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	Eesti Keele Instituut
	Indrek Vainu	OÜ Tarkvara Tehnoloogia Arenduskeskus
	Uuno Vallner	Majandus- ja kommunikatsiooniministeerium
	Kadri Vider	Eesti Keeleressursside Keskus, Tartu Ülikool
	Jan Villemson	AS Cybernetica
	Oliver Väärtnõu	ELIKO Tehnoloogia Arenduskeskus OÜ

Programmi koduleht www.keeletehnoloogia.ee

EKT 2012 KONVERENTSI AJAKAVA

	teisipäev, 2. oktoober 2012	
10.10-10.30	avatakse registreerimine ja pakutakse kohvi	
10.30-11.00	**konverentsi avamine, META-NETi algatuste ja keeleraportite esitlus**	Jaak Vilo, Kadri Vider
11.00-14.00	**avatud seminar „Keeleressursside õiguslikud aspektid“**	
	Rahvusvaheline koostöö keeleressursside arendamisel ja kasutamisel	Kadri Vider
	Keeleressursside arendamise õiguslikud väljakutsed ja võimalused	Aleksei Kelli
	Creative Commons’i ja teiste tüüplitsentsilepingute kasutamise head ja vead	Peeter P. Mõtsküla
12.15-12.30	kohvipaus	
	Litsentseerimise olemus ja sellega kaasnevad ohud	Triin Tuulik
13.00-14.00	**Paneeldiskusioon keeleressursside loomisest ja kasutamisest**	Panelistid Kadri Vider, Peeter P. Mõtsküla, Triin Tuulik ja Liina Jents. Modereerib Aleksei Kelli
14.00-15.00	lõunapaus	
15.00-17.00	**ettekanded, juhatab Kadri Muischnek**	
15.00-16.00	Eesti Keeleressursside Keskus	Kadri Vider, Krista Liin, Margus Treumuth, Nee-me Kahusk
16.00-16.20	Võru ja seto keelekorpus	Sulev Iva
16.20-16.40	Eesti keele spontaanse kõne foneetilise korpuse arendused	Pärtel Lippus
16.40-17.00	Eesti-prantsuse paralleelkorpus	Madis Jürviste
17.00-18.30	**kohvipaus diskussioonideks, kohtumisteks**	

	kolmapäev, 3. oktoober 2012	
10.30-11.30	**ettekanded, juhatab Hille Pajupuu**	
	Kõnetuvastus	Tanel Alumäe
	Audiovisuaalse kõnesünteesi prototüüp	Einar Meister
	Kõnesünteesiliidesed	Meelis Mihkla
11.30-12.30	**kohvipaus + postrid, juhatab Arvi Tavast**	
	Uued ressursid masintõlkes	Heiki-Jaan Kaalep
	Semantika vahendid eesti keelele	Neeme Kahusk
	Autentse meditsiinikeele korpuse alusel radioloogia elektroonse piltsõnastiku koostamine	Eola Valdre
	Eesti avatud paralleelkorpus	Martin Luts
	Subtiitrite helindamise ja tele-eetrisse edastamise tarkvaralahendus	Meelis Mihkla
12.30-13.30	**ettekanded, juhatab Einar Meister**	
	Eestikeelse dialoogi pragmaatika analüsaator	Mare Koit
	Eestikeelsete dialoogsüsteemide loomise raamistik	Margus Treumuth
	Mallipõhine faktituletus tekstikorpustest	Timo Petmanson
13.30-15.00	lõunapaus	
15.00-16.00	**ettekanded, juhatab Heiki-Jaan Kaalep**	
	Vahendid teksti mitmekihiliseks märgendamiseks (rakendatuna Koondkorpusele)	Kadri Muischnek
	Leksikograafi töökeskkonna modifitseerimine	Arvi Tavast
	Eesti Wordnet'i täiendamine	Heili Orav
16.00-16.30	kohvipaus	
16.30-17.30	**ettekanded, juhatab Tanel Alumäe**	
	Kõne ja teksti emotsionaalsuse statistilised mudelid	Hille Pajupuu
	Kõne- ja multi-modaalsed korpused	Einar Meister
	Suulise eesti keele audiovisuaalse suhtluskorpuse kogumine ja päringusüsteemi arendamine	Tiit Hennoste
17.30-18.00	**konverentsi lõpetamine**	
20.00 -	**EKT2012 ja HLT2012 osalejate pidulik kohtumine AHHAA näitustesaalis**	

AVATUD SEMINAR
KEELERESSURSSIDE ÕIGUSLIKEST ASPEKTIDEST

11.00 - 11.15	**Kadri Vider**. Rahvusvaheline koostöö keeleressursside arendamisel ja kasutamisel
11.15 - 11.45	**Aleksei Kelli**. Keeleressursside arendamise õiguslikud väljakutsed ja võimalused
11.45 - 12.15	**Peeter P. Mõtsküla**. Creative Commons ja teiste tüüplitsentsilepingute kasutamise head ja vead
12.15 – 12.30	Kohvipaus
12.30 – 13.00	**Triin Tuulik**. Litsentseerimise olemus ja sellega kaasnevad ohud
13.00 – 14.00	**Paneeldiskusioon** keeleressursside loomisest ja kasutamisest. Panelistid Kadri Vider, Peeter P. Mõtsküla, Triin Tuulik ja Liina Jents. Modereerib Aleksei Kelli

ESINEJATE TUTVUSTUS

Kadri Vider on Eesti Keeleressursside Keskuse (EKRK), riikliku tähtsusega teaduse infrastruktuuri tegevjuht, CLARIN ERIC (*Common Language Resources and Technology Infrastructure European Research Infrastructure Consortium*) riiklik koordinaator Eestis ja TÜ arvutiteaduse instituudi keeletehnoloogia teadur. EKRK tähtsamaid ülesandeid on mh riiklike keeletehnoloogia teadus- ja arendusprogrammide EKKTT (2006-2010) ja EKT (2011-2017) projektide tulemusena loodud ressursside ja tarkvara haldamine, kättesaadavaks tegemine ning litsentsimine või sellealane konsulteerimine.
Varasemal ametikohal Haridus- ja Teadusministeeriumi teadusosakonna peaeksperdina esindas Kadri Vider HTMi ja teaduskollektsioonide huve ja seisukohti Kultuuriministeeriumi juures tegutsevas kultuuripärandi digitaalse säilitamise nõukogus, mille üks strateegilisi ülesandeid on ka autoriõiguse ja isikuandmete kaitse küsimuste reguleerimine digitaalse kultuuripärandi kontekstis.

Aleksei Kelli on intellektuaalse omandi dotsent (Tartu Ülikool) ning intellektuaalse omandi õiguse kodifitseerimise töögrupi juht (Justiitsministeerium). Dr. Kelli tegeleb digitaalsete keeleressursside õiguslike küsimustega Tartu Ülikoolis ning Eesti Keele Instituudis ning ta on intellektuaalse omandi ekspert CLARIN-i õiguskomitees.
Ta on töötanud külalisteadurina Columbia (Columbia Law School, Kernochan Center for Law, Media and the Arts) ja Uppsala (Uppsala University Innovation) Ülikoolis.
Dr. Kelli on panustanud intellektuaalse omandi eksperdina Majandus- ja Kommunikatsiooniministeeriumi ja Kultuuriministeeriumi tegevusse. Ta on ÜRO intellektuaalse omandi ekspertgrupi (*Team of Specialists on Intellectual Property. United Nations Economic Commission for Europe*) ja uue autoriõiguse seaduse väljatöötamise töögrupi liige ning teadus- ja innovatsioonipoliitika seire programmi IO seirevaldkonna põhitäitja. Aleksei Kelli on osalenud ka mitmetes EL-i ja Eesti teadusprojektides IO eksperdina, esinenud rahvusvahelistel ja siseriiklikel konverentsidel ning avaldanud rahvusvahelistes ajakirjades IO, innovatsiooni, IO ärimudelite ja teadmussiirde teemalisi artikleid.

Peeter P. Mõtsküla on Tartu Ülikooli õigusteaduskonna doktorant (ekstern). Ta töötab teenuste valdkonna juhina AS-s Proact Estonia ning osaleb eksperdina Justiitsministeeriumi intellektuaalse omandi õiguse kodifitseerimise töögrupis.
Hr Mõtsküla on õigusteaduse magister (Tartu Ülikool, 2009) ja diplomeeritud majandusinsener (Tallinna Tehnikaülikool, majandusteaduskond, infotöötluse eriala, 2000). Autoriõiguse ning info- ja sidetehnoloogia arengutega seotud küsimusi on ta põhjalikult uurinud alates 2006. aastast, mil kirjutas bakalaureusetöö tarkvara õiguskaitse perspektiividest võrgustatud ühiskonnas.
Hr Mõtsküla on eksperdina osalenud Creative Commons'i litsentside eestindamise projektis (Eesti Infotehnoloogia SA ja advokaadibüroo GLIMSTEDT, 2010) ja riigi IT koosvõime raamistikku kuuluva tarkvara raamistiku uuendamise projektis (RISO, 2011) ning esinenud mitmetel konverentsidel ja seminaridel ettekannetega autoriõigusest infoühiskonnas. 2012. aasta alguses võttis hr Mõtsküla aktiivselt osa ühiskondlikust diskussioonist võltsimisvastase kaubanduslepingu (ACTA) teemal.

Triin Tuulik on Tartu Ülikooli õigusteaduskonna doktorant ning töötab Advokaadibüroos GLIMSTEDT intellektuaalse omandi õiguse valdkonna advokaadina. Ta omab praktilisi kogemusi litsentsilepingute koostamisel, klientide esindamisel intellektuaalse omandi õigusi puudutavates vaidlustes ning igapäevasel nõustamisel peamiselt autori- ja kaubamärgiõiguse küsimustes. Oma doktoritöös analüüsib Triin Tuulik looja-keskset intellektuaalse omandi õiguse süsteemi ettevõtjate perspektiivist.

Liina Jents on Tartu Ülikooli doktorant ja töötab Advokaadibüroos Borenius intellektuaalse omandi valdkonna advokaadina. Liina Jents on lõpetanud Tartu Ülikooli Õigusinstituudi ja omandanud Stockholmi Ülikoolis magistrikraadi Euroopa intellektuaalse omandi õigus alal. Tema tegevusvaldkond advokaadina ulatub traditsioonilisest autoriõigusest ja autoriõigusega kaasnevatest õigustest kuni IT ja muude tehnoloogia küsimusteni, hõlmates endas ka kaubamärkide, patentide ja domeeninimedega seonduvat. Oma doktoritöös keskendub Liina Jents intellektuaalse omandi piirangutega seotud probleemidele.
Lisaks doktoriõpingutele ja advokaaditööle jagab Liina Jents oma kogemusi ja teadmisi intellektuaalsest omandist lektorina Tartu Ülikooli Viljandi Kulutuuriakadeemias ning erinevatel koolitustel ja seminaridel. Lisaks sellele osaleb Liina Jents eksperdina intellektuaalse omandi õiguse kodifitseerimise töögrupis (Justiitsministeerium) ning on aktiivne AIPPI (*International Association for the Protection of Intellectual Property*) Eesti töörühma liige.

„EESTI KEELETEHNOLOOGIA 2011-2017“ PROJEKTID EKT 2012 KONVERENTSIL

	PROJEKTI NIMI	PROJEKTI JUHT	ASUTUS	KESTUS
EKT10	**Eesti Keeleressursside Keskus**	Kadri Vider	TÜ, Matemaatika-informaatikateaduskond	2011-2014
EKT13	**Võru ja seto keelekorpus**	Sulev Iva	Võru Instituut	2011-2014
EKT4	**Eesti keele spontaanse kõne foneetilise korpuse arendused**	Pire Teras	TÜ, Filosoofiateaduskond	2011-2014
EKT19	**Eesti-prantsuse paralleelkorpus**	Antoine Chalvin	Eesti-Prantsuse Leksikograafiaühing	2011-2012
EKT18	**Kõnetuvastus**	Tanel Alumäe	TTÜ Küberneetika Instituut	2011-2014
EKT17	**Audiovisuaalse kõnesünteesi prototüüp**	Einar Meister	TTÜ Küberneetika Instituut	2011-2014
EKT20	**Kõnesünteesiliidesed**	Meelis Mihkla	Eesti Keele Instituut	2011-2014
EKT11	**Uued ressursid masintõlkes**	Heiki-Jaan Kaalep	TÜ, Matemaatika-informaatikateaduskond	2011-2013
EKT12	**Semantika vahendid eesti keelele**	Neeme Kahusk	TÜ, Matemaatika-informaatikateaduskond	2011-2014
EKT6	**Autentse meditsiinikeele korpuse alusel radioloogia elektroonse piltsõnastiku koostamine**	Eola Valdre	TÜ, Filosoofiateaduskond	2011-2014
EKT35	**Eesti avatud paralleelkorpus**	Margit Kurm	Tilde Eesti OÜ	2012-2014
EKT37	**Subtiitrite helindamise ja teleeetrisse edastamise tarkvaralahendus**	Meelis Mihkla	Eesti Keele Instituut	2012-2013
EKT5	**Eestikeelse dialoogi pragmaatika analüsaator**	Mare Koit	TÜ, Matemaatika-informaatikateaduskond	2011-2013
EKT38	**Eestikeelsete dialoogsüsteemide loomise raamistik**	Margus Treumuth	TÜ, Matemaatika-informaatikateaduskond	2012-2014
EKT22	**Mallipõhine faktituletus tekstikorpustest**	Sven Laur	TÜ, Matemaatika-informaatikateaduskond	2011-2013
EKT7	**Vahendid teksti mitmekihiliseks märgendamiseks (rakendatuna Koondkorpusele)**	Kadri Muischnek	TÜ, Matemaatika-informaatikateaduskond	2011-2014
EKT24	**Leksikograafi töökeskkonna modifitseerimine**	Arvi Tavast	Eesti Keele Instituut	2011-2012
EKT2	**Eesti Wordnet'i täiendamine**	Heili Orav	TÜ, Filosoofiateaduskond	2011-2014
EKT1	**Kõne ja teksti emotsionaalsuse statistilised mudelid**	Hille Pajupuu	Eesti Keele Instituut	2011-2014
EKT3	**Kõne- ja multi-modaalsed korpused**	Einar Meister	TTÜ Küberneetika Instituut	2011-2014
EKT8	**Suulise eesti keele audiovisuaalse suhtluskorpuse kogumine ja päringusüsteemi arendamine**	Tiit Hennoste	TÜ, Filosoofiateaduskond	2011-2014

Projekti eesmärgid ja metoodika

Keskuse (EKRK, vt ka http://keeleressursid.ee/) tegevuse eesmärgiks on luua taristu, mis teeb eestikeelsed keeleressursid ja keeletehnoloogilise tarkvara (sõnastikud, teksti- ja kõnekorpused, keeleandmebaasid, keeletarkvara) eelkõige teadlastest huvilistele kättesaadavaks. Keskuses pakutakse keeletehnoloogia vahendeid kui veebiteenust, mis kasutab arhiveeritud andmeid. Seeläbi tehakse olemasolevate ressursside kasutamine ja kombineerimine uute väärtuste loomisel lihtsamaks, samuti tagatakse loodud ressursside säilimine.

Keskusele on pandud kohustus arhiveerida ja teha kättesaadavaks ka riiklike programmide EKKTT ja EKT projektide tulemused.

Et tagada keeleressursside pikemaajaline kasutusvõimalus, võimalus eri ressursside kombineerida, võrrelda ja kasutada koos erinevate Eesti-siseste või ka välismaiste rakendustega, selleks viiakse keskusse kogutavad ressursid vastavusse üldlevinud standarditega, dokumenteeritakse ning tehakse nende metaandmed registris ja repositooriumis kättesaadavaks ja automaatselt töödeldavaks.

Keskuses töötatakse välja litsentsilepingud ja viiakse sisse autentimissüsteem, et lubada ressursside kasutust võimalikult lihtsalt, järgides kasutuslepingute tingimusi ja kaitstes võimaluste piirides ressursiomanike huve. Autentimissüsteemi haaratakse esmajärjekorras teadus- ja arendusasutused, sõlmitakse lepingud vastastikuseks juurdepääsuks välismaiste riiklike akadeemiliste identiteedipakkujate liitudega (Identity provider federation), lisaks võimaldatakse kasutajakontod ning juurdepääs vabalt kasutatavatele ressurssidele ka mitteteadlastest kasutajatele. Keeleressursid tehakse eri liiki litsentsilepingute ja eraldi kokkulepetega kättesaadavaks ka avalikule ja erasektorile.

Saavutatud tulemused

Riiklike programmide projektide tulemusena või muul toel loodud TÜ, TTÜ KübI, EKI ja Filosofti keeleressursside kohta on andmed koondatud ühtse registrimalli alusel, mida hakatakse kasutama Keskuse registris. Tartu Ülikooli META-SHARE repositooriumivõrgustiku sõlmes http://metashare.ut.ee/ on hetkel 21 eesti keelt sisaldava ressursi põhjalikud meta-andmed, lisaks on sama portaali kaudu võimalik ligi pääseda enam kui 200 keeleressursile projekti META-NORD partnerite juures. Katsetamisjärgus on kasutajate ligipääs Eesti haridus- ja teadusasutustevahelise autentimise ja autoriseerimise taristu TAAT (http://taat.edu.ee/) kaudu.

Keeleressursside litsentsimine on osutunud eeldatust keerukamaks ülesandeks tihedate seoste tõttu autoriõiguse ja intellektuaalomandi digiajastule jalgu jäävate regulatsioonidega. Eesti parimaid valdkonna õiguseksperte kaasates on lootust lähitulevikus sõnastada toimivad tüüplitsentsid Creative Commons baasil ja mõjutada samas ka valdkondlike õigusaktide muutusi nii Eestis kui Euroopa Liidus.

Tartu Ülikooli, Eesti Keele Instituudi ja Tallinna Tehnikaülikooli Küberneetika Instituudi ühisettepaneku põhjal arvati EKRK „Eesti teaduse infrastruktuuride teekaardi“ oluliseks objektiks ning kolme asutuse konsortsiumina saadi rahastust ka EL tõukefondide „Riikliku tähtsusega teaduse infrastruktuuri kaasajastamine“ investeeringute kava alameetmest. Konsortsiumina peab EKRK täitma ka CLARIN ERIC (*Common Language Resources and Technology Infrastructure European Research Infrastructure Consortium,* vt ka www.clarin.eu) Eesti riikliku keskuse kohustusi.

VÕRU JA SETO KEELEKORPUS

Projekti eesmärgiks on ette valmistada võru ja seto keelele keeletehnoloogilise toe loomist läbi võru ja seto keeleressursside kogumise ja korraldamise ühtseks keelekorpuseks.

Võru ja seto keele arendamist ja laialdasemat kasutust on peetud tähtsaks nii kohalikul kui riiklikul tasandil. 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.

Keeletehnoloogiline esmavajadus võru ja seto keele puhul oleks võru ja seto keele nii kirjaliku kui suulise korpuse loomine ja selle põhjal võru-seto automaatkorrektuuri, -poolitaja jt vajalike rakenduste loomine. Seejuures on esimesteks töödeks võru ja seto keeleressursside koondamine, korrastamine ja täiendamine ühtseks keelekorpuseks, vajalike otsingumootorite jm kasutajaliideste loomine ning edaspidi loodud korpuse täiendamine ja laiendamine ning uute rakenduste lisamine.

Korpuse kirjaliku keele poolde on plaanitud võru ja seto ajakirjanduskeele osa (ajalehtede Uma Leht ja Setomaa elektrooniliste arhiivide sisu põhjal) ja Võru Instituudis jm säilitatavate muude võru kirjakeele allikate osa (õpikud, ilukirjandusväljaanded jm).

Korpuse suulise keele ossa kogutakse ühistöös TÜ murdekorpuse ja suulise kõne korpuse arendajate ning TÜ Lõuna-Eesti keele- ja kultuuriuuringute keskusega kokku nii murdekorpuses juba olemasolevaid tekste kui ka päris uusi jäädvustusi nii heli- ja videofailide kui litereeringutena. Korpusele liidetakse ka olemasolevad võru lastekeele salvestised.

Nii kirjalike kui suuliste keeleressursside osas tuleb lisaks olemasoleva materjali koondamisele ja korrastamisele korpust 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.

Korpuse eesti- ja võrukeelne koduleht koos lastekeele videonäite ja korpuse võru kirjakeele ossa kogutud Uma Lehe artiklitega asub Võru Instituudi kodulehe juures aadressil:

eesti k http://wi.ee/index.php/keelekorpus-et

võru k http://wi.ee/index.php/keelekorpus-vro

EESTI KEELE SPONTAANSE KÕNE FONEETILISE KORPUSE ARENDUSED

Eesmärgid ja metoodika

Antud projekti eesmärgiks on arendada nii eesti spontaankõne foneetilist korpust kui korpuse otsingusüsteemi. Korpust saab kasutada keeletarkvara väljatöötamiseks, kõnetuvastuse ja kõnesünteesi arendamiseks. Projekt on loogiliseks 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 kasvatada korpuse salvestuste maht vähemalt 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 varasemate 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 kättesaadav kõigile Internetis: http://www.murre.ut.ee/foneetikakorpus/ .

Põhitulemused

Nii sõna- ja häälikutasandil on märgendatud 9 tundi kõnet ehk 224 045 segmenti (93 506 segmenti sõnatasandil ja 130 539 segmenti häälikutasandil). Koostöös Tanel Alumäega on katsetatud automaatset kõnetuvastust, et kiirendada märgendamist ja saada sõnatasandist ülevaade enne käsitsi märgendamist. Märgendajate seniste kogemuste põhjal võib öelda, et automaatne tuvastus eksib spontaansete dialoogide puhul rohkem, aga on täpsem monoloogide puhul, hõlbustades ja kiirendades eelkõige viimaste märgendamist.

Lisaks sõna ja häälikutasandi märgendamisele on jätkatud varem sõna- ja häälikutasandil märgendatud failide märgendamist silbi- ja taktitasandil, millega seoses on kontrollitud ja ühtlustatud kahe esimese tasandi märgendamist. Muudel kui sõna- ja häälikutasandil märgendatud kihtidel on märgendatud 140 280 segmenti.

Praegu on spontaanse kõne foneetilises korpuses sõna- ja häälikutasandil märgendatuna kokku 37 tundi kõnet (sõnatasandil 273 514 segmenti ja häälikutasandil 660 105 segmenti), millest 70% on märgendatuna ka silbitasandil ning 45% taktitasandil, mis teeb korpuse kogumahuks 1 365 298 segmenti.

Uusi salvestusi on tehtud nii välitöödel kui stuudios, kusjuures peamiselt on salvestatud dialooge. Hetkel on tehtud salvestusi kokku 17 tundi ja 20 minutit, kusjuures käesoleva aasta sügisel kestendutaksegi uute salvestuste tegemisele.

Koostöös Heiki-Jaan Kaalepiga on kogu korpus morfoloogiliselt märgendatud, kasutades Filosofti morfanalüsaatorit. Tegu on täisautomaatse märgendusega, mis tehnilistel põhjustel on ühestamata. Ühestamatusest hoolimata on lisatud morfoloogilise info kiht ka veebiotsingusse.

Projekti eesmärgid ja metoodika

Projekti (2011–2012) algsed eesmärgid olid:

1) viia lõpule eesti-prantsuse paralleelkorpus ja täiendada korpuse veebiliidest (http://corpus.estfra.ee);
2) liita korpus masintõlkesüsteemiga;
3) uurida võimalusi korpuse kasutamiseks keeleõpperakendustes.

Programmi raames saadud vahenditest osutus võimalikuks rahastada ainult esimest eesmärki.
Korpus on joondatud lausete tasandil. Kirjandus- ja humanitaaralaste tekstide joondamiseks kasutasime joondamisprogrammi Hunalign. Automaatse joondamise kvaliteedi parandamiseks koostasime eesti-prantsuse elektroonilise abisõnastiku, mida täiendasime Estmorfi abil automaatselt genereeritud muutevormidega. Käsitsi on parandatud automaatjoondamise vead. Euroopa Parlamendi arutelude tekstid joondasime täisautomaatselt Gargantua-nimelise joondamisprogrammiga.
Kõik tekstid on morfoloogiliselt märgendatud ja ühestatud. Märgendamisel on kasutatud järgmist tarkvara: Estmorf eesti keele morfoloogiliseks analüüsiks, TreeTagger prantsuse keele morfoloogiliseks analüüsiks.
Tekstid on varustatud bibliograafiliste viidetega ja statistikaga tekstide mahu kohta.

Saavutatud ja oodatavad tulemused

Aastatel 2011 ja 2012 oleme korpust täiendanud 59 miljoni sõne võrra.
Praeguse seisuga sisaldab korpus 61 miljonit sõnet, mis jagunevad järgmiselt :

1) eesti ilukirjandus (3 miljonit)
2) prantsuse ilukirjandus (1,2 miljonit),
3) eesti humanitaaralased tekstid (132 000),
4) prantsuse humanitaaralased tekstid (715 000),
5) Euroopa Liidu seadusandlus (26,3 miljonit)
6) Euroopa Parlamendi istungid (28,2 miljonit)
7) Piibel (1,4 miljonit)

Lähinädalatel lisame veel 4 miljonit sõnet ilukirjanduslikke või humanitaaralaseid tekste, nii et aasta lõpuks peaks sõnede arv kasvama 65 miljonini, mis jääb korpuse lõplikuks mahuks.
2012. a. lõpuks täiendame kasutajaliidese järgmiste uuendustega:

1. Iga eesti lemma/sõnavormi otsingu puhul pakub kasutajaliides ülevaadet selle lemma/sõnavormi kollokatsioonidest.
2. Iga eesti või prantsuse lemma otsingu puhul pakub kasutajaliides nimekirja kõige tõenäolisematest tõlkevastete kandidaatidest.
3. Kasutajal on võimalik näha iga tekstilõigu eelnevaid ja järgnevaid lõike (maksimaalselt 2 lõiku) nii prantsuse kui ka eesti keeles.

Eesmärgid

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;
- parem kvaliteet spontaanse ja aktsendiga kõne puhul;
- mitmesuguste nimede parem tuvastus.

Pikkade kõnesalvestuste transkribeerimise osas on selle projekti eesmärgiks vähendada vigade arvu umbes 25% võrreldes 2010. a tasemega.

Lisaks eelnevale on kavas tegeleda kõnetuvastuse väljundi struktureerimise meetoditega, mis võimaldaksid kõnetuvatuse väljundi "kirjavahemärgistamist", nimega üksuste identifitseerimist ning automaatset teemadeks segmenteerimist.

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

Programmi raames loodavate rakenduste osas on plaanis tähelepanu pöörata järjest populaarsemaks saavate nutitelefonide rakendustele.

Tulemused

Projekti raames on valminud neli kõnetuvastustehnoloogial põhinevat rakendust Android nutitelefoniplatvormile. Rakendus "Kõnele" lubab eestikeelse kõne abil sisestada teksti kõikides Androidi rakendustes. Rakendus "Arvutaja" kasutab samuti eestikeelset kõnetuvastust, kuid oskab ka kasutaja poolt öeldule arukalt reageerida: selle abil saab teha matemaatilisi tehteid, teha ühikuteisendusi, otsida Eesti kohanimesid jms. Rakenduse "Diktofon" abil saab salvestada pikki kõnelõike (näiteks intervjuud) ning neid automaatselt tekstiks teisendada. Rakendus „Inimesed" lubab otsida kõne abil telefoni kontaktide andmebaasist.

Projekti raames implementeeritava tuvastustehnoloogia kvaliteedi paranemist illustreerib allolev tabel, kus on toodud eri tüüpi kõnesalvestuste tuvastusvigade osakaalu vähenemise progress aastate lõikes. Võrdluseks on toodud ka projekti eesmärkides seatud kvaliteedisihid. Telefoniintervjuude puhul on seatud eesmärk juba saavutatud.

Kõne tüüp	2010	2011	2012	2014 (eesmärk)
Vestlussaated	28,6	27,1	25,6	21,5
Konverentsiettekanded	37,1	33,9	33,0	28,0
Telefoniintervjuud	-	29,1	26,6	35,0

Lingid projekti raames loodud tarkvarale on toodud projekti kodulehel.

AUDIOVISUAALSE KÕNESÜNTEESI PROTOTÜÜP

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

Saavutatud ja oodatavad tulemused

1. Valdkonna taustauuringud ja parameetrilise mudeli valik

2012 seis: On tutvutud mitmete audiovisuaalse kõnesünteesi mudelitega ja valitud edasiseks arenduseks kaks mudelit – MASSY ja LUCIA, mis algselt on loodud saksa ja itaalia keele AV- sünteesiks. On uuritud ka erinevaid tarkvarapakette uue parameetrilise mudeli loomiseks, nt Render, MeshLab, FaceGen Modeller.

2. Töövahendite valik ja arendus

2012 seis: Audiovisuaalsete salvestuste töötluseks ja märgendamiseks on valitud tarkvarapakett ELAN (http://www.lat-mpi.eu/tools/elan/), 3D artikluatsiooniandmete analüüsiks, segmenteerimiseks ja animeerimiseks on valitud programm VisArtico (http://visartico.loria.fr/); 3D andmestiku esmaseks töötluseks on loodud tarkvarapakett MotionAnalyzer.

3. Multimodaalse andmestiku töötlus ja analüüs

2012 seis: Video- ja 3D salvestustest on mõõdetud huulteartikulatsiooni kirjeldavad tunnused eri viseemide korral (huulte avatus, suu laius, alahuule asend), nende alusel on koostatud eesti viseemide numbriline kirjeldus MASSY mudeli jaoks ja tehtud viseemide klasteranalüüs.

4. AV-sünteesi prototüübi loomine

2012 seis: Koostöös MASSY- mudeli autoriga on loodud eestikeelse AV-sünteesi esmane prototüüp, milles eestikeelne difoonsüntesaator on liidestatud MASSY-mudeliga. Uuritakse LUCIA-mudeli kohandamist eesti keelele.

5. Tajueksperimendid

2012 seis: MASSY mudeli abil on sünteesitud audiovisuaalsed stiimulid ja ettevalmistamisel on tajueksperimendid viseemide ja nende kombinatsioonide loomulikkuse hindamiseks.

KÕNESÜNTEESILIIDESED

Projekti eesmärgid

Projekti üheks eesmärgiks on 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).

Teiseks projekti eesmärgiks on erinevate kõnesünteesi rakenduste loomine: veebisõnastike helindamisliidesed, heliraamatute genereerimine, subtiitrite helindamine, nutitelefoni rakendused jms

Tulemused (2011-2012)

1. Eestikeelne formantsüntees realiseeriti avatud koodiga kõnesünteesi arendussüsteemis eSpeak. Lisaks kompaktsusele ja mitmekeelsusele tagab eSpeaki arendussüsteem märgenduskeelte toe ja kõnesüntesaatorite töö eri platvormidel.

2. Aasta lõpus valmib HTS-mootori ja eestikeelsete sünteeshäälte sobitusliides Sapi 5-le, mis võimaldab kasutada HTS-et sünteeshääli Windowsi platvormil.

3. Korpuspõhiste häälte arendustööd on keskendunud lingvistilise töötluse moodulite täiendamisele ja programmeerimistöödele. Teisel aastal loodi lapse sünteeshääl Luukas. Korpuspõhiseid hääli saab kuulata ja võrrelda aadressil http://heli.eki.ee/syntees/.

4. Esimese aasta pilootprojekti katsetustest subtiitrite helindamiseks nägemispuudega inimestele käivitus teisel aastal koostöös Eesti Rahvusringhäälingu ja Eesti Pimedate Liiduga integreeritud tarkvaraprojekt (vt EKT37 „Subtiitrite helindamise ja tele-eetrisse edastamise tarkvaralahendus“).

5. Kuna subtiitrite projekti tarkvaralahenduse mooduleid saab kastutada ka heliraamatute genereerimiseks, siis paralleelselt on alustatud teksti ettevalmistusmooduli loomist, mis on mõeldud erinevate tekstiformaatide mittevajalikust infost (pildid, lingid jms) puhastamiseks, lühendite, numbrite, erimärkide ja võõrnimede lahti kirjutamiseks.

6. Android- operatsioonisüsteemi kasutavatele mobiiltelefonidele loodi rakendus, mis loeb sünteeshäälega ette uudiseid. Kasutajal on võimalus valida kolme sünteeshääle ja kolme lugemiskiiruse vahel. Vt http://heli.eki.ee/uudistelugeja/

7. Loodi veebisõnastike helindamisliides, mida katsetati põhisõnavara sõnastiku märksõnade, muutevormide ja näitelausete helindamiseks.

Plaan

1. Korjata uusi paralleelkorpusi, mis kajastavad loomulikumat keelt (nt. tõlgitud subtiitrid).

2. Katsetada, kas uute korpuste peal treenimine tõstab tõlkekvaliteeti.

Täitmine

1. Subtiitrite korjamisest loobuti, sest selle töö tegi ära Jörg Tiedemann, kelle poolt 2011 korjatud OpenSubtitles v. 2 (54 keelt, 1,4 miljonit faili, 8,3 miljardit sõna) sisaldab eesti-inglise paralleelseid subtiitreid mahus 4800 faili, vastavalt 25,8/30,8 miljonit sõna.

Omapoolse korpuse korjamise asemel kontrolliti OpenSubtitles v.2 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. Tulemused on väljas OPUSe kodulehel.

2. Katsete tegemiseks kombineeriti erinevas vahekorras alljärgnevaid korpusi (sõnu eesti/inglise keeles, miljonites): EU Journal (41,6/55,8), OpenSubtitles v.2 (25,8/30,8), JRC Acquis (18,4/25), DGT-TM (17,7/21,1), Europarl (12,9/17,3), EMEA (9,6/11,1), ECB (2,1/2,8), KDE4 (1,6/1,9), Maarika Traadi korjatud korpused (1,2/1,5).

Eestikeelne Europarl ning OpenSubtitles v2 on palju lähedamad loomulikule keelele kui kõik eelmised korpused. Need mõlemad uued korpused isegi eraldi võttes annavad parema kvaliteedi kui varasemate aastate korpused, aga kõik korpused kokku segatult annavad veelgi paremaid tulemusi: BLEU 24,8. Google'ga võrreldes (BLEU 25,7) jäädakse veel natuke taha, aga vahe pole enam kuigi suur.

Eesmärk

Projekti eesmärgiks on luua ressursse ja vahendeid eesti keele semantika tarbeks. Projekti käigus luuakse:

- ühestatud sõnatähendusega korpus (500 000 sõna)
- Freimileksikon
- Verbide valentsileksikon
- EKSSist tuletatud WordNeti tüüpi andmebaas
- Sõnatähenduste ühestaja ja teisi vahendeid, mis aitavad loodud ressursse kasutada

Tulemused

Projekti tulemusena on aastatel 2011 - 2012 tehtud:

- ühestatud sõnatähendustega korpus 500 000 sõna
- Freimileksikon 2461 leksikaalset üksust
- Automaatne sõnatähenduste ühestamine:
    - Reeglipõhine: On leitud üle 90 reegli, mille alusel saab tuvastada sõna tähendust
    - Automaatne: On katsetatud masinõppesüsteemi, mis on saavutanud kuni 80% täpsuse sõnade tähenduste äratundmisel

Projekti tulemused on saadaval projekti kodulehel aadressil:

http://www.keeletehnoloogia.ee/ekt-projektid/semantika-vahendid-eesti-keelele

AUTENTSE MEDITSIINIKEELE KORPUSE ALUSEL RADIOLOOGIA ELEKTROONSE PILTSÕNASTIKU KOOSTAMINE

Eesmärk ja tähtsus

Koostada tegelikku, s.o töist, keelekasutust kajastav radioloogia elektroonne piltsõnastik, mis hõlbustaks terviseandmete automaatset analüüsi, radioloogiaõpet, vabatekstipõhiseid päringuid ning haiglainfosüsteemi ja/või tõlkerakenduste loomist ja kaasajastamist.

Kliinilises töös kasutatav keel on võrreldes akadeemilisega palju lakoonilisem, lühenditerohkem, vahel kirjakeele norme eirav ja sageli ebahariliku lauseehitusega. Terviseandmed registreeritakse kas struktuurselt või vabatekstina. Viimase puhul on andmete automaatne töötlus ja/või analüüs tüsilik. Uurimata sõnavara ja standardimata keelekasutuse tõttu on vabatekstist vajalikku teavet raske leida. Seetõttu on oluline teada, mis terminitega (sh lühendid) mõistet kirjeldatakse ja milline on nende kasutussagedus.

Materjal

1. AS Ida-Tallinna Keskhaiglas aastatel 2009–2011 tehtud isikustamata radioloogilised kirjeldused on koondatud korpuseks (11,8 miljonit sõnet). Korpuses on esitatud vabatekstilised andmed (kliinilised andmed, leid ja kokkuvõte) ning uuringu identifikaator.

2. Võrdlusmaterjaliks on arstiteaduskonna radioloogialoengud (ARHO.01.033 „Radioloogia (uurimismeetodid, radio-anatoomia, kliinilise radioloogia algkursus)“ ja ARHO.02.009 „Kliiniline radioloogia“).

Tulemused

1. Koostatud on autentsete radioloogiliste kirjelduste korpus (http://www.cl.ut.ee/korpused/medkorpus/). Korpus on valideeritud ja XML-märgendatud. Korpuses on 207 534 teksti (11,8 miljonit sõnet). Materjal jaguneb järgmiselt: röntgenuuringud (139 998 uuringut, 4 663 958 sõnet), ultraheliuuringud (34 020 uuringut, 2 970 399 sõnet), kompuuteruuringud (20 725 uuringut, 2 751 990 sõnet), magnetuuringud (11 037 uuringut, 1 293 070 sõnet), stsintigraafiauuringud (1754 uuringut, 185 939 sõnet).

2. Radioloogialoengutest on koostamisel võrdlusmaterjali korpus (praegu 72 000 sõnet), failide esitusvorming (ppt) on muudetud tekstivorminguks (txt), tekstid on liigenduse säilitamiseks märgendatud.

3. Koostatud on anatoomilist normivarianti esitavate röntgenpiltide kogu.

4. Alustatud on lühendite ja lühendamisviiside analüüsi.

Koostööpartnerid: TÜ Eesti ja üldkeeleteaduse instituut, AS Ida-Tallinna Keskhaigla, SA TÜ Kliinikumi radioloogiakliinik

Eetikaluba: Tallinna Meditsiiniuuringute Eetikakomitee luba nr 2169

EESTI AVATUD PARALLEELKORPUS

Projekti 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.

Projekti mõõdetav tulem on kogutud ja korrastatud paralleelkorpuste maht. Projekti esimese aasta jooksul kogutud vähemalt 2,5 miljonit ühikut (sõna), projekti lõpuks vähemalt 15 miljonit ühikut.

Projekti tegevused
Tegevus 1. Paralleelkorpuse materjali kogumine.
Domeenis .ee on seisuga august 2012 registreeritud ca 66 000 alamdomeeni millest ca 50 000 on kättesaadavad ning millest omakorda hinnanguliselt 4000–7000 veebisaiti on ingliskeelse sisuga, mis on kasutatav paralleelkorpuse loomisel.
Tegevus 2. Sisendtekstiformaatide teisendamine formaatimata tekstiks.
Tegevus 3. Joondamine (alignment).
Tegevus 4. Tulemuste levitamine.
Kogutud korpus läbib kvaliteedikontrollid ning tehakse seejärel kättesaadavaks masintõlkesüsteemidele.

Tulevikuvisioon
Projekti teisel ja kolmandal aastal kaalume paralleelkorpuste kogumiseks ja töötlemiseks *crowdsourcing*'u kasutamist Teeme Ära! näitel.

Projekti täitjast
Tilde on Euroopa juhtiv keeletehnoloogia teadus- ja arendustegevuse ning väiksematesse keeltesse lokaliseerimisega tegelev ettevõte. Eestis tegutseme aastast 2000.
Tilde põhitegevusalad on:

(i) keeletehnoloogia, sh masintõlge http://www.letsmt.eu/ , õigekeelsustööriistad, kõnetehnoloogia, mitmekeelsed fondid, terminiandmebaasid;
(ii) tarkvara lokaliseerimine ja tõlketeenused.

keeletehnoloogia@tilde.ee

SUBTIITRITE HELINDAMISE JA TELE-EETRISSE EDASTAMISE TARKVARALAHENDUS

Projekti eesmärk

Subtiitrite helindamise ning tele-eetrisse edastamise tarkvaralahenduse eesmärk on televisioonis kasutatavate subtiitrifailide alusel kõnesüntesaatoriga helifailide genereerimine ning eraldi helikanalis digitelevisiooni eetrisse edastamine. Ühisprojektis osalevad kolm asutust: Eesti Keele Instituut (EKI), Eesti Rahvusringhääling (ERR) ja Eesti Pimedate Liit (EPL)

Subtiitrite helindamise tarkvaralahendus koosneb kahest loogilisest komponendist:

1) subtiitrite helindamise toimetamise tarkvara, mille sisendiks on STL vormingus subtiitrifail ning väljundiks ajakooditäpne helifail (WAV või MP3);

2) subtiitri- ja helifailide haldamise, eetrisse planeerimise ning eetrisse edastamise tarkvara.

Tarkvaralahenduse esimese komponendi arendab EKI ning teise komponendi ERR. Koostöös juurutatakse reaalselt töötav lahendus, võttes arvesse subtiitrite helindamise teenuse sihtgrupi vajadusi. Peamiseks sihtgrupiks on nägemispuude või tavakirjas teksti lugemist takistava puudega inimesed. Kuid lisaks neile on teenuse võimalikud kasutajad ka eakad, lapsed ja näiteks eesti keele õppijad. EPL osaleb projektis eelkõige nägemispuudega inimeste vajaduste vahendajana ning erinevate lahendusvariantide testijana.

ERR-i telekanalite subtiitrite helindamise tarkvara on tasuta kasutatav ka teistele telestuudiotele. Tarkvaralahendus (lähtekood, sünteeshääled) tehakse teistele kastutatavaks Eesti Keeleressursside Keskuse vahendusel.

Saavutatud ja oodatavad tulemused

1. Valminud on subtiitrite helindamise tarkvaralahenduse funktsionaalsuse ning toimimise töö- ja andmevoode üksikasjalik kirjeldus .
2. Loodud on binaarse subtiitrite faili STL automaatne teisendaja ajakoodidega tekstifailiks ja tekstifaili põhjal subtiitritega sünkroonse helifaili genereerija.
3. Proovifilmidele ja –saadetele on genereeritud subtiitrite alusel erinevaid sünteeshääli, mis on miksitud programmiheliga videofailidesse.
4. Käimas on proovifilmide testid, et leida sobivaim sünteeshääl ja lugemiskiirus. Aasta lõpus tahetakse esimesi helindatud subtiitritega saateid testida tele-eetris.
5. Loomisel on tekstianalüsaator võõrnimede ja lühendite tuvastamiseks. EKI-is ja ERR-is on kokku lepitud võõrnimede transkriptsioonireeglites. Võõrnimede häädusbaasi haldamiseks on loomisel toimetaja liides.

EESTIKEELSE DIALOOGI PRAGMAATIKA ANALÜSAATOR

Dialoogi pragmaatiline analüüs leiab rakendamist kasutajaga eesti keeles suhtlevates dialoogsüsteemides, aga samuti lingvisti töövahendina dialoogi uurimisel.

Eesmärgid ja metoodika

Projekti käigus kavandatakse järgmiste pragmaatilise analüüsi osaülesannete lahendamine:

1. teadmuse automaatne ekstraheerimine eestikeelsest tekstist (dialoogist),
2. dialoogiaktide automaatne tuvastamine,
3. dialoogi struktuuri automaatne analüüs,
4. dialoogistrateegiate automaatne analüüs.

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

- Varasemas projektis töötati välja andmebaasides olevat infot vahendava intelligentse kasutajaliidese kontseptsioon ja valmis seda realiseeriv programm – asünkroonsete dialoogsüsteemide raamistik. Teadmuse ekstraheerimine on osa raamistiku abil uue dialoogsüsteemi loomisest. Teadmus ekstraheeritakse küsimus-vastuskomplektidest ja esitatakse regulaaravaldistena raamistiku jaoks sobival kujul.
- Dialoogiaktide automaatsel tuvastamisel TÜ tüpoloogia kohaselt võetakse aluseks Bayesi klassifitseerimismeetod. Programmi treenitakse ja testitakse Eesti dialoogikorpusel.
- Dialoogi struktuuri analüüs viiakse läbi kindlaksmääratud ainevaldkonnas. Märgendatakse alamdialoogid, tuginedes varem tuvastatud dialoogiaktidele.
- Dialoogistrateegiate analüüs keskendub dialoogis osalejate kindlat tüüpi käitumismustrite leidmisele ja nende automaatsele märgendamisele, kasutades nii dialoogiakte kui ka dialoogi struktuuri.
- Eesti dialoogikorpust täiendatakse erinevat liiki dialoogidega, märgendatakse dialoogiaktid.

Tulemused

- Teadmuse ekstraheerimine eestikeelsest tekstist.
- Programm, mis koostab regulaaravaldisi KKK rubriikide alusel (Raul Sirel). Kasutusel dialoogsüsteemide raamistikus.
- Dialoogiaktide tuvastamine.
- Dialoogiaktide poolautomaatse märgendamise veebipõhine tarkvara (http://ats.cs.ut.ee/darec/www1/ , Sven Aller). Treenitud infotelefonidialoogide alamkorpusel ja testitud ka kauplusedialoogidel.
- Dialoogikorpuse laiendamine: argivestlused, MNS-vestlused, võlur Ozi dialoogid.

Tulevane töö

- Dialoogi struktuuri tuvastamine.
- Dialoogistrateegiate tuvastamine.

Projekti eesmärgid

Varasemate projektide käigus on loodud eestikeelsete dialoogsüsteemide raamistik, mille abil saab luua dialoogsüsteeme kitsas ainevaldkonnas.

Käesoleva projekti eesmärk on pakkuda laiem juurdepääs dialoogsüsteemide loomisele. Selleks luuakse raamistikule administreerimisliides.

Projekti tulemusel tekib mugav võimalus kasutada raamistiku põhifunktsionaalsust, mis on ennast õigustanud mitmes ainevaldkonnas. Kasutaja saab seadistada omanäolise dialoogsüsteemi, kasutada inim-abi liidest teadmusbaasi kogumiseks, kasutada teadmusbaasi täiendamiseks administreerimisliidest, vaadata vestluslogisid. Valminud dialoogsüsteemi saab kasutaja integreerida oma asutuse veebilehega.

Saavutatud ja oodatavad tulemused

Projekt alles algas. Projekti tulemusel tekib veebipõhine keskkond, kus arendaja saab luua endale dialoogsüsteemi. Seni oli see võimalik vaid raamistiku autori kaasabil ning ka siis ei olnud võimalik loodud dialoogsüsteemi teadmusbaasi hilisem täiendamine. Tekib võimalus teha ka väljavõtteid vestluslogidest.

Projekti tulemusel saab olema kättesaadav dialoogsüsteemide põhifunktsionaalsus tulevastele arendajatele.

Olemasolev põhifunktsionaalsus on orienteeritud eesti keelele ja sisaldab järgnevaid komponente:

- kasutaja sisendis automaatne õigekirjavigade parandamine teadmusbaasi reeglite suhtes;
- inim-abi liides teadmusbaasi kogumiseks, st dialoogsüsteemi omanik saab sekkuda vestlusesse, seejuures vestluse hilisem analüüs võimaldab teadmusbaasi täiendamist;
- vestluse algusest teavitamine SMS-i vahendusel (oluline inim-abi kasutamisel;)
- asünkroonne vestlusmudel;
- eestikeelne morf-analüüs, et lihtsustada teadmusbaasi reeglite loomist, piirdudes reeglites vaid algvormidega;
- sõnajärjestuse probleemi lahendus, et lihtsustada teadmusbaasi reeglite loomist, piirdudes reeglites vaid kindla järjestusega ning lubades teisi permutatsioone automaatselt;
- integreeritud kõnesüntees.

MALLIPÕHINE FAKTITULETUS TEKSTIKORPUSTEST

Projekti eesmärk on luua tarkvarakomponent, mis suudab vabatekstidest õppida erinevaid seoseid ning nende abil eraldada struktureeritud infot. Seosed võivad olla lihtsad nagu isikunimed ja organisatsioonid või keerulisemad nagu firmade peakontorite asukohad.

Meetod vajab sisendiks korpust, milles on meid huvitav seos märgendatud. Seejärel leitakse automaatselt sobivad mallid ja koostatakse mudel, mis antud seost võimalikult hästi tuvastaksid. Tulemusena saame märgendamata vabatekstidest leida uusi seosele vastavaid näiteid.

Põhitulemused 2011-2012

- Mallikaeve algoritm ning selle omaduste teoreetilised tõestused.
- Tarkvaraprototüüp seoste märgendamiseks, treenimiseks ning kasutamiseks. Vahendid korpuste eeltöötluseks ning sobivale kujule teisendamiseks.
- Juhtumiuuringud isikunimede, organisatsioonide ja asukohtade tuvastamiseks tekstist.
- Juhtumiuuringud näidete poolautomaatseks laiendamiseks (aktiivõpe).

Eeldatavad tulemused 2012-2013

- Faktituletuse algoritmide edasiarendus (korpuste eelklasterdamine, negatiivsed mallid) ning efektiivsuse hindamine.
- Faktituletuse komponendi kasutamine NER lahenduste täiustamiseks (valepositiivsete filtreerimine).
- Kasutajaliidese edasiarendamine ning mallide jaotuste visualiseerimine.
- Faktituletuse meetodite kasutamine meditsiiniandmete analüüsimiseks.

VAHENDID TEKSTI MITMEKIHILISEKS MÄRGENDAMISEKS (RAKENDATUNA KOONDKORPUSELE)

Eesmärk: koondada senised korpuse märgendamiseks kasutatud tarkvaraprototüübid ühtseks standardiseeritud programmide koguks ja nende abil muuta eesti keele Koondkorpus mitmetasandiliselt märgendatud korpuseks.

Tulemused

Morfoloogiline ja süntaktiline analüüs

On integreeritud kitsenduste grammatikal (CG) põhinevad morfoloogiline ühestaja ja süntaksianalüsaator. Viimane on kohandatud ka statistilise morfoloogilise ühestaja väljundile, st töötab nüüd mõlema eesti keele jaoks olemasoleva morfoloogilise ühestaja väljundiga.
CG süntaksianalüsaator on adapteeritud EKI morfanalüsaatori väljundile, st töötab nüüd mõlema eesti keele jaoks olemasoleva morfoloogiaanalüsaatori väljundiga.
Täiustatakse jätkuvalt CG sõltuvussüntaktilise analüsaatori reegleid.
2012. aasta lõpuks märgendatakse morfoloogiliselt kogu Koondkorpus ja süntaktiliselt Tasakaalus korpus (Koondkorpuse 15 miljoni sõnaline allosa).
Pidevalt toimub käsitsi sõltuvussüntaktiliselt märgendatud puudepanga loomine (praegu ca 80 000 sõna).

Praktiline semantiline analüüs

On loodud programm nime- ja numbriüksuste märgendamiseks, mis märgistab tekstis isikud, kohad, aadressid, organisatsioonid, suurtähelised lühendid e. akronüümid, telefoninumbrid, hinnad, kogused, mitut liiki registreerimisnumbrid ning ajaväljendid.
Semantiline märgendus on lisatud ka Keeleveebi (www.keeleveeb.ee) kaudu kasutatavale Koondkorpuse versioonile.

Tekstiliigi automaatse tuvastamise eeltööd

Tasakaalus korpuse põhjal on koostatud sõnavormide ja lemmade sagedusloendid allkorpuste kaupa, vt lähemalt http://www.cl.ut.ee/ressursid/sagedused1/

Koondkorpuse enda ja tema kasutusvõimaluste edasiarendamine

On täiustatud kollokatsioonide tuvastajat (www.rabauti.ee/clc), mis nüüd võimaldab otsida osalauses esinevate sõnavormide või lemmade koosesinemisi. Nii sisestava lemma või sõnavormi kui ka otsitavate kollokaatide ringi saab piirata nende sõnaliigilise kuuluvusega.

LEKSIKOGRAAFI TÖÖKESKKONNA MODIFITSEERIMINE

Projekti alguseks oli EELex töötav vabavaraline sõnastikusüsteem, milles hallati 25 sõnastikku (EKI-s ja mujal) ja mis täitis oma kasutajate põhivajadused. Nüüdseks (september 2012) on sõnastike arv kasvanud 48-ni, sh 7 oskussõnastikku. Sõnastike arvu kiiret suurendamist takistanud asjaolud on kõrvaldatud järgmisel viisil:

- Süsteem on ümber kirjutatud brauserisõltumatuks, välja arvatud harva vajaminevad funktsioonid: skeemieditor, morfoloogia lisamine ja automaatne klaviatuurivahetus. EELexi väga vanade versioonidega loodud sõnastikele saab brauserisõltumatust rakendada pärast nende korrastamist.
- EELexist on nüüd saadaval kasutaja enda serverisse installitav versioon (http://eelex.eki.ee/pub/Install/eeLex/eeLex_pakk_02juun11.tgz) ja olemas ka esimene kasutaja.
- Süsteemi dokumentatsioon on värskendatud ja lisatud EELexi kodulehele.
- Uue sõnastiku lisamine ja võimaliku olemasoleva materjali importimine ei vaja enam arendaja sekkumist.
- Mõistepõhiste struktuuride võimaldamiseks on sõnastiku lähtekeel viidud baasistruktuurist vaate parameetriks.
- Uute sõnastike baasistruktuuride ühtlustamise eesmärgil on loodud standardskeemid.

Eesti keele toe ning viidete ja päringute funktsionaalsuse osas on tehtud järgmised arendused:

- Morfoloogia andmebaasi (MAB) esialgne struktuur on valmis. Andmed lisamise järjekorras: tüvebaas, vormierandid, palatalisatsioon ja sõnamoodustuse andmed.
- Päringuvõimaluste laiendus EELexis: sama märksõna teistes sõnastikes, märksõna ühendid, link Google'i või teiste sõnastike veebiversioonide päringusse.
- Komplekspäring (kompromiss x-path'i ja mugava kasutajaliidese vahel) võimaldab teha keerukamaid päringuid: korraga mitut tunnust, otsingupiirkonna valik jm.

Taotluses lubatule lisaks on tehtud järgmised arendused:

- xmlStats on EELexi kõrvale loodud (pea)toimetaja tööriist sõnastiku XML faili analüüsimiseks ja kontrollimiseks ning skeemi otstarbekuse kohta järelduste tegemiseks.
- Osasõnastiku loomise vahend. Osasõnastiku genereerimine on mõeldud selleks, et võtta olemasolev rikkaliku infoga sõnastik taaskasutusse, luues temast mitmeid uusi, kitsama teemaga sõnastikke.

EESTI WORDNET'I TÄIENDAMINE

Tartu Ülikoolis on alates 1998. aastast koostamisel uuema põlvkonna arvutitesaurus – Eesti Wordnet. See on leksikaal-semantiline andmebaas, kus tähendusüksused seostatakse omavahel 45 erineva semantilise seose, nagu alam-/ülemmõisted, antonüümia, osa-terviku, põhjuslikkuse jm suhte, kaudu. Mõistetele on lisatud ka nende ingliskeelsed vasted.

Käesoleva projekti eesmärgiks on arvutitesauruse suurendamine, täiendamine ja olemasoleva kontrollimine ning parandamine. Projekt kestab neli aastat – 2011-2014.a..

Praeguseks on projekti kahe aasta jooksul lisandunud tesaurusesse u 16 500 uut mõistet - Eesti Wordnetis on (seisuga september 2012) üle 57 500 mõiste (projekti alguses 2011.a. oli u 40 700 mõistet). Töö tesauruse täiendamisel on kulgenud mitmeti. Esiteks oleme täiendanud tesaurust kitsaste valdkondade kaudu (nt meditsiin, filosoofia jms). Teiseks lisanud uusi või puuduvaid tähendusi koondkorpuse sagedustele tuginedes. Ja kolmandaks on toimunud andmebaasi täiendamine sõnatähenduste ühestamise andmete põhjal.

Eesti keele tesauruse lehitsemiseks töötavad lingid http://www.cl.ut.ee/ressursid/teksaurus/ või www.keeleveeb.ee.

Eesmärk: automaatselt ära tunda emotsioon kõnes ja kirjas. Realiseeritakse kahe prototüübina: 1) veebipõhine kirjaliku teksti emotsioonituvastaja; 2) kõnelejakohane emotsioonituvastaja.

1. Veebipõhine kirjaliku teksti emotsioonituvastaja

Lähtekoht: Sõnakasutus mõjutab lugeja meeleolu ja käitumist enam kui teksti sisu (vt nt Keysar jt 2012). Emotsioonituvastaja hindab teksti emotsionaalsust tema sõnalise koostise põhjal. Teksti positiivsus, negatiivsus, vastuolulisus või neutraalsus tekitab potentsiaalselt samu tundeid ka lugejas. Emotsioonituvastus toimub ortograafiliste lõikude kaupa, sest lõigu piires meeleolu tavaliselt ei muutu (Pajupuu jt 2011). Tuvastusmeetod: leksikonipõhise ja statistilise hübriid.

Loodud on emotsioonisõnade leksikon: 1019 sõna, neist 413 positiivset, 606 negatiivset, enamik on sagedased sõnad. On suur tõenäosus, et mõni neist emotsionaalses lõigus esineb ning võimaldab lõigule emotsioonihinnangut anda. Leksikon võtab arvesse polüseemiat ning seda, et sõna eri vormidel võib olla erinev valents. Tuvastuses rakenduvad valentsipöörajad (nt eitus).

Emotsioonituvastuseks on loodud Emotsioonidetektor (v 0.4) http://peeter.eki.ee:5000/valence ning vastav Google Chrome'i laiendus. Arendus jätkub.

Detektoriga on analüüsitud ja lugejahinnanguga võrreldud "Postimehe" 6 rubriigi 1500 lõiku ja loodud neist korpus. Emotsioonihinnangu andmine analüüsitavale tekstile praegu leksikonipõhine. Tuvastuse õigsus keskm 82%. Korpuse põhjal loomisel lõigu statistiline hindaja.

Kasutusala: Enda ja teiste kirjutatud tekstide (nt meilide, leheartiklite) võimaliku emotsionaalse mõju hindamine, tekst-kõne sünteesi sisendteksti emotsionaalsuse määramine sobiva kõneesituse leidmiseks jt.

2. Kõnelejakohane emotsioonituvastaja

Meetod statistiline. Kasutame vabavaralist kõneemotsioonituvastajat OpenSMILE http://www.openaudio.eu

Treeningmaterjal: esilekutsutud emotsioonidega laused (naishääl Eesti emotsionaalse kõne korpusest). Katsetatakse nii kategoriaalset klassifitseerimist (viha, rõõm, kurbus), kui ka dimensioonilist (positiivne – negatiivne, aktiivne – passiivne). Dimensioonilise tuvastuse jaoks kõigi korpuse lausete valents ja aktiivsus kuulamistestidega määratud.

Töötav versioon plaanis projekti lõpuks.

Kirjandus

Pajupuu, H., Kerge, K., Altrov, R. (2012) Lexicon-based detection of emotion in different types of texts: preliminary remarks. Estonian Papers in Applied Linguistics, 8, 171-184.
Keysar, B., Hayakawa, S.L., An, S. G. (2012) The foreign-language effect: Thinking in a foreign tongue reduces decision biases. Psychological Science, 661-668.

KÕNE- JA MULTI-MODAALSED KORPUSED

Projekti eesmärgiks on 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.

Saavutatud ja oodatavad tulemused

Loengukõne korpus: eesmärk 30 tundi uusi märgendatud salvestusi.
2012 seis: kogutud 25 tundi uusi salvestusi, märgendatud 40 tundi.
Aktsendikorpus: eesmärk salvestada 40 uut eri keeletaustaga keelejuhti.
2012 seis: salvestatud 25 uut keelejuhti.
Raadiointervjuude korpus: eesmärk 80 tundi märgendatud salvestusi.
2012 seis: kogutud ja märgendatud 40 tundi salvestusi.
Noorukite kõnekorpus: eesmärk salvestada 200 keelejuhti vanuses 8-15 ja märgendada kogu korpus.
2012 seis: tekstikorpus, mobiilne salvestuskomplekt, läbirääkimised koolidega, isikukaitse küsimused, proovisalvestused.
Nimega üksuste korpus: eesmärk salvestada ja märgendada kuni 50000 nimega üksust kuni 200 keelejuhiga.
2012 seis: tekstikorpus, proovisalvestused.
Eriliigilised kõnekorpused: kavandatakse ja salvestatakse sõltuvalt esilekerkivatest uurimisvajadustest.
2012 seis: koostatud tekstikorpus fookusrõhu akustiliste tunnuste uurimiseks, salvestatud 11 keelejuhiga.

Multimodaalsed korpused:

- **kõneproduktsiooni andmebaas:** eesmärgiks on eestikeelse kõne artikulatsiooni kirjeldava andmebaasi salvestamine kasutades erinevaid mõõtesüsteeme – larüngograaf, palatograaf, EMA (elektro-magneetiline artikulograafia); 2 keelejuhti, ca 4 tundi kõnet; 2012 seis: koostatud tekstikorpus, salvestatud 1 keelejuhiga, jätkub märgendamine;
- **põhiviseemide korpus:** eesmärk on salvestada eesti põhiviseemide korpus audiovisuaalse kõnesünteesi projekti jaoks;
  2012 seis: korpus salvestatud 1 keelejuhiga;
- **viipekeele korpus:** eesmärgiks on eesti viipekeele baaskorpuse loomine kasutades video ja 3D ruumilist liikumist registreerivat mõõtesüsteemi; korpuse loomist alustatakse 2014.

SUULISE EESTI KEELE AUDIOVISUAALSE SUHTLUSKORPUSE KOGUMINE JA PÄRINGUSÜSTEEMI ARENDAMINE

Projekti eesmärgid on

a) filmida ja salvestada suulise eesti keele kasutust tegelikes suhtlussituatsioonides;
b) translitereerida tekstid ja varustada taustakirjeldusega keelekasutust mõjutavate keeleväliste nähtuste kohta;
c) arendada arvutitarkvara, mis võimaldab otsida korpusest erinevaid keelelisi nähtusi ning neid analüüsida.

Ainult sellise korpuse abil saab analüüsida tegelikku kõnekeelt ja suhtlust ning leida seal järgitavad keele- ja suhtlusnormid. See omakorda võimaldab modelleerida inimese ja arvuti dialoogi pragmaatikat, nii et suhtlus oleks võimalikult sarnane inimeste tegelikule suhtlusele.

Korpuse kogumise metoodika.

1) Kõigepealt salvestatakse videokaamera(te)ga reaalseid suhtlussituatsioone.
2) Seejärel transkribeeritakse nende vokaalne osa (sõnad, lausungid, pausid, kõnetempo ja hääle muutused jms).
3) Transkriptsioon ja salvestuse heli sünkroniseeritakse kõnevoorude kaupa programmiga Transana. (Mittevokaalse osa transkribeerimine ilma konkreetse uurimisülesandeta ei ole mõttekas, kuna töö maht on liiga suur).
4) Neile lisatakse taustakirjeldus (sugu, vanus, sotsiaalne staatus, haridus; argi/avalik suhtlus, silmast silma/telefonisuhtlus; dialoog/monoloog jms). Selle täitmiseks on vastav blankett.
5) Suhtluses osalejad täidavad lepingu, millega nad annavad materjali korpusesse uurijatele kasutamiseks.

Otsingutarkvara arendamine. Päringusüsteem võimaldab otsida korpuse litereeringutest sama sõnavormi erinevaid variante (nt hobune, obene jms) seotuna kasutajate erinevate sotsiaalsete parameetritega (naised/mehed, erinev haridus jms). Jätkame päringusüsteemi arendamist:

a) teeme täpsemaks variantide otsingu ja lisame otsimisel kasutatavaid parameetreid,
b) loome võimaluse otsida üksiksõnade kõrval mitme lähestikuse sõna järjendeid,
c) loome võimaluse otsida sõnade erinevaid grammatilisi vorme,
d) loome võimaluse otsida erinevaid lauseliikmeid.

Viimase kahe lisanduse tarvis tuleb integreerida süsteemi suulise keele morfoloogia ja süntaksi analüsaatorid. Nende kasutamiseks loome programmi, mis võimaldab viia korpuse ortograafilisele tasandile.

Kõik tegevused on pooleli ja kulgevad plaanipäraselt.