EKKTT 2010

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

EESTI KEELE KEELETEHNOLOOGILINE TUGI

25.–26. november 2010
Konverentsikeskus Dorpat, Tartu

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

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

Programmi raames rahastati keeletehnoloogiaalast teadus- ja arendustegevust alates ressursside loomisest kuni keeletehnoloogiliste rakenduste prototüüpide loomiseni. Programmi tulemusena tekkinud intellektuaalne omand on avalik omand.

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

Programmi juhtkomitee koosseis

Juhtkomitee esimees: Jaak Vilo (Arvutiteaduse instituut, Tartu Ülikool)

Juhtkomitee aseesimees: Einar Meister (Tallinna Tehnikaülikooli Küberneetika Instituut)

Liikmed: Heiki-Jaan Kaalep (Arvutiteaduse instituut, Tartu Ülikool)

Kaili Müürisepp (Arvutiteaduse instituut, Tartu Ülikool)

Karl Pajusalu (Eesti ja üldkeeleteaduse instituut, Tartu Ülikool)

Indrek Reimad (Haridus- ja Teadusministeerium)

Urmas Sutrop (Eesti Keele Instituut ja Tartu Ülikool)

Uuno Vallner (Majandus- ja kommunikatsiooniministeerium)

Kadri Vider (Haridus- ja Teadusministeerium)

EKKTT koordinaator: Neeme Kahusk (Arvutiteaduse instituut, Tartu Ülikool)

Grammar-based interactive pedagogical programs for Sami and other Finno-Ugrian languages

Lene Antonsen,
Ciprian Gerstenberger and
Trond Trosterud,
University of Tromsø.

The last decade or so has witnessed a large paradigm shift within language technology, from grammar-based to statistically based approaches. The field now presents itself as somewhat diveded, with grammar-based theoretical projects like for example LFG and HPSG based parsing in academic institutions, and statistical projects in the industry. Also, to the extent that academic institutions need working applications for basic work (such as POS tagging of larger corpora), they rely upon statistical methods. Seen from a linguistic point of view, the grammatical approaches represent explicit models of the languages in question, in essence empirically falsifyable hypotheses about the grammars. As such they are cleary more interesting than the black boxes provided by statistical approaches.

Our approach represents a third, middle road, combining the robustness of the statistical approaches with the transparency and explicitness of grammatical approaches. North Sami, like Estonian, is a language combining a rich agglutinative morphology with complex suprasegmental morphological processes. We model it by using finite-state transducers. For morphological disambiguation and syntactic analysis we use the constraint grammar framework.

Based upon this basic analysis we have built a long range of practical applications, ranging from spellcheckers to machine translation systems. The talk will look closer at one of the applications, oahpa.uit.no, a set of interactive pedagogical programs for North Sami. Oahpa consists of a set of pedagogical programs. Two of the programs are system-governed dialogue system, accepting free input from the user. The input will then be subject to grammatical error analysis, and the user will receive adequate comments.

The Oahpa programs are presently being ported to South Sami. All the components needed to build an oahpa system are already present for Estonian. The talk will result in a suggestion for building language learning programs for Estonian along the lines presented here.

Spoken Language Processing: tomorrow's technology today or today's technology tomorrow?

Prof ROGER K MOORE
Chair of Spoken Language Processing
Vocal Interactivity Lab (VILab)
Speech & Hearing Research Group (SPandH)
Department of Computer Science, UNIVERSITY OF SHEFFIELD

Recent years have seen steady improvements in the quality and performance of speech-based human-machine interaction driven by a significant convergence in the methods and techniques employed. Spoken language processing has finally emerged from the research laboratory into the real-world, and members of the general public now regularly encounter talking and listening machines in their daily lives. However, whilst several niche markets have been established, there is general concern that spoken langauge technology is still not sufficiently robust for a range of valuable applications, and that the capabilities of contemporary spoken language systems are falling short of what users expect and the market needs. This talk will review the current state-of-the-art in spoken language processing, and will take a critical look into the future taking into account the results of various surveys conducted by the author into the opinions of the top practitioners in the field. The talk will conclude with a discussion of what the author believes are the key scientific barriers to progress.

Eesti keeleressursside keskus

Vastutav täitja	Tiit Roosmaa
Teised põhitäitjad	Krista Liin
Finantseerimine	1,38 miljonit krooni
Kestus	2008-2010

Käesoleva projekti toel tehti ära ettevalmistav töö Eesti keeleressursside keskuse käivitamiseks. Eesti keeleressursside keskus on Eesti tugistruktuur, mis käivitunult omab nii riistavaralisi ressursse kui ka teadmisi, kuidas juba olemasolevaid keeleressursse kasutades eri valdkondade teadlasi parimal moel keeletehnoloogiliselt toetada.

Käesolevas kontekstis mõistame keeleressursside all nii keeletarkvara (s.o. keeletöötlusprogramme) kui keeletehnoloogilisi ressursse (elektroonilised sõnastikud, keeleandmebaasid, teksti- ja kõnekorpused). Loomuliku keele ressursid on erinevate soovijate/huvi-liste poolt kasutatavad ainult siis, kui olemasolevad keeleressursid on korralikult doku-menteeritud ja arhiveeritud ning avalikult kättesaadavad. Selliste, kohati keeleressursside loojatele tarbetutena tunduvate tegevuste toetamiseks on vaja teatud infrastruktuuri olemasolu, mis korraldaks ja koordineeriks sellealast tööd Eestis.

Loodava keskuse põhifunktsioonid on olemasolevate ressursside kaardistamine ja kogumine, neile juurdepääsu ja selle säilimise tagamine, kasutajate vajaduste väljaselgitamine, keeletehnoloogiliste standardite väljatöötamine ja fikseerimine, keeleressursside kasutamiseks vajalike juriidiliste lepingute ja litsentside koostamine.

Et tagada keskusse kuuluvate keeleressursside pikemaajaline kasutusvõimalus, võimalus eri ressursside kombineerida, võrrelda ja kasutada koos erinevate eestisiseste või ka välismaiste rakendustega, viiakse ressursid vastavusse üldlevinud standarditega, dokumenteeritakse ning tehakse nende metaandmed kättesaadavaks ja automaatselt töödeldavaks nii eesti kui inglise keeles. Keskuses töötatakse välja litsentsilepingud ja viiakse sisse kvaliteetne 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. Avaliku sektori puhul panustatakse eelkõige riigi infosüsteemide arendamisse. Erasektoril võimaldatakse erikokkulepete alusel keeleressursse kasutada ja rakendada ka oma toodetesse.

Eesti keeleressursside keskus on kavandatud osana üle-euroopalisest võrgustikust, mille loomiseks on käivitunud ESFRI projekt CLARIN (Common Language Resources and Technology Infrastructure, http://www.clarin.eu), kus üheks kolmekümne kuuest partnerist on Eesti ametliku esindajana ka Tartu Ülikool. Osalemine CLARINi võrgustikus annab meile unikaalse võimaluse kaasa-

ta oma probleemide lahendamisse üle-euroopaline kogemus. Hetkel käivad läbirääkimised Eesti liitumiseks CLARIN-ERIC (European Research Infrastructure Consortium) võrgustikku asutajaliikmena. Eesti keeleressursside keskuse osalemine rahvusvahelises võrgustikus annab meie teadlastele juurdepääsu keskus-te võrgustikus olevate keskuste teiste keelte jaoks loodud ressurssidele.

Eesti keeleressursside keskus nimetati vabariigi valitsuse poolt 2010 aastal üheks kahest Eesti teaduse infrastruktuuri teekaardi objektiks humanitaar- ja sotsiaalteaduste alal.

Eesti keele koondkorpuse esituse ja kasutusvõimaluste arendamine

Vastutav täitja	Kadri Muischnek
Teised põhitäitjad	Kaarel Veskis, Kristel Uiboaed, Katrin Tsepelina
Finantseerimine	550 000 krooni
Kestus	2010

Projekti eesmärk

Korpusekogumise projekt "Eesti keele koondkorpus" lõppes 2009. aastal. Selle tulemusena on valminud ligi 250 miljoni sõnaline avalikult vabalt kasutatav kirjaliku eesti keele kogu.

Praegune, käimasolev projekt ühendab endas mitut ülesannet, mille ühiseks eesmärgiks on Koondkorpuse täiustamine ja tema kasutusvõimaluste laiendamine.

Ülesanded

1. Koondkorpust koguti küllaltki pika aja jooksul ja sellest tingitud ebaühtlused märgenduses ja kodeeringutes ühtlustati projekti viimasel, 2009. aastal. Koondkorpuse koostamise aastate jooksul on aga muutunud ka standardid; kõige levinumaks märgenduskeeleks on meie korpuses kasutatava SGML asemel saanud XML. Samuti on meie korpuses kasutatava ASCII kooditabeli + olemite (entities) süsteemi asemel standardkooditabeliks saanud UTF-8.

   Projekti esimeseks ülesandeks planeeritigi üleminek UTF-8-le ja XML-keelele. See töö on lõpetatud, praegu toimub kolmetasemelisele päiste (header) süsteemile üleminek.

2. Koondkorpuse kasutajaliidesele lisatakse käesoleva projekti raames kollokatsioonide leidja esialgne versioon. Kollokatsiooni all mõeldakse siin selliseid sõnavorme (või ka lemmasid), mis esinevad tekstis koos (st esinevad üksteise naabruses) sagedamini kui võiks eeldada nende eraldisesinemise sageduste põhjal.

   Kollokatsioonide leidja esialgne versioon asub aadressil

   http://www.rabauti.ee/clc/ , hiljem hakkab ta asuma aadressil

   http://www.keeleveeb.ee .

3. Koondkorpuses on 22 miljonit sõna nn uue meedia keelekasutust (jututoad, kommentaarid, uudisgrupid, foorumid). Kui muud Koondkorpuse tekstid on automaatselt morfoloogiliselt märgendatud, siis uue meedia tekstide leksika, ortograafia ja kohati ka morfoloogia on normeeritud kirjakeelest niivõrd erinevad, et kirjakeele analüüsiks loodud morfoloogiline analüsaator ja ühestaja t3mesta teeb nende

analüüsil liiga palju vigu. Projekti kolmandaks ülesandeks oligi morfoloogiaanalüsaatori kohandamine uue meedia keelekasutuse töötlemiseks. Praeguseks on välja töötatud analüsaatori tekstiklassiga kohandamise meetod ja käib tekstide märgendamine.

Töötajad

Projekti kallal töötavad Kaarel Veskis, Kristel Uiboaed ja Katrin Tsepelina ning Kadri Muischnek vastutava täitjana.

Korpusepäring keeleveebis (www.keeleveeb.ee)

Vastutav täitja	Heiki-Jaan Kaalep
Teised põhitäitjad	Rene Prillop, Tarmo Vaino, Katrin Tsepelina, Andrei Dementjev
Finantseerimine	1,3 miljonit krooni
Kestus	2006 - 2010

Eesmärgid

Eesmärgiks on võimaldada sõnastike ja tekstikorpuste mugavat kasutamist üle interneti.

1. Suurendada veebis kättesaadavate erialasõnastike arvu, teisendades olemasolevaid standardsele XML-põhisele kujule TBX (Term Base eXchange,

   http://www.lisa.org/standards/tbx/), mis on just termini-sõnastike ja –baaside jaoks mõeldud standard.

2. Luua eesti keele morfoloogiat arvestav korpuse päringu süsteem.

3. Siduda sõnastiku- ja korpusepäringud omavahel, nii et lisaks sõnastikuvastele saab kasutaja ka näiteid otse korpusest.

Käesolevas projektis ühendatakse olemasolevad eesti keele ressursid — tekstikorpus, morfoloogiline analüsaator koos ühestajaga ning sõnastikud — uueks tervikuks, mis annab seni puudunud võimalused nii uurijatele, eesti keele õppijatele kui ka muidu huvilistele.

Liidetakse järgmised ressursid, mille kasutamiseks antud projekti raames ei ole autoriõiguslikke takistusi: TÜ eesti keele koondkorpus (http://www.cl.ut.ee), Filosofti morfoloogiline analüsaator (http://www.filosoft.ee) ja ühestaja ning mitmete eri autorite sõnastikud (http://www.keeleveeb.ee).

Morfoloogilise analüsaatori ja ühestaja kasutamine võimaldab otsida korpusest sõnu, ilma et peaks muretsema sõna muutevormide rohkuse pärast. See omakorda võimaldabki teha korpuse-päringut samasuguse lihtsusega kui sõnastikupäringut.

Tulemused

Kõik eesmärgid täideti.

1. Keeleveebis (http://www.keeleveeb.ee) on tehtud tasuta kasutatavaks 30 erialasõnastikku kogumahuga 200 000 mõistet. Kõik erialasõnastikud on kasutatavad ühispäringus, millesse on hõlmatud ka 30 keeleveebi-välist sõnastikku. See tähendab, et saab otsida sõna või terminit kuni 60 sõnastikust korraga.

2. Eesti keele koondkorpus (http://www.cl.ut.ee) on morfoloogiliselt analüüsitud ja ühestatud, indekseeritud sõnavormi, lemma ja grammatilise info järgi ning kasutamiseks väljas. Korpuse päring käib lausete kohta; saab kombineerida mitut sõna ja/või grammatilist kategooriat ning piirata nende koosesinemist osalausega. Otsing on realiseeritud Filosofti omaloodud päringumootoril.

3. Lihtne korpusepäring käib täpselt sama moodi kui lihtne sõnastikupäring.

Edasised suunad

Korpuse kasutajaliidest saaks muuta veel kasutajasõbralikumaks, kui korpusesse lisada olemasolevale lingvistilisele märgendusele juurde uusi kihte. Käesoleva projekti käigus märkisime osalausete piirid, mis võimaldab mitmesõnalised päringud muuta tunduvalt täpsemaks. (Seda ülesannet projektis algselt ei olnud.) Kui edaspidi lisada kasvõi osalist süntaktilist infot, nt märgendades atribuudid, või semantilist infot, nt märgendades nimeüksused, annaks see võimaluse uurida keelt viisil, mis praegu veel pole võimalik.

Administriviaalset

Projekti kestus: 2006-2010. Kulud: 1,3 miljonit krooni. Vastutav täitja: Heiki-Jaan Kaalep. Projektis osalejad: Rene Prillop, Tarmo Vaino, Katrin Tsepelina, Andrei Dementjev

Leksikograafi töökeskkond

Vastutav täitja	Ülle Viks
Teised põhitäitjad	Andres Loopmann, Indrek Hein, Ain Teesalu, Kristina Koppel, Kati Sein, Merike Koppel, Julius Juurmaa, Sven-Olav Paavel
Finantseerimine	6,5 miljonit
Kestus	2006-2010

Veebiaadress: http://eelex.eki.ee/

Eesmärgid ja tähtsus

Projektil on kolm põhieesmärki:

1. Luua leksikograafidele sobiv interaktiivne töökeskkond e sõnastike haldussüsteem EELex, st töövahendid, mis ühilduvad kehtiva rahvusvahelise märgistusstandardiga (XML) ja rakendavad nii universaalseid kui ka eesti keele põhiseid keeletehnoloogia vahendeid: keeleressursse ja keeletarkvara.

2. Koostada eesti lähtekeele andmebaas uute kakskeelsete sõnaraamatute jaoks ehk Eesti–X-keele sõnastik.

3. Anda projekti tulemused avalikku kasutusse: (a) süsteemi kuuluvate sõnastike avalikud veebiversioonid, (b) sõnastike haldussüsteemi laiatarbeversioon.

Tähtsus

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

Põhitulemused

Projekti tulemusena on loodud veebipõhine leksikograafi töökeskkond EELex, mis ühendab leksikograafile vajaliku tarkvara ja keeleressursid, toetab rühmatööd ja pakub eesti keele tuge.

Töökeskkond

EELexi tarkvara on sõnastike haldussüsteem, mis võimaldab sõnastikke koostada, toimetada ja küljendada, teha lihtsaid ja keerulisi struktuuripõhiseid päringuid ning päringutulemusi sortida. Leksikograaf kasutab korraga kaht omavahel seotud tööakent, ühes sõnaartikli tekst koos struktuurimärgendusega ja teises küljendatud kujul. Toimetaja töö hõlbustamiseks on loodud mitmeid tööriistu,

nt ristviidete kontroll, hulgiparandused kogu sõnastikus, eesti morfoloogia andmete genereerimine, küljendusvaate kujundus, sõnastikuteksti eksport Wordi jms.

Alanud on koostöö kirjastustega ja teiste asutustega, kus sõnastikke koostatakse või uuteks rakendusteks (nt telefonisõnastikud) ette valmistatakse — nii Eestis kui ka väljaspool.

Leksikaalsed ressursid

Projekti käigus on loodud kakskeelsete sõnastike pooltoode – Eesti-X sõnastiku andmebaas, kus on olemas eesti pool (eesti märksõna kohta käivad andmed, nt sõnaliik, muutevormid, tähendusjaotus, näitelaused jm). Sihtkeele (tõlkevastete) info lisab uue kakskeelse sõnastiku koostaja. EELexi leksikaalsete ressurside hulka kuuluvad kõik sõnastike haldussüsteemis koostatud või sinna üle viidud sõnastike andmebaasid. Need on standardse XML märgendusega taaskasutatavad keeleressursid, mida saavad kasutada leksikograafid ja keeletehnoloogid uute sõnastike ja andmebaaside koostamiseks. Seni on EELexi töökeskkonnas trükivalmis saanud 7 sõnastikku, leksikograafide käsutuses (koostamisel ja toimetamisel) on 14 sõnastikku, keeletehnoloogilise arenduse ja testimise järgus on 3 sõnastikku.

Avalik kasutus

1. Tarkvara. Professionaalse leksikograafi töökeskkonna baasil on selle kõrvale loodud EELexi avalik laiatarbeversioon (http://exsa.eki.ee/), mille abil saab oma sõnastikku luua ja toimetada veebis ka tavakasutaja.

2. Sõnastike veebiversioonid. Olulisemad EELexis valminud sõnastikud on koos struktuuripõhise päringu võimalustega tehtud avalikult kättesaadavaks veebis (http://portaal.eki.ee/).

Eesti vanema kirjakeele elektroonilised kogud

Vastutav täitja	Külli Habicht
Teised põhitäitjad	Valve-Liivi Kingisepp, Pille Penjam, Külli Prillop, Kristel Ress
Finantseerimine	95 000 krooni
Kestus	2009-2010

Eesmärgid ja tähtsus

Viis aastat kestnud projekti põhieesmärgiks oli eesti vanema kirjakeele elektrooniliste kogude täiendamine ning elektrooniliste ressursside kasutamiseks mõeldud paindlike kasutajaliideste loomine. Projekti käigus on täiendatud 18. sajandi tekstide valikkorpust, pandud alus 19. sajandi tekstide korpusele, lemmatiseeritud ja morfoloogiliselt märgendatud 17. sajandi tekste ning arendatud märgendustarkvara ja elektrooniliste kogude kasutajaliideseid. Vana kirjakeele perioodi olulisemate tekstide laiemasse kasutusse andmine võimaldab edaspidi säästa uurijate töövaeva ja hõlbustab diakroonilist keeleuurimist. Vanade tekstide ja sõnastike elektrooniline kogu on lisaks uurimisotstarbele oluline ka eestlaste keelelise ja kultuurilise identiteedi säilitajana nüüdisaegses infoühiskonnas.

Projekti tulemusi saab lisaks lingvistilisele uurimistööle edukalt kasutada ka e-hariduses (nt on TÜ teaduskool koos vana kirjakeele uurimisrühma teaduritega käivitanud õpilastele mõeldud programmi vanema põhjaeestilise kirjakeele kohta) ja e-infopäringutes (etümoloogid; kirjastused, kel huvi teatud perioodi vanade tekstide vastu; leksikograafid (diakrooniline mõõde); piiblitõlke uurijad).

Tekstikorpus koos päringusüsteemiga on kasutatav Tartu Ülikooli vana kirjakeele uurimisrühma veebilehel http://www.murre.ut.ee/vakkur/Korpused

Korpusepäringu tarbeks on välja töötatud kasutajaliidesed, mis võimaldavad päringuid teha nii märgendamata kui ka märgendatud tekstidest. Tekstide lemmatiseerimiseks on kasutusel Külli Prillopi disainitud originaaltarkvara "Vakker", mis võimaldab vanu ebaühtlase ortograafia ja vormistikuga tekste paindlikult märgendada.

Põhitulemused

Keeleressursside laiendamine:

- 18. sajandi tekstide valikkorpuse ja 19. sajandi esimese poole tekstide korpuse täiendamine umbes 750 000 tekstisõna mahus.
- Heinrich Stahli tekstide lemmatiseerimine ja morfoloogiline märgendamine 80 000 tekstisõna mahus.
- Vanimate sõnastike koondandmebaasi loomine (Stahl 1637, Gutslaff 1648, Göseken 1660).
- Eesti kirjakeele esmaesinemussõnastiku andmebaasi loomine.

Tarkvara arendamine:

- Vana kirjakeele korpuse kasutajaliidese pidev arendamine ja kasutajasõbralikumaks muutmine, k.a veebipõhise kasutajaliidese loomine lemmatiseeritud ja morfoloogiliselt märgendatud vana kirjakeele korpuse jaoks.
- Ebaühtlases kirjaviisis tekstide lugemise hõlbustamiseks mõeldud veebirakenduse loomine – see võimaldab lihtsa ligipääsu olemasolevale leksikaalsele ja grammatilisele lisateabele.
- Märgendamisprogrammi "Vakker" täiustamine.

Eesti fraseologismide elektrooniline alussõnastik (FES)

Vastutav täitja	Katre Õim
Teised põhitäitjad	Asta Õim
Finantseerimine	0,5 miljonit krooni
Kestus	2008-2010

Eesti fraseologismide elektrooniline alussõnastik on mõisteline sõnaraamat, kus fraseologismid on korraldatud mõisteseoste järgi. Sõnastik kui semantilis-leksikaalne keeletehnoloogiline ressurss sisaldab 20755 fraseologismi, mida on analüüsitud kognitiivse keeleteaduse põhimõtetel.

Sõnastikus käsitatakse fraseologismina konventsionaalset sõnaühendit, mille mõistmiseks ei piisa ainult keele grammatikareeglite ja sõnavara tundmisest.

Sõnastiku alusandmed (fraseologismid, nende kogumiskihelkond, kasutus-kontekst) pärinevad Eesti kõnekäändude ja fraseologismide andmebaasist (http://haldjas.folklore.ee/justkui/).

Tegemist on Eesti Rahvaluule Arhiivi kõnekäänukartoteegi keelematerjaliga, mis on kogutud põhiliselt 20. sajandi esimesel poolel, uusimad andmed piirduvad 1980ndate aastatega.

Sõnastik järgib ühtset ontoloogilist mõistehierarhiat, kust on mõistetevaheliste semantiliste seoste näitamiseks vahet tehtud kategooriatel, klassidel, üldmõistetel ja mõistetel. Tegemist on sõnastiku koostajate tehtud alusandmete üldistusega. Mahukamate mõistete puhul suureneb ka tõenäosus, et kasutaja suudab neid otsitavate fraseologismidega seostada.

Sõnastiku põhiüksused on 1964 mõisteartiklit, mille koostamisel on lähtutud Eesti kõnekäändude ja fraseologismide mõistestikus (http://haldjas.folklore.ee/justkui/moiste.php) eristatud mõistetest. Viimaseid on olulisel määral konkretiseeritud ja toimetatud. Mõisteartikli keskmes on vastava mõistega seotud ja seega suhteliselt lähedase tähendusega fraseologismid. Polüseemilised või homonüümsed fraseologismid (mida on 6058) kuuluvad mitmesse mõisteartiklisse. Seega on sõnastikus kokku 26813 tähendusüksust. Lisaks üldisele tähendusele antakse mõisteartiklis infot fraseologismide kasutuse, süntaktilise struktuuri, vormistiku ja leviku kohta.

Fraseologismid esitatakse sõnastikus lemma kujul. Tüüpjuhul on fraseologismi lemma kõige selgem, tavalisem ja sagedasem morfosüntaktiline kuju, millega fraseologism oma idiomaatilist tähendust ja funktsiooni aktiveerib.

Sõnastikus näited 1) selgitavad ja täpsustavad fraseologismi tähendust või osutavad fraseologismi kasutusvaldkonnale, 2) selgitavad fraseologismi kasutamist ning formaalset (s.o morfosüntaktilist, süntaktilist ja leksikaalset) teisenemist konkreetsete kasutussündmuste käigus ja loomulikes seostes. Kui näited sisaldavad fraseologismide murdevariante, on olulisemate murdesõnade juures toodud nende kirjakeelne vaste. Vajadusel on näiteid lühendatud ja keeleliselt toimetatud. Näited puuduvad, kui alusandmed ei ole näidanud fraseologismide kasutust ja teisenemist loomulikus keeles.

Sõnastikus on fraseologismi kui terviku funktsioneerimist ja lemmat arvestades vahet tehtud fraseologismidel, mis väljendavad lausega tähistatava situatsiooni komponente, ja lausekujulistel fraseologismidel.

Sõnastikus esitatakse fraseologismide lemmade osiste struktuur (nt tüvi, järelliide, lõpp), sõnaliik ja kääne või pööre. Fraseologismide lemmad on analüüsitud morfoloogilise analüsaatoriga ESTMORF. ESTMORFi jaoks tundmatute sõnade analüüsimisel on kasutatud oletajat, kuid paljud murdesõnad on morfoloogiliselt analüüsimata.

Fraseologismide leviku piirkond esitatakse sõnastikus kihelkonna täpsusega. Kuna sõnastiku alusandmete kogumine ei ole olnud kuigi süsteemne, siis ei ole fraseologismide levikuandmeid ja näiteid üksteisega seotud.

Eesti fraseologismide elektroonilise alussõnastiku kasutajaliides asub aadressil

http://haldjas.folklore.ee/justkui/sonastik/

TÜ eesti keele tesauruse (eesti wordneti) täiendamine

Vastutav täitja	Heili Orav
Teised põhitäitjad	Kadri Kerner, Sirli Parm, Piia Taremaa, Lauri Eesmaa
Finantseerimine	1 563 674 krooni
Kestus	2007 - 2010

Eestis on erineva sisu ja ülesehitusega sõnastikke palju, kuid mõistelisi sõnastikke kaks: esimene Andrus Saareste 1978. a. Uppsalas välja antud ajaloolise tähtsusega "Eesti keele mõisteline sõnaraamat" ja teine Tartu Ülikoolis koostatav uuema põlvkonna arvutitesaurus — Eesti Wordnet.

Eesti keele tesauruse (Eesti Wordneti) loomine käivitati 1998 aastal EuroWordNeti projekti (http://www.illc.uva.nl/EuroWordNet/) käigus, kus samade põhimõtete järgi koostati 8 erineva keele tesaurused. Tähendusüksuste vahel kehtestatakse 45 erinevat semantilist seost, nagu alam-/ülemmõisted, antonüümia, osa-terviku suhe jms. Mõisted on seotud ka nende inglisekeelsete vastetega. Töö eestikeelse tesaurusega jätkub siiani.

Käesoleva riikliku keeletehnoloogia projekti eesmärgiks on olnud andmebaasi suurendamine, täiendamine ja olemasoleva kontrollimine ning parandamine. Projekt kestis neli aastat ja finantseerimine neil aastatel on olnud kokku 1 563 674 krooni.

Projekti alguses prognoosisime tesaurusesse vähemalt 25 000 uut mõistet. See eesmärk on täidetud - Eesti Wordnetis on praeguse seisuga (november 2010) üle 42 500 mõiste (projekti alguses oli u 15 500 mõistet). Töö on kulgenud mitmesuunaliselt. Esiteks oleme lisanud uusi sõnaliike – adjektiive ja adverbe. Teiseks oleme täiendanud tesaurust kitsaste valdkondade kaudu (nt transport, isiksuseomadused, liikumine jms). Selline valdkonna-spetsiifiline lähenemine muudab mõistete ja nende vaheliste semantiliste suhete lisamise täpsemaks. Ja kolmandaks on toimunud andmebaasi täiendamine sõnatähenduste ühestamise andmete põhjal.

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

Masintõlge I-II

Vastutav täitja	Heiki-Jaan Kaalep
Teised põhitäitjad	Mark Fišel, Harri Kirik, Kaarel Veskis, Katrin Tsepelina
Finantseerimine	3,07 miljonit krooni
Kestus	2004-2010

Masintõlke projektide eesmärk on arendada automaatset tõlget eesti ja inglise keele vahel. Esimese projekti ("Masintõlge I") tulemuseks sai koostatud seaduste tõlkenäiteid sisaldav paralleelkorpus. Samuti oli loodud interneti teel kättesaadav eesti-inglise masintõlke süsteem (http://masintolge.ut.ee/). Jätkuprojekti ("Masintõlge II") peamised eesmärgid olid uurida tüüpilisi probleeme eesti-inglise masintõlkes ning parandada olemasoleva tõlkesüsteemi väljundi kvaliteeti.

Eesti-inglise masintõlke olukord on praegu järgmine: lisaks TÜ tõlkesüsteemile on olemas Google'i tõlketeenus, kus on üle 40 omavahel tõlgitava keele, s.h. ka eesti keel: http://translate.google.com/. Nii Google'i kui TÜ versioon kasutavad statistilist masintõlkemeetodit.

Mitmete katsete tulemusena, kus prooviti erinevaid korpusi ja erinevaid morfoloogilise analüüsi viise, on selgunud, et:

1. Eestikeelsete sõnade morfoloogiline analüüs, mille käigus sõnad on tükeldatud tüvedeks ja lõppudeks, aitab kaasa õigete ingliskeelsete fraaside leidmisele ja seega ka paremale tõlkele.

2. See ei aita parandada tõlkeprobleeme, mille põhjuseks on eesti ja inglise keele erinev sõnajärg.

Käesoleva projekti käigus arendati meetodit, kuidas parandada sõnajärje modelleerimist ilma süntaksi analüüsi kasutamata. Meetod põhineb tüüpilistel sõnaliikide järjenditel ning kasutab fakti, et sõnaliikide puhul on andmed vähem hõredad, mis tähendab mudelite parameetrite stabiilsemat statistilist õppimist. Katsed näitavad, et sellel on hea mõju suulise keele tõlkekvaliteedile ning halvem mõju spetsiifilisema kirjakeele (s.t. seadustekstid) tõlkekvaliteedile.

Samuti sai TÜ tõlkesüsteemi abil uuritud tüüpilisi lauseid, mida kasutajad tõlkida soovivad. Peamine probleem on seejuures see, et seadustekstide peal õpetatud tõlkesüsteem ei saa loomulikuma tekstiga väga hästi hakkama; see aga osutab teistest valdkondadest tekste sisaldavate korpuste vajadusele.

Paralleelselt sai arendatud eesti-inglise interneti kaudu kasutatav tõlkeabisüsteemi (http://kde.teataja.ee/). Süsteemi põhikomponent on leksikon, mis on koostatud automaatselt paralleelkorpuse põhjal. Süsteem otsib vasteid päringusõnale, ning näitab neid koos sobivushinnangute ja korpuses esinemise näidetega.

Tuleviku arendamissuunad sisaldavad tüüpiliste tõlkevigade automaatset analüüsi ning süntaksanalüsaatoriga tõlkimise katsetamist. Samas vajab vastupidine tõlkesuund (inglise keelest eesti keelde) põhjalikku veaanalüüsi.

Intelligentne kasutajaliides andmebaasidele

Vastutav täitja	Mare Koit
Teised põhitäitjad	Mark Fišel, Olga Gerassimenko, Riina Kasterpalu, Krista Mihkels, Andriela Rääbis, Siiri Pärkson, Margus Treumuth
Finantseerimine	1,01 miljonit krooni
Kestus	2009-2010

Eesmärgid ja tähtsus

Projekti eesmärk on luua kasutajaliides, mis võimaldaks hõlpsat kohandamist erinevate ainevaldkondadega ja seostamist erinevate andmebaasidega. Liidest saab minimaalsete täienduste tegemise teel häälestada uutele ainevaldkondadele ja siduda andmebaasidega, andes kasutajale võimaluse pöörduda andmebaaside poole eesti keeles ning saada vastuseks adekvaatset infot.

Dialoogihalduris realiseeritakse infodialoogi juhtimise üldine mudel, mis võtab arvesse erinevates praktilistes infodialoogides kehtivad üldised seaduspärasused. Loodavat liidest saab kasutada ka nn võlur Ozi režiimis, kus arvuti rolli mängib inimene; see võimaldab lihtsasti koguda andmeid liidese häälestamiseks uuele ainevaldkonnale, määrata, missuguseid kasutaja lausungeid ja missuguseid dialoogiakte peab intelligentne liides hiljem suutma käsitleda ning kuidas nendele reageerima. Intelligentses liideses lõimitakse mõned olemasolevad ja/või teiste keeletehnoloogiaprojektide toel loodavad eesti keele automaattöötluse vahendid. Liidese loomise käigus täiendatakse ühtlasi Eesti dialoogikorpust ja selle töötlemise tarkvara.

Põhitulemused

1. Dialoogiaktidega märgendatud dialoogikorpuse maht on 2010.a lõpuks 245 000 tekstisõna (inimestevahelised suulised ja võlur Ozi meetodil kogutud kirjalikud dialoogid). Dialoogiaktide märgendamise käigus on korrastatud dialoogiaktide tüpoloogiat, täpsustatud ja täiendatud dialoogiaktide märgendusjuhendit. Dialoogikorpus on kättesaadav veebis (parooliga kaitstud).

   http://www.cs.ut.ee/~koit/Dialoog/EDiC.html.

2. On analüüsitud dialoogikorpuses leiduvate infodialoogide ülesehitust, sh telefonikõnede ning kirjaliku Interneti-suhtluse alustamist ja lõpetamist, partneri algatatud parandusi, sagedamini esinevaid dialoogiakte ja nende väljendamist eesti keeles, eesmärgiga modelleerida loomulikku dialoogi ja leida keelelisi märguandeid, mida dialoogsüsteem saaks kasutada dialoogiaktide automaatsel tuvastamisel.

3. Eesti dialoogikorpuse analüüsimise hõlbustamiseks on täiendatud varem väljatöötatud, veebis kasutatavat tarkvara, mis võimaldab otsida dialoogikorpusest ja loendada mitmesuguseid dialoogides esinevaid nähtusi: sõnajärjendeid, transkriptsioonielemente, dialoogiakte,

andes ette akti nime, osaleja tähise, sõne; teha automaatselt morfoloogilist analüüsi, määrata alamdialooge (partneri algatatud parandussekventse ja vastuse tingimuste täpsustamise alamsekventse). Tööpinki on lõimitud dialoogiaktide poolautomaatse märgendamise moodul. http://www.dialoogid.ee/dialoogid/.

4. On loodud veebipõhine dialoogi juhtimise tarkvara (intelligentne liides andmebaasidele), mis on häälestatav erinevatele ainevaldkondadele. Tarkvarasse on lõimitud eesti keele morfoloogiline analüüs ja süntees, eestikeelsete ajaväljendite ja pärisnimede tuvastamine, õigekirjakontroll ja ortograafiavigade parandamine, kõnesüntees. Tarkvara on testitud kinoinfo ja hambaraviinfo andmebaasidel.
http://www.dialoogid.ee/kinoagent/

5. Liides on kohandatud võlur Ozi eksperimentide läbiviimiseks, selle abil on kogutud 75 infodialoogi. http://www.dialoogid.ee/aivo/

Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine

Vastutav täitja	Tanel Alumäe
Teised põhitäitjad	Toomas Kirt
Finantseerimine	1,9 miljonit krooni
Kestus	2008 - 2010

Eesmärgid

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

Põhitulemused

Viimaste aastate jooksul on keskendutud suure sõnavaraga kõnetuvastusele, ning täisautomaatse pikkade kõnesalvestuste transkribeerimissüsteemi loomisele. Praeguseks on arendatud välja süsteem, mis automaatselt segmenteerib pika helisalvestuse lühemateks lõikudeks, klassifitseerib saadud lõigud kõneks ja mitte-kõneks (näit. muusika), grupeerib kõnelõigud kõnelejate järgi, ning transkribeerib kõnelõigud, kasutades kolmesammulist tuvastust, kus pärast igat sammu arvutatakse igale kõnelejale uued adapteeritud akustilised mudelid. Adapteerimise abil õnnestub tuvastusvigade arvu umbes 15% võrra vähendada. Sellise tuvastusstrateegia abil saavutatavat tuvastuskvaliteeti on hinnatud kolme liiki testandmete põhjal: raadiouudiste salvestuste puhul on valesti tuvastatud sõnu 14.9%, raadiote vestlussaadete puhul 28.6% ning keeletehnoloogia konverentsi ettekannete salvestuste puhul 37.1%.

Sel aastal valmis ka kõnesalvestuste transkriptsioonide sirvimist, salvestuste kuulamist, ning nendest otsingut võimaldav veebirakendus, millega saab tutvuda aadressil

http://bark.phon.ioc.ee/tsab .

Veebirakenduse lähtekood on saadaval AGPL litsensi alusel.

Projekti raames arendatud meetodeid kasutades implementeeriti ka kõnetuvastussüsteemi prototüüp radioloogidele (koostöös AS-iga Cybernetica), mis esmastes eksperimentides on andnud väga häid tulemusi (vähem kui 10% sõnavigu adapteerimata ning reaalajalise tuvastuse puhul).

Varasemalt on selle projekti raames arendatud kaks eesti keele tuvastustehnoloogiat kasutavat rakendusprototüüpi: autosegmenteerija ning häälega juhitav kalkulaator.

Kõnekeele ressursid ja kõnetehnoloogia andmebaasid

Vastutav täitja	Einar Meister
Teised põhitäitjad	Lya Meister, Rainer Metsvahi, Martin Külvik
Finantseerimine	2,9 miljonit krooni
Kestus	2006-2010

Eesmärgid

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

Põhitulemused

Uudistekorpus

Korpus sisaldab ca 300 tundi Eesti Raadio lühiuudiste salvestusi ja üle 8000 lk digitaliseeritud uudistetekste. Korpuse märgendamiseks on välja arendatud töökeskkond vabavaralise programmi Transcriber (http://trans.sourceforge.net) baasil, märgendatud on 30 tundi uudistesalvestusi. Märgendamine koosnes kahest etapist: 1. automaatse kõnetuvastuse abil genereeriti signaalifailidele vastavad tekstifailid, 2. Transcriberi abil kontrolliti automaatselt tuvastatud tekstide ja salvestuste vastavust ning korrigeeriti käsitsi tuvastusvead. Korpus on kättesaadav LAMUS-süsteemi kaudu.

Vestlussaadete korpus

Korpus sisaldab ca 20 tundi raadiote vestlussaadete (Rahvateenrid, Olukorrast riigis, Reporteritund, Kukkuv õun, Vastasseis, Linnatund, Välismääraja, Nädala tegija jt) salvestusi, mis kõik on käsitsi märgendatud (Transcriberi abil). Korpus on kättesaadav LAMUS-süsteemi kaudu.

Loengukõne korpus

Korpus sisaldab umbes 350 tundi eri ainevaldkondade akadeemiliste loengute salvestustusi (erinevate lektorite arv on 33) ja üle 20 tunni konverentsiettekandeid (45 isikut). Konverentsiettekannetest on märgendatud (Transcriberi abil) 24 isiku salvestused kogumahuga ca 13 tundi. Korpus on kättesaadav LAMUS-süsteemi kaudu.

Aktsendikorpus

Aktsendikorpus sisaldab eri emakeelega inimeste eestikeelse kõne salvestusi. Salvestatud on umbes 160 keelejuhi kõnematerjal, kelle keeletaust on järgmine: vene (50), soome (30 keelejuhti), saksa (15), prantsuse (12), itaalia (5), inglise (4), leedu (3), hispaania (2), taani (2), hollandi (2), slovaki (2), jaapani (2), rootsi

(1), poola (1), läti (1), šoti (1), iiri (1), aserbaidžaani (1), portugali (1), võrdlusmaterjalina on salvestatud 20 eesti emakeelega keelejuhi kõnenäited. Korpus on kättesaadav LAMUS-süsteemi kaudu.

Infrastruktuuri kaasajastamine

On välja ehitatud ja sisustatud kõnesalvestusstuudio, kõnekorpuste tarvis on paigaldatud eraldi server. Kõnekorpuste haldamiseks ja neile ligipääsu loomiseks on kohandatud Hollandis Max Planck'i Psühholingvistika Instituudis välja töötatud korpuste haldussüsteem LAMUS (Language Archive Management and Upload System, http://www.lat-mpi.eu/tools/lamus/).

Kõne analüüs ja variatiivsuse mudelid

Vastutav täitja	Einar Meister
Teised põhitäitjad	Lya Meister, Margus Muskat, Jüri Kuusik
Finantseerimine	2,6 miljonit krooni
Kestus	2006-2010

Eesmärgid

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

Põhitulemused

Kõne mikroprosoodiliste nähtuste uurimine Uuriti vokaalide omakestuse rolli vokaalikategooriate ja kontrastiivsete kestuskategooriate lühike vs. pikk eristamisel. Leiti, et (1) vastupidiselt üldlevinud seisukohale mõjutab vokaali kestus vokaalikategooria taju ka kvantiteedikeeltes (nt eesti ja soome); (2) vokaali omakestus on vokaalikategooriate eristamisel täiendavaks tunnuseks juhul kui spektraalne informatsioon ei taga piisavat pertseptiivset kontrasti; (3) vokaali omakestus mõjutab lühike/pikk kategooriapiiri taju – kõrgete vokaalide puhul tajutakse kategooriapiiri lühema stiimuli kestuse korral võrreldes madalate vokaalidega.

Tulemused on avaldatud järgmistes publikatsioonides

1. Werner, Stefan; Meister, Einar (2008). Microdurational influences on perceived vowel quality. In: The Third Baltic Conference on Human Language Technologies : Proceedings, October 4-5, Kaunas, Lithuania: (Toim.) Cermák, F.; Marcinkevicienè, R.; Rimkutè, E.; Zabarskaitè, J.. Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 2008, 335 - 342.

2. Meister, Einar; Werner, Stefan (2009). Vowel category perception affected by microdurational variations. In: Proceedings of Interspeech 2009 : [Speech and Intelligence], 6-10 September 2009, Brighton, UK: ISCA, 2009, 388 - 391.

3. Meister, Einar; Werner, Stefan (2009). Duration affects vowel perception in Estonian and Finnish. Linguistica Uralica, 45(3), 161 - 177.

Kõne makroprosoodiliste nähtuste uurimine

Loengukõne temporaalse struktuuri analüüsil klassifitseeriti ja mõõdeti segmenteeritud kõnematerjalis erinevate prosoodiliste üksuste (teema, hingamisrühm,

prosoodiline rühm) ja neile vastavate pauside keskmised kestused ning esitati loengukõne temporaalse struktuuri mudel. Tehti ettevalmistavaid töid andmekaeve meetodite rakendamiseks loengukõne korpuse temporaalse struktuuri uurimiseks – loodi trifoonide klasteranalüüsimudelid eraldi mees- ja naiskõnelejatele, viidi läbi tunnuste (kepstrikordajate) peakomponentide analüüs jm (teema jätkub J. Kuusiku doktoritöö raames).

Tulemusi on avaldatud ühes konverentsiartiklis

1. Meister, Einar; Lippus, Pärtel (2007). On temporal organization of spontaneous Estonian: preliminary analysis results of lecture speech. In: The Third Baltic Conference on Human Language Technologies 2007: The Third Baltic Conference on Human Language Technologies, Kaunas, October 4-5, 2007. Kaunas:, 2007, 28 - 28.

Aktsendiga kõne uurimine

Uuriti eesti (L1) ja vene (L2) emakeelega keelejuhtide eesti keele fonoloogiliste kategooriate – vokaalid, lühike/pikk vastandus ja välted – taju.

Vokaalikategooriate tajueksperimendid näitasid, et:

1) eesti vokaalide /i, e, u, o, a, ä/ vastenduvad üks-üheselt vastavate vene vokaalidega fonemaatilisel või allofoonilisel tasemel ja L2 ning L1 kuulajad tajuvad neid sarnaselt;

2) vokaalid /ü, ö, õ/ assimileeruvad osaliselt vene vokaaliga (/ɨ/) ja seetõttu on L2 kuulajatel raskusi nende eristamisega – kategooriapiirid on L2 katsealuste tajuruumis tunduvalt hägusamad võrreldes L1 katsealuste vastavate piiridega. Kestuskategooriate (lühike vs pikk vokaal) katsetes leiti, et L2 kuulajad tajuvad lühike/pikk kategooriapiiri pikemate vokaalikestuste korral võrreldes L1 katseisikutega.

Vältevastandusi sisaldavad katsed näitasid, et L1 katsealused kasutavad väldete eristamiseks nii kestust kui põhitooni, L2 katsealused aga eristavad välteid eelkõige kestusparameetri alusel.

Akustilise analüüsi tulemused näitasid, et erinevused L1 ja L2 rühma vokaalikategooriate, lühike/pikk ja vältevastanduste tajus tulevad esile ka katseisikute häälduses.

Tulemusi on esitatud järgmistes publikatsioonides

1. Meister, Lya; Meister, Einar (2007). Perceptual assessment of Russian-accented Estonian. In: ICPhS XVI : Proceedings of the 16th International Congress of Phonetic Sciences, 6-10 August 2007, Saarbrücken Germany: Saarbrücken: Universität des Saarlandes, 2007, 1717 - 1720.

2. Meister, Lya (2009). Eesti vokaalikategooriate piirid vene ja eesti emakeelega kõnelejate tajuruumis. In: Eesti Rakenduslingvistika Ühingu aastaraamat 5 = Estonian Papers in Applied Linguistics 5: (Toim.) Metslang, Helle; Langemets, Margit; Sepper, Maria-Maren; Argus, Reili. Tallinn: Eesti Keele Sihtasutus, 2009, 143 - 156.

3. Meister, Lya; Meister, Einar (2010). Perception of Estonian vowel categories by native and non-native speakers. In: Proceedings of INTERSPEECH 2010 Spoken Language Processing for All : 26-30 September 2010, Makuhari, Chiba, Japan: International Speech Communication Association, 2010, 1870 - 1873.

4. Meister, Lya; Meister, Einar (2011). Perception of short vs. long phonological category in Estonian by native and non-native subjects. Journal of Phonetics (avaldamisel).

5. Meister, Lya. Vene aktsent eesti keeles: taju ja akustika. Doktoritöö. (valmimisel).

Lisaks eelnevatele on käsitletud ka järgmisi alateemasid:

- automaatse kõnekvaliteedi hindamise algoritmide võrdlus: võrreldi kahe algoritmi – PESQ (Perceptual Evaluation of Speech Quality) ja 3SQM (Single Sided Speech Quality Measure) tulemusi erinevate signaalimoonutuste ja kodeerimisviiside puhul; uuriti, kuidas on erinevaid psühhoakustilisi tajumehhanisme algoritmiliselt modelleeritud ja pakuti välja võimalusi nende edasiarenduseks (koostöö Skype OÜga). Tulemused on avaldatud artiklis:

  – Muskat, Margus; Meister, Einar (2008). Quality estimation of timescale modified signals. In: The Third Baltic Conference on Human Language Technologies: Proceedings, October 4-5, Kaunas, Lithuania: (Toim.) Cermák, F.; Marcinkevicienè, R.; Rimkutè, E.; Zabarskaitè, J.. Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 2008, 197 - 204.

- vokaalide formantsageduste variatiivsus sidusas kõnes: on testitud erinevaid formantsageduste mõõtmis- ja normaliseerimisalgoritme, tulemusi on kavas avaldada 2011.

Eesti emotsionaalse kõne korpus

Vastutav täitja	Hille Pajupuu
Teised põhitäitjad	Rene Altrov, Kairi Tamuri, Ago Kuusik, Jaan Pajupuu
Finantseerimine	3,5 miljonit krooni
Kestus	2006-2010

Veebilent: http://peeter.eki.ee:5000

Projekti eesmärk

1. olla korpuspõhise emotsionaalse tekst-kõne sünteesi akustiline baas;
2. olla usaldusväärne andmekogu kõnes ja kirjas avalduvate emotsioonide uurimiseks.

Korpuse materjal

Korpus sisaldab rõõmu, viha ja kurbuse emotsiooni kandvaid lauseid ning neutraalseid laused. Laused on pärit ajakirjanduslikest tekstilõikudest, mille on korpusesse lugenud mittenäitleja (naishääl). Kontekstist eraldatud korpuslausete emotsioon on määratud kuulamis- ja lugemistestidega. Kuulamis- ja lugemistesti on läbinud ~ 1700 lauset.

Vt http://peeter.eki.ee:5000 Aruanded / Testide tulemused Emotsioon on loetud äratuntuks, kui vähemalt 51% testijatest on olnud lause emotsioonis ühte meelt.

Korpuse infrastruktuur

Korpuse veebipõhises rakenduses on kasutatud vabavara Linux, PostgreSQL, Python, Praat, Pylons. Korpus koosneb andmebaasist ja kuulamis- ja lugemistestide läbiviimise ning tulemuste statistilise hindamise vahenditest. Rakendus on installeeritav nii Windows'i kui ka Linux'i keskkonda. Veebiliides on eesti-, inglise-, soome- ja lätikeelne. Vt korpuse tehniline kirjeldus http://peeter.eki.ee:5000/docs

Päringud

Korpuses on võimalik eristada laused, kus

1. emotsiooni kannab ainult hääl;
2. emotsiooni äratundmist võib olla mõjutanud tekst.

Vt http://peeter.eki.ee:5000 Aruanded / Lausete tulemused. Korpusest saab alla laadida heli (wav), TextGrid'i (lausetasand, sõnatasand, häälikutasand, sõnaliigitasand).

Korpus on integreeritud mitme süsteemiga: Praat, kõneandmebaaside süsteem EMU, WaveSurfer ja R.

Projekti raames kaitstud magistritööd

1. Rene Altrov, teadusmagistrikraad, 2007, Emotsionaalse kõne korpuse loomine eesti keele tekst-kõne sünteesi jaoks. Tekstimaterjali evalvatsioon viha näitel, juhendajad Haldur Õim, Hille Pajupuu, Tartu Ülikool.

2. Kairi Tamuri, magistrikraad, 2007, Pausid etteloetud ilukirjandustekstis, juhendaja Hille Pajupuu,Tallinna Ülikool.

Käimasolevad doktoritööd

1. Rene Altrov, Eesti emotsionaalse kõne korpuse loomine ja emotsioonide taju, juhendajad Hille Pajupuu, Urmas Sutrop,Tartu Ülikool.

2. Kairi Tamuri, Eesti põhiemotsioonide akustiline analüüs ja modelleerimine, juhendajad Hille Pajupuu, Karl Pajusalu, Tartu Ülikool.

Korpuse kohta

1. Altrov, Rene (2008). Eesti emotsionaalse kõne korpus: teoreetilised toetuspunktid. Keel ja Kirjandus, 4, 261–271.

2. Altrov, Rene; Pajupuu, Hille (2008). The Estonian Emotional Speech Corpus: Release 1. The Third Baltic Conference on Human Language Technologies. Vilnius: Vytauto Didžiojo Universitetas; Lietuviu kalbos institutas, 9–15.

3. Altrov, Rene; Pajupuu, Hille (2010). Estonian Emotional Speech Corpus: Culture and age in selecting corpus testers. Inguna Skadina, Andrejs Vasiljevs (Eds.). Human Language Technologies — The Baltic Perspective — Proceedings of the Fourth International Conference Baltic HLT 2010. Amsterdam: IOS Press, 25–32.

4. Altrov, Rene;Pajupuu, Hille (2011, ilmumas). Estonian Emotional Speech Corpus: Content and options. R.I.L.A. - Rassegna Italiana di Linguistica Applicata, 1-2.

5. Tamuri, Kairi (2010). Kas pausid kannavad emotsiooni? Eesti Rakenduslingvistika Ühingu Aastaraamat, 6, 297–306.

Eestikeelne korpuspõhine kõnesüntees

Vastutav täitja	Meelis Mihkla
Teised põhitäitjad	Indrek Kiissel, Tõnis Nurk, Liisi Piits
Finantseerimine	4,45 miljonit krooni
Kestus	2006 - 2010

Eesmärgid ja tähtsus

Projekti eesmärgiks on keskmiste (50-100 minutit kõnet ühe keelejuhi kohta kohta) ja suurte (120-480 minutit kõnet) kõnekorpuste baasil genereerida uusi sünteeshääli test-kõne sünteesiks. Uute häälte loomiseks kasutatakse erinevaid arendussüsteeme Festival, HTS, Clustergen. Sünteeshäälte loomise kõige töömahukam osa on kõnekorpuse märgendamine ja kõneüksusteks segmenteerimine, kus rakendati automaatset segmenteerimist (Sphinx, Ehmm) järelkontrolliga ja väiksemate korpuste korral ka käsitsi segmenteerimist.

Projekti teiseks eesmärgiks oli sünteeskõne loomuliku rütmi ja kõla parandamine. Selleks modelleeriti erinevate statistiliste meetoditega (regressioon, CART, närvivõrgud) häälekõrgust ja kõne ajalist struktuuri sidusa kõne korpustel.

Korpussünteesi saab rakendada inimene–masin dialoogsüsteemide osana. Eelkõige vajavad kõnesünteesi nägemispuudega inimesed arvuti vahendusel info hankimiseks ja suhtlemiseks.

Põhitulemused

1. Kõnesünteesi tarvis on salvestatud ja märgendatud kuue keelejuhi kõnekorpused, mis kokku sisaldavad 20,9 tundi kõnet.

2. Eri meetoditel on genereeritud viis korpuspõhist sünteeshäält (Riina, Tõnu, Einar, Liisi, Tõnis).

3. Kõneprosoodia vallas on uuritud statistiliste meetoditega kõne ajalise struktuuri ja põhitooni modelleerimist ning on välja töötatud vastav metodoloogia. Uurimuste tulemusena on ilmunud seitse publikatsiooni. Prosoodia uurimine jätkub teiste sidusprojektide raames.

4. Koostöös Põhja-Eesti Pimedate Ühinguga ja Eesti Pimedate Raamatukoguga loodi eestikeelsete teabetekstide ettelugemise süsteem nägemispuudega inimestele, mille abil pimedad saavad lugeda uudiseid, ajalehti, raamatuid ja kuulata heliajakirju ning audioraamatuid.

Tartu Ülikooli suulise eesti keele korpuse projektid

Vastutav täitja	Tiit Hennoste
Teised põhitäitjad	Olga Gerassimenko, Riina Kasterpalu, Siim Orasmaa, Andriela Rääbis, Krista Mihkels.
Finantseerimine	1,88 miljonit krooni
Kestus	2006-2010

Eesti kõnekeele korpuse kogumine ja translitereerimine (kokku 2004-2008, käesolevast programmist 2006-2008).

Tartu ