute vastuste isikustamata vabatekstidest **meditsiinikeele korpus**. Selles on 207 534 radioloogiauuringute vastuse teksti (11,8 miljonit sõnet), millest röntgenuuringuid on 139 998 (4 663 958 sõnet), ultraheliuuringuid 34 020 (2 970 399 sõnet), kompuuteruuringuid 20 725 (2 751 990 sõnet), magnetuuringuid 11 037 (1 293 070 sõnet), stsintigraafiauuringuid 1754 (185 939 sõnet). Keskmiselt oli vastuses 57,2 sõnet. Keskmine sõnede arv oli kõige väiksem (33,1) röntgenuuringute, suurim (132,8) kompuuteruuringute puhul. Lühimad leiu kirjeldused koosnesid kõigest ühest sõnest (nt lileus, Normis, Normileid), pikimas oli 4882 sõnet. Sõneks loeti tühikutega eraldatud märgijada, kus oli vähemalt üks tähemärk või number.

**2012. a** koostati **radioloogilise normanatoomia sõnastik** ning seda kirjeldav **pildikogu**, TÜ arstiteaduskonna radioloogialoengute põhjal koostati **võrdlusmaterjali korpus** (65 719 sõnet) ning alustati korpusepõhist **lühendite analüüsi**. Projekti radioloog valis ja koondas pildikoguks radioloogilise normanatoomia suhtes representatiivsed radioloogiliste uuringute pildid ning kirjeldas neil olevad struktuurid. Pildikogu on arhiveeritud. Võrdlusmaterjali korpuse koostamiseks saime loa kasutada TÜ arstiteaduskonna III kursuse radioloogialoenguid ja VI kursuse kliinilise radioloogia loenguid (õppeained ARHO.01.033 ja ARHO.002.009). Selleks et esitada korpuses loengute tekst täielikult (sh kirjeldada ka slaidide illustratiivse materjali tekst (s.o radioloogilised kujutised, tabelid, skeemid, joonised jne) ning säilitada slaidi visuaalsed mõtteüksuse piirid (ei kattu enamasti lausepiiridega) ka tekstikorpuses, tuli kõik loengud slaidhaaval läbi vaadata (2455 slaidi) ning neil olev tekst käsitsi segmentida ning märgendada.

Sõnavara uurimist alustati lühendite analüüsist, sest lühendamine on radioloogiavastustes väga sage. Enamasti lühendatakse sageli kasutatavaid sõnu, pikki sõnu või sõnaühendeid:

radioloogiauuringute vastustes seega eelkõige termineid. Seetõttu on vastuste lühendite ja lühendamise uurimise kaudu võimalik kaardistada suur osa erialaspetsiifilisest terminikasutusest. Põhiprobleemid lühendite uurimisel olid: a) mis on lühend (nt mitmesõnalise termini puhul, mille iga sõna on lühendatud, kuid tekstis olev mõiste on nende tulem: teda kirjeldab vaid kõigist neist lühenditest koosnev lühend ja mitte selle üksikkomponendid: nt lühendite fr. vert. Th, corp. vert. Th, colum. vert. lumb., a. vert. puhul ei saaks lühendi üksikkomponente eraldi vaadeldes aru, millest räägitakse (sh mida vert. igal konkreetsel juhul tähendab), samas koondlühendina on mõiste selgelt määratletud); b) kuidas lühendit korpusest leida: ülesannet raskendas oluliselt lühendamise arbitraarsus ja lühendite suur variatiivsus (sh tingitud ka näpuvigadest, suur- ja väiketähtede, kirjavahemärkide ja tühikute kasutamisest, käändelõppude lisamisest, sõnede (sh ka lühendite) liitumisest ning lühendatud termini keelest (nt mõiste ühisreiearter esines lühendatult kokku 68 korda, lühendid: AFC (48), CFA (16), A.FC (1), a. fem.com (1), a. fem.comm. (1), A fem.comm (1)); c) lühendite kattuvus eri mõistete kirjeldamisel (nt l/s: muutused l/s rinnaosas (l/s=lülisamba), lülisamba l/s osas (l/s=lumboskaraal-); suurenenud l/s kaelal (l/s=lümfisõlm); d) mida lühend kontekstis tähendab – kuidas arvestada vahetut (lühendi tähendus trigrammide põhjal) ja laiemat konteksti (patsiendi eripära (vanus, sugu), anatoomiline piirkond, radioloogilise uuringu liik (röntgen, ultraheli jne) ja kasutatud metoodika (nt kiire suund) määravad, mida on võimalik visualiseerida ja kirjeldada) ja lõpuks e) teksti autori mõju lühendi kasutussagedusele korpuses. Lühendite korpusest leidmiseks koostati reeglid, mille alusel õnnestus välja sõeluda 14 961 lühendikandidaati, mida oli kokku kasutatud 1 250 260 korda (s.o 10,5% korpuse kõigi sõnede suhtes). Läbivaatamisel osutus neist lühenditeks vaid 10 606, esinemissagedusega 446 158 korda (s.o 3,8%). Lühendid rühmitati ühe- ja mitmetähenduslikeks lühenditeks.

**2013. a** jätkati **lühendite analüüsi**. Ühetähenduslike mitmemärgiliste lühendite rühmast analüüsiti kõik sagedusega kuni viis korda korpuses esinenud sõned nii, et alguses määras nende tähenduse üks arst ning seejärel vaatas kogu töö üle teine arst (radioloog). Nende lühenditega oli ühetähenduslikult lühendatud 672 mõistet, kokku kasutati ühetähenduslikke mitmemärgilisi lühendeid 334 505 korda (korpuse sõnedest umbes 2,8%). Mitmetähenduslikest lühenditest analüüsiti vahetus kontekstis (trigrammides) mittetõstutundlikult ja lisatud kirjavahemärkidest sõltumatult kõigi ühetäheliste lühendite kõik tähendused. Üksiktähti esines korpuses kokku 80 571 korda. Neid kasutati lühendina 44 034 korda ja tähisena 34 741 korda. Üksiktähed osutusid näpuvigadeks 305 korral, tähendus jäi teadmata 906 korral (ligikaudu 1%). Üksiktähed võisid olla ka muutelõpud (402) või mitmetäheliste lühendite osad (183), mis jäid tekstis tühikute või kirjavahemärkide tõttu eraldi. Uuritutest olid kolm sagedaimat lühendit üksiktäht X (x) tähenduses „korda″ (31 177), P-A tähenduses „posterio-anterioorne″ (26 123) ja sh. tähenduses „sealhulgas″ (25 316). Kümne sagedaima lühendi hulgas oli lühend X (x) ainsana mitmetähenduslik (tähendused: korda, oktoober, kümnes, x (iks)). Lühendite analüüs võimaldas tekstist ühetähenduslikult kindlaks teha 811 mõistet, neid kirjeldavaid termineid oli kokku lühendatud 2453 korda (keskmiselt 3 lühendit mõiste kohta). Mõisteid, mida kirjeldavaid termineid lühendati alati vaid ühtmoodi, oli 46,7% (nt uuringute metoodika spetsiifilised lühendid nagu MRA, TRUS, PNB, MRCP, FLAIR, MTF jne). Mõisteid, mille terminite lühendamisel kasutati palju erinevaid lühendivariante, oli eriti palju anatoomiliste struktuuride nimetuste seas: nt mõiste viies metatarsaalluu (nii selle eesti- kui ka ladinakeelsete terminite) lühendamiseks kasutati 25, mõiste ühissapijuha (termin ductus choledochus) lühendamiseks 24 ja mõiste õndlaarter (termin arteria poplitea) lühendamiseks 27 lühendivarianti. Mõiste kohta kasutatud eri lühendite rohkust mõjutas ka lühendi algupära keel: ladinakeelsetel terminitel oli üldiselt rohkem lühendivariante. Anatoomiliste struktuuride kohta kasutati eelistatult ladinakeelseid termineid ning nende rohkuse tõttu tekstides moodustasid ladina algupära lühendid enam kui kaks kolmandikku kõigist uuritud lühenditest (69,8%). Lühendite analüüsi tulemused on esitatud avaldamiseks Eesti Arstis.

# SUULISE EESTI KEELE AUDIOVISUAALSE SUHTLUSKORPUSE KOGUMINE JA PÄRINGUSÜSTEEMI ARENDAMINE

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Tiit Hennoste |
| **Teised täitjad** | |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 20 000 € (2012), 20 000 € (2013) |

## PROJEKTI EESMÄRK

Arendada edasi kõnekeele korpuse kogumise seniseid projekte aastatest 2004-2010 (täpsutada ja täiendada videomaterjali kogumise juhendeid, transkribeerimise süsteemi, taustakirjelduste skeemi, hoida korpust töökorras ja aidata soovijatel seda kasutada jne)

Filmida ja salvestada suulise eesti keele kasutust tegelikes suhtlussituatsioonides (multimodaalsed videosalvestused, mis võimaldavad analüüsida verbaaalse ja mitteverbaalse suhtluse koostööd, institutsionaalne telefonisuhtlus, sh suuline materjal Dialoogikorpuse tarvis, meediasuhtlus)

Transkribeerida tekstid ja varustada taustakirjeldusega keele kasutust mõjutavate keeleväliste nähtuste kohta. See tegevus võtab põhilise osa tööajast.

Arendada arvutitarkvara, mis võimaldab otsida korpusest erinevaid keelelisi nähtusi ning neid analüüsida (lisada juurde erinevaid sõnavariantide otsimise parameetreid.

Korpuse olemus:

- avatud pragmaatilis-suhtluslik keelekorpus;
- sisaldab reaalsetes situatsioonides toimuvat suhtlust: argine-avalik, silmast-silma, telefoni- ja meediasuhtlus, dialoog-monoloog, spontaanne-redigeeritud suhtlus;
- koosneb salvestustest, transkriptsioonidest, taustakirjeldustest ja otsingutarkvarast.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Projekti materjali kogumise ja transkribeerimise osa on üldiselt täidetud.

Korpuse maht (2013 lõpp) on umbes 3200 vestlust, sellest litereeritud 2300 vestlust, umbes 1 900 000 sõna. Juurde on transkribeeritud ca 400 000 sõna materjali. Videokorpuses on 75 salvestust, litereeritud audioosa 13 tervikuna, 40 osaliselt. Juurde on tulnud ca 25 tundi salvestusi. Osalt on täitmata avalike meediatekstide korpusesse toomise osa.

### Korpuse kasutamine

Projekti abiga täiendatud korpus on kasutatav ainult uurimiseks ja õppetööks kõigile uurijatele (v.a. vähesed väga tundlikud materjalid, nagu arsti-patsiendi vestlused, mis on väljapoole töörühma suletud). Materjalide saamiseks tuleb esitada soov korpuse administraatorile ja allkirjastada ülikooli juristi poolt kinnitatud konfidentsiaalsuskohustus. Päringusüsteem ja selle veebiliides praegu väljapoole kasutatav ei ole.

Projekti keskseks koostööpartneriks oli Mare Koidu projekt “Eestikeelse dialoogi pragmaatika analüsaator“. Projektis loodud vahendeid ja teadmist on kasutatud “Võru ja seto keelekorpuse projektis”.

Korpuse materjalide põhjal on valminud kõik uurimistööd, mis suulise eesti keele kohta Eestis on tehtud. Korpuse materjale kasutatakse eesti keele võõrkeelena õpetamise jaoks, et muuta õpetatavat dialoogi tegeliku keele lähedaseks.

Korpuse materjale on kasutatud teistes riikides, kus uuritakse suulist eesti keelt ja suhtlust (Soome, Rootsi).

**Korpuse säilitamine**

Korpust säilitatakse Tartu ülikoolis. Salvestused säilitatakse DVD-del (vanad lisaks ka analooglintidel), kahes arvuti välismälus ja dialoogikorpuse tööpingis www.dialoogid.ee/dialoogid. Litereeringud ja taustakirjeldused asuvad välismäludes ja arvutites. Kõik materjalid on mitmekordselt dubleeritud.

Päringusüsteemi lähtekood asub BIIT repositooriumis: https://biit-dev.cs.ut.ee/trac/tools/browser/SpokenLangCorpus/trunk. Töötav rakendus jookseb BIIT-dev serveris aadressil https://biit-dev.cs.ut.ee/~orasmaa/dev/suulinekone/. Mõlemad on parooli all.

# UUED RESSURSID MASINTÕLKES

| | |
|---|---|
| **Organisatsioon** | Tartu Ülikool |
| **Vastutav täitja** | Heiki-Jaan Kaalep |
| **Teised täitjad** | Mark Fišel, Kaarel Veskis, Urmo Visk, Siim Orasmaa |
| **Projekti kestus** | 2011-2013 |
| **Finantseerimine** | 127 000 € |

## PROJEKTI EESMÄRK

2011 alguseks oli TÜ-s üles seatud platvorm fraasipõhise SMT eksperimentide läbiviimiseks (Moses) ja treenitud teda olemasolevate paralleelkorpuste peal (eesti/inglise sõnu, miljonites): JRC-Acquis – EL seadusandlus, 18,4/25; EMEA – meditsiin, 9,6/11,1 KDE4 – infotehnoloogia, 1,6/1,9; OPUS – varia, 1/1). Kuid katsetega selgus, et nende paralleelkorpuste peal treenitud süsteemid ei saa hästi hakkama selle keelega, mida kasutajad soovivad: palju on tundmatuid sõnu ja ka lauseehitus on erinev. Ehk teiste sõnadega - need süsteemid ei ole hästi porditavad.

Projekti eesmärgiks oli arendada eesti keele statistilist fraasipõhist masintõlget sel moel, et anda talle paremaid materjale – korpusi ja sõnastikke – mille pealt treenida.

Keskenduti inglise-eesti keelepaarile.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Süsteemi kvaliteedi adekvaatseks hindamiseks loodi nn. TempEst korpus: inglise keelde tõlgitud eestikeelsed laused, mille kasutajad on andnud masintolge.ut.ee lehele tõlkida. Korpuses on kokku 2800 lauset, eesti keeles 23 000 sõna. Ta peaks esindama keelt, mida kasutajatel tõepoolest vaja läheb (erinevalt nt. EL seadustekstide korpusest).

Uute paralleelkorpuste korjamisest loobuti, sest see töö on ära tehtud mujal, muude keelte paralleelkorpuste korjamise kontekstis (eelkõige OPUS, http://opus.lingfil.uu.se/). Selle asemel kontrolliti paralleelkorpuste kvaliteeti automaatselt, täpsemalt korpusefailide omavahelist kattuvust ja sarnasust (korpuses esineb nii sama algteksti alternatiivseid tõlkeid kui lihtsaid duplikaate, mis erinevad üksteisest nt ajastuskoodide poolest või algustiitrite olemaolu/puudumise poolest, aga ka juhtumeid, kus sama mitmeseerialine film on esitatud kord ühe terve subtiitrifailina, kord aga eraldi seeriatena). Selle tegevuse eesmärk oli automaatselt tuvastada nii duplikaadid (mis rikuvad statistilist jaotust) kui ka paralleelistusvead. Informatsioon korduvuste kohta on väljas OPUSe kodulehel (http://opus.lingfil.uu.se/OpenSubtitles2011/overlaps/et/). Paralleelistusvigade automaatne tuvastamine ebaõnnestus: ei õnnestunud leida tunnuseid, mille alusel saaks öelda, kummas kahest paralleelfailist leidub vähem vigu. Seejuures oli ootamatuks avastuseks asjaolu, et peale sihtkeele failide eri versioonide võib lausestus olla erinev ka sama filmi lähte-versioonides.

Lisaks paralleelkorpustele lisati treenimismaterjali hulka ka Eesti Keele Instituudi inglise-eesti sõnastik ftp://ftp.eki.ee/pub/keeletehnoloogia/inglise-eesti/), mille sõnadest genereeriti muutevormid, et parandada süsteemi võimet katta korpusest puuduvaid sõnavorme. Seejuures genereeriti muutevormid ka mitmesõnalistest sõnastikuvastetest, arvestades nende tüüpe. Kõige tüüpilisemad juhtumid olid: a) muutumatu osa, millele järgneb üks käänduv/pöörduv sõna („aafrika kannike“, „jalga laskma“); b) omadussõna + nimisõna („kaunis neiu“) - käänduvad mõlemad kõigis vormides, v.a. omadussõna neljas viimases käändes; c) nimisõna, millele järgneb muutumatu sõna („mees metsast“) – ainult nimisõna muutub.

Selgus, et sõnastiku lisamine andis tõlkekvaliteedile juurde väga vähe.

Kõiki olemasolevaid ressursse kasutades treeniti uued süsteemid ja pandi nad avalikult kasutamiseks välja:

1. eesti -> inglise ja inglise -> eesti (http://masintolge.ut.ee)

2. eesti -> prantsuse ja prantsuse -> eesti. (http://masintolge.ut.ee/fr/) Seejuures kasutati EKT projektis „Eesti-prantsuse paralleelkorpus“ loodud korpust.

# VÕRU JA SETO KEELEKORPUS

| | |
|---|---|
| **Organisatsioon** | Võru Instituut |
| **Vastutav täitja** | Sulev Iva |
| **Teised täitjad** | Mariko Faster, Laivi Org, Kaur Männamaa jt |
| **Projekti kestus** | 2011-2014 |
| **Finantseerimine** | 16 000 € (2012), 25 000 € (2013), 25 000 € (2014) |

## PROJEKTI EESMÄRK

- Projekti eesmärgiks on ette valmistada võru ja seto keelele keeletehnoloogilise toe loomist läbi võru ja seto keeleressursside koondamise ja süstematiseerimise ühtseks keelekorpuseks. Võru ja seto keele arendamist ja laialdasemat kasutamist on peetud tähtsaks nii kohalkul (maakondade ja omavalitsuste arengukavades) kui ka riiklikul tasandil (Võru Instituudi töö, kultuuriministeeriumi Vana Võrumaa ja Setomaa programm, Eesti keele strateegia ja keelseaduse sätted eesti keele piirkondlikest erikujudest).
- On üldiselt teada, et tänapäeva maailmas ei saa säilida ega jätkusuutlikult areneda keeled, millele pole loodud vähimatki keeletehnoloogilist tuge. See kehtib ka võru ja seto keele kohta, mis on 2009. aastal kantud UNESCO ohustatud keelte nimekirja. Setokeelne leelotraditsioon on samas kantud ka UNESCO maailma vaimse kultuuripärandi nimekirja. Nüüd, kui eesti keeletehnoloogias on saavutatud eesti kirjakeelele keeletehnoloogilise toe loomisel juba arvestatav tase, oleks igati loomulik ja tänuväärne rakendada loodud baasi ja kogemusi ka Eesti põliste kohakeelte revitaliseerimisel.
- Keeletehnoloogiline esmavajadus võru ja seto keele puhul oleks võru kirjakeelel ja kirjalikul seto keelel põhineva keelekorpuse ja sellest lähtuva võru-seto automaatkorrektuuri ja -poolitaja loomine. Teatud alus selleks tööks on loodud juba 1995. aastal ilmunud doktoritöös seto verbi grammatika ja sõnastikega (Toomsalu 1995), oma grammatikaosas suuresti sellest lähtunud võru-eesti sõnaraamatu (Iva 2002) ja praegu koostamise lõppjärgus oleva eesti-võru sõnaraamatuga, mis sisaldavad nii võru, seto kui ka laiemalt lõunaeestilisel keelekasutusel põhinevat võru kirjakeele sõnavara ja grammatikat. Keeletehnoloogiliseks uurimis- ja arendustööks vajalik võru kirjakeele muutemorfoloogia põhjalikum käsitlus leidub doktoritöös Võru kirjakeele sõnamuutmissüsteem (Iva 2007), kuid väga kasulik eeltöö on selleks tehtud ka EKI-s morfoloogiliselt märgendatud korpuseks arendatud Salme Nigoli Hargla konsonantismi käsitluse näol. Hulgaliselt morfoloogiliselt märgendatud võru ja setu tekste leidub TÜ murdekorpuses ja suulise kõne korpuses.
- Esimese etapina ülalnimetatud võru ja seto keeleressursside koondamine, süstematiseerimine ja täiendamine ühtseks keelekorpuseks ja vajalike kasutajaliideste loomine ning edaspidi loodud korpuse täiendamine ja laiendamine ning selle põhjal keeletehnoloogiliste rakenduste (automaatkorrektuur, poolitaja jm) loomine ongi käesoleva projekti tööülesandeiks. Lisaks nimetatud ressurssidele on plaanis korpusele lisada võru ja seto ajakirjandustekstide osa (ajalehtede Uma leht ja Setomaa elektrooniliste arhiivide sisu põhjal) ja Võru Instituudis säilitatavate võru kirjakeele allikate osa (õpikute, ilukirjanusteoste jm elektrooniliste tekstide põhjal). Koostöös TÜ murdekorpuse ja suulise kõne korpuse arendajatega saab korpusele liita seal olemasoleva võru ja setu materjali nii tekstina kui helifailidena. Keeleressursside kogumise ja litereerimise osas on plaanitud koostöö TÜ Lõuna-Eesti keele- ja kultuuriuuringute keskusega. Korpusega saaks liita ka (Triin Iva doktoritöö raames tehtud) teadaolevalt ainsad võru väikelastekeele salvestised. Nii kirjalikke kui suulisi keeleressursse tuleb lisaks olemasoleva materjali koondamisele ja

süstematiseerimisele pidevalt täiendada uue keelematerjali kogumise, litereerimise ja märgendamisega. Korpuse suulise kõne pool loob aluse selleks, et tulevikus saaks võru ja seto keelega arvestada ka eesti kõnetuvastuse ja -sünteesi arendamisel.

**PROJEKTI TULEMUSED (aasta 2013 lõpuks)**

Korpuse kolmandal tööaastal (2013) on jätkatud plaanitud mahus projekti eelmiste aastate põhilisi töösihte ja lisaks alustatud mitme uue tegevusega:

- Jätkatud on kirjalike ja suuliste tekstide kogumist.

- Suurimas mahus on kogutud ajakirjandustekste:

- võrukeelse ajalehe Uma Leht tekstikorpust on täiendatud ca 262 000 sõna mahus

- loodud on setokeelse ajalehe Setomaa tekstikorpus mahuga ca 279 000 sõna seto keeles ja ca 228 000 sõna eesti keeles.

- Setomaa lehe korpuse kasutamiseks on loodud uus kasutajaliides, mis asub aadressil: synaq.org/seto

- Uusi salvestisi on tehtud ca 9 tundi, peamiselt spontaanseid argivestlusi.

- Litereeringuid on tehtud ca 36 000 sõna mahus. Litereeritud on peamiselt võrukeelseid spontaanseid argivestlusi. Kõik uued, 2013. aastal tehtud litereeringud on tehtud helifailiga seotult ja ajaliselt joondatult programmi ELAN abil.

- Alustatud on eesti-võru paralleelkorpuse loomist, esialgu tekstiarhiivina, mis hetkel sisaldab 43 teksti kummaski keeles (kokku 20 684 sõna).

- Alustatud on ettevalmistusi eesti-võru masintõlke, sh masintõlke esialgne katserakenduse loomiseks.

- Koostöös TÜ foneetikalabori ja EKI kõnesünteesi töörühmaga on alustatud ettevalmistusi võrukeelse kõnesünteesi esimeste prototüüpide loomiseks. Toimunud on konsultatsioonid EKI spetsialistidega võru difooniandmebaasi koostamispõhimõtete osas.

- Täiendatud on kirjakorpusse tekstiarhiivina kogutud ligi 100 000-sõnalist ilu- ja tarbekirjanduse ning dokumentide allkorpust.

# EESTI AVATUD PARALLEELKORPUS

| | |
|---|---|
| **Organisatsioon** | Tilde Eesti OÜ |
| **Vastutav täitja** | Margit Kurm |
| **Teised täitjad** | keeletehnoloogid Sander Vahter, Tiina Köömnemägi ja Martin Luts, keeletoimetajad Tiit Päeva ja Tiina Altküla, tehnilised töötajad |
| **Projekti kestus** | 2012-2014 |
| **Finantseerimine** | 77 505 € |

## PROJEKTI EESMÄRK

Projekti „Eesti avatud paralleelkorpus" eesmärk on luua oluline kogus keeleressursse statistiliste masintõlkesüsteemide parendamiseks. Projekt aitab kaasa olukorra saavutamisele kus:

(i) Erinevad kommerts- ja kogukondlikud masintõlkesüsteemid pakuvad kvaliteetset tõlketeenust.

(ii) Masintõlkesüsteemide teenused on lõppkasutajatele võimalikult väheste piirangutega (tasu, maht, kasutatavad platvormid) kättesaadavad.

(iii) Sõltuvus üksikutest masintõlketeenuste kommertsteenusepakkujatest ei ole kriitiline ja on asendatav avatud ning vabavaraliste lahendustega.

(iv) Masintõlkealaseks teadus- ja arendustegevuseks on kättesaadav piisavalt paralleelkorpuseid.

(v) Eesti keele keeletehnoloogia ressursid on Euroopa keeletehnoloogia taristu kaudu saadaval kõrvuti teiste keelte ressurssidega.

Projekti mõõdetavad tulemid on:

(i) Kogutud ja korrastatud paralleelkorpuste maht. Projekti lõpuks vähemalt 15 miljonit ühikut.

(ii) Kogutud korpuste täiendavalt olemasolevatele korpustele abil treenitud masintõlkesüsteemide kvaliteedinäitajate parenemine (mõõdetakse koostöös masintõlkesüsteemide omanikega).

(iii) Kogutud korpused aktsepteeritud ja publitseeritud METASHARE (http://www.metanet.eu/metashare) ja CLARIN (http://www.clarin.eu/external/) baasides.

Projekti tulemina loodava paralleelkorpuse omadused:

• Lause tasandil joondatud inglis-eesti paralleelkorpus.

• Erinevate ainevaldade katvus.

• Korpus on kättesaadav tasuta ja piiranguteta kasutamiseks kommerts- ja vabavararakendustes, edasiarendusteks jm. Korpus on allalaaditav nii METASHARE taristu kui ka CLARINi võrgustiku kaudu.

Avatud ja tasuta paralleelkorpus on olulisim statistilise masintõlke ressurss. Selliste korpuste vähesus ja ühekülgsus (keskendumine kitsale ainevallale, nt seadusandlus) ning mittevastavus lõppkasutajate keelele on statistiliste masintõlkesüsteemide madala teenuskvaliteedi olulisimad põhjused. Käesolev projekt on loodud nende põhjuste mõju vähendamiseks. Kogutud paralleelkorpus on kasutatav ka teistes valdkondades, sh terminoloogiatöö.

Projekti tulemuste rakenduskohad:

1. Keeletehnoloogia programmi raames. Projekti käigus loodavad paralleelkorpused on rakendatavad projekti Masintõlge jätkuprojektides (masintolge.ut.ee).

2. Väljaspool keeletehnoloogia programmi. Projekti tulemid on kasutatavad statistiliste masintõlkesüsteemide omanike poolt. Taotleja pakub projekti tulemeid proaktiivselt huvilistele. Projekti tulemid on kasutatavad ka terminoloogiatöös jm rakendustes (terminology extraction, named entity extraction, translation studies).

**PROJEKTI TULEMUSED (aasta 2013 lõpuks)**

Arvestades juhtkomitee 2012 aasta soovitust "EKT juhtkomitee on väga huvitatud ka teiste keelte paralleelkorpustest" laiendati Eesti Avatud Rööpkorpusesse kogutavate keelepaaride hulka – lisaks 2012 aastal kogutud eesti-inglise rööptekstidele koguti 2013 aastal eesti-vene ja eesti-läti rööptekste. Neist viimast koguti koostöös Eesti Keele Instituudiga ning korpust kasutatakse (lisaks masintõlkesüsteemide treenimisele) eesti-läti, läti-eesti sõnaraamatu koostamisel EKI ja Läti Keele Keskuse piiriüleses ühisprojektis.

Eesti-inglise rööptekstide kogumisel keskenduti valdkondadele millised annavad enim väärtust statistiliste masintõlkeprogrammide treenimisel – so võimalikult kasutajate tõlgitavate tekstide lähedaste korpuste kogumisele. Arvestades küllastatust avaliku sektori materjalide hulgas (kümned miljonid segmendid Euroopa Liidu Parlamendi, õigusaktide, Keskpanga jt asutuste materjalid) koguti 2013 aastal valdavalt ilukirjandust, so tõlketekste, arvestades autoriõiguslikke piiranguid. Kogutud tekstide maht on 2013 aasta lõpuga niivõrd mahukas et osutus võimalikuks treenida Google Translate'i masintõlkega võrreldava ja kohati parema tõlkekvaliteediga masintõlkesüsteem, millise avas 8. novembril 2013 Eesti Vabariigi president Toomas Hendrik Ilves (http://www.tilde.ee/presidentilvestutvustildemasintolketehnoloogiaga).

Töö tulemusi on tutvustatud ülikoolides (sh Tallinna Ülikoolis tõlkemagistritele), erialakonverentsidel ning avaliku ja erasektori organisatsioonides (masintõlke seminarid Riigi Infosüsteemi Ametis, Eesti Infosüsteemide audiitorite teabepäevadel jm), osaletud masintõlkega seotud magistritööde juhendamisel ja oponeerimisel.

Korpuses on eesti-inglise rööptekste ca 6,5 miljonit sõna, vene-eesti 1 ja läti-eesti 1 miljonit sõna.

Kogutud korpuste abil treenitud masintõlkesüsteemide kvaliteedinäitajad paranesid niivõrd et tulemused on, sõltuvalt hindamismetoodikast (BLEU ja kasutatav hindamiskorpus, inimeste poolt läbiviidud pimetestid kus tuli valida etteantud kahe masintõlke vahel parim) samaväärsed või kohati paremad kui nt Google Translate'i ja MS Bingi masintõlkesüsteemidel.

* Inglise-eesti tõlkesuund, üldvaldkond

Kogutud rööpkorporaga treenitud süsteem – BLEU 24.22 ACCURAT Balanced Evaluation Set +

Google – BLEU 21.46 ACCURAT Balanced Evaluation Set +

* Eesti – inglise tõlkesuund, üldvaldkond

Kogutud rööpkorporaga treenitud süsteem – BLEU 22.77 Tempest; 37.97 ACCURAT

Google – BLEU 23.06 Tempest; 35.30 ACCURAT Balanced Evaluation Set +

Microsoft – BLEU 30.14 ACCURAT Balanced Evaluation Set +

Tartu Ülikool – 26.11 ACCURAT Balanced Evaluation Set +

Kogututi tekste ilukirjanduse valdkondades saamaks tervikuna tasakaalus korpust (võrdle omadussõnade sagedust nt seadustekstides ja ilukirjanduses). Tulenevalt autoriõiguste piirangutest korrastati töö tulem ning avaldati ngram'dena millised ei võimalda rööpkorpustest algset teost taastada. Selline lähenemine võimaldab koguda parimat saadaolevat, igapäevasele tööalasele keelele lähimat ning masintõlkesüsteemide lõppkasutajale enim väärtust loovat masintõlkesüsteemide treeningmaterjali õiguslikus raamistikus.

Korpuse kogumisel keskenduti ilukirjandusele. 2013 aasta lõpu seisuga võib hinnata et e-raamatute osa on ammendumas (suurusjärgus 100 e-raamatut kaardistatud), seetõttu alustasime sügisel 2013 piloteerimist paberraamatute skanneerimise ja OCR tehnoloogiatega. DIGAR andmebaasi andmetel on eesti-inglise rööptekstide maht ca 18000 nimetust, st tööpõld on siin järgnevateks aastateks kindlustatud – mis lõppkokkuvõttes aitab oluliselt tõsta masintõlke kvaliteeti.

Valdkondade (avalik sektor, seadusaktid, ilukirjandus) mitmekesistamiseks alustasime koostööd ICC Eestiga ärivaldkonna tekstide (nt INCOTERM) kaasamisega rööpkorpustesse. Samuti on lisatud firmade Nokia, Samsung, Apple jt avalikke juhendmaterjale.

Avalikust sektorist on kogutud ainult mitmekesistavat materjali, nt Terviseameti ja Keskkonnaameti, Ettevõtluse Arendamise Sihtasutuse materjale millised sisaldavad varem korpuses puudunud valdkondi ja sõnavara.

# EESTI KEELERESSURSSIDE KESKUS

| | |
|---|---|
| **Organisatsioon** | Konsortsium: Tartu Ülikool, TTÜ Küberneetika Instituut, Eesti Keele Instituut |
| **Vastutav täitja** | Kadri Vider |
| **Teised täitjad** | TÜ töörühm: Krista Liin, Neeme Kahusk, Margus Treumuth, Lauri Jesmin, Indrek Jentson, Rait Talvik, Aleksei Kelli; |
| **Projekti kestus** | EKT10 projekti kestus 2011-2014 |
| **Finantseerimine** | 182 448 € |

## PROJEKTI EESMÄRK

Eesti Keeleressursside Keskus (EKRK) on teadustaristu, mis teeb huvilistele kättesaadavaks eesti keele digiressursid ja –tehnoloogia (http://keeleressursid.ee, http://ee.clarin.eu)

Keskuse tegevuse eesmärgiks on koondada olemasolevad digitaalsed keeleressursid (sõnastikud, teksti- ja kõnekorpused, keeleandmebaasid) ja keele töötlemise vahendid (tarkvara) vastastikku toimivaiks ning oskusteabega varustatud teenusteks, mida kasutajad saavad vajaduse korral ka oma tarbeks kohandada. Kõik keeleressursid peavad vastama rahvusvahelistele standarditele ja olema varustatud kasutuslitsentsiga. Eestisiseselt on taristu avatud nii keeleressursside omanikele, arendajatele kui ka kasutajatele, kes nõustuvad kasutus- ja litsentsitingimustega. Eelistatud on teaduskasutajad ning eesti keele kasutajaskonnale avalikes huvides pakutavad keeleressursid ja keeletehnoloogilised vahendid. Erasektoril võimaldatakse erikokkulepete alusel keeleressursse kasutada ja rakendada ka oma toodetesse.

Keskusele on pandud kohustus arhiveerida ja teha kättesaadavaks ka riiklike programmide EKKTT ja EKT projektide tulemused. Keskuse üks ülesandeid on korraldada ka EKT riikliku programmi haldamise tegevus.

Tartu Ülikooli, Eesti Keele Instituudi ja Tallinna Tehnikaülikooli Küberneetika Instituudi konsortsiumina on EKRK „Eesti teaduse infrastruktuuride teekaardi“ riikliku tähtsusega teadustaristu, mida rahastatakse ka ERFi vahenditest ja CLARIN (Common Language Resources andTechnology Infrastructure) ERIC (European Research Infrastructure Consortium) riiklik konsortsium. Eesti Keeleressursside Keskuse osalemine rahvusvahelises võrgustikus annab meie teadlastele juurdepääsu keskuste võrgustikus olevate keskuste teiste keelte jaoks loodud ressurssidele.

Suurem osa EKRK kui konsortsiumi kavandatud mahus käivitamiseks vajaminevast rahast tuleb 2012-2015 EL ERFi alameetmest "Riikliku tähtsusega teaduse infrastruktuuri kaasajastamine", riigieelarvelistest vahenditest kaetakse kõigi partnerite tõukefondide omafinantseering ning mõningaid abikõlbmatuid, kuid vajalikke tegevusi.

## PROJEKTI TULEMUSED (aasta 2013 lõpuks)

Paljuski on tulemused seotud EL ERFi struktuuritoetuse projekti tegevuskavas kavandatust ja CLARIN ERICu asutamise järel vajalikest tegevustest saamaks Eesti riiki esindavaks keeleressursside keskuseks Euroopa teaduse infrastruktuuris.

1. Keskseid teenuseid pakkuma hakkavad serverid TÜ juures on hangitud ja tööks seadistatud. Loodud ja tulevikus loodavate virtuaalserverite funktsionaalne süsteem on loodud.

Virtuaalservereid on kasutamiseks jagatud ka EKT programmiga seotud projektidele arendustegevusteks, ja majutamiseks.

2. Esialgu on repositooriumitarkvarana kasutusel META-NETi võrgustikus loodud META-SHARE (https://metashare.ut.ee/), seadistamist ja pealiskihi liidestamist ootab Fedora-põhine repositooriumitarkvara, mis vahendab paremini CLARINi võrgustiku nõuetele vastavat meta-andmestikku CMDI formaadis.

META-SHARE registrisse on kantud ligi 30 Eesti keeleressursi detailsed meta-andmed, mille hulka kuulub info nii ressursside keeletehnoloogilise sisu, tehniliste parameetrite, kättesaadavuse kui ka tõestatud kasutusõiguste (litsentside) kohta.

3. Nn SSO-login ehk kasutajate autentimine ja autoriseerimine TAATi kaudu (http://taat.ee/main/ ) on META-SHARE repositooriumis loodud ja seadistatud, teiste EKRK teenuste liidestamine käib. Töö partnerite (TTÜ ja EKI) liidestamiseks identiteedipakkujatena käib. EKRK teenuste pakkumine rahvusvaheliselt on korraldatud CLARIN ERIC teenusepakkujate föderatsiooni (SPF) kaudu.

4. Töötab PID-teenus ehk arhiveeritavatele andmetele püsi-identifikaatorite omistamine ja nende lahendamine. TTÜ KübI kõneressurssidele on omistatud juba üle 52000 PIDi.

5. CLARIN ERICu riikliku keskuse tegevus: riiklike koordinaatorite foorumi (NCF) töös osaleb Kadri Vider, tehniliste keskuste komitees Krista Liin, meta-andmete ja standarditega tegelevad Neeme Kahusk ja Tõnis Nurk, õigusasjade komisjoni töös osaleb aktiivselt Aleksei Kelli. EKRK kui CLARINi Eesti tehnilise keskuse hindamise (assessment) protsess on käivitatud. Osaletud on aastakoosolekutel ja CLARIN ERIC Üldkogul.

6. Kasutuslitsentside ja ressursside intellektuaalomandi õiguslikud küsimused, kasutatavate lepinguvormide koostamine: Ressursside kasutuslepingute alusena on otsustatud võtta kasutusele Creative Commonsi laadsed lepingud, mistõttu ei hakata välja töötama eraldi lepinguid, vaid kohandatakse vajadusel lisapiiranguid. Eesmärk on EL-i tasemel erandi kehtestamine keeleressursside loomiseks. Ka Eesti uude autoriõiguse eelnõusse on lisaks üldisele teadustöö erandile viidud sisse keeleressursside erand.

7. Keskuse ja ressursside tutvustamine kasutajatele, kasutajate harimine: Keskuse töötajate idee põhjal on valminud EKRK logoraamat ja logo, mis kujutab murakat. Keskuse koduleht on selle põhjal saanud ka uue kujunduse (vt www.keeleressursid.ee). Eelkõige CLARINi võrgustikus levitamiseks valmis ingliskeelne EKRK-d tutvustav infovoldik, hiljem kohandati see eestikeelseks. TÜ töötajad on Keskust tutvustanud ja kasutajaid harinud igapäevase tegevuse käigus, Eesti keeleressursse on ettekannetes esile toodud ka CLARINi aastakogunemistel.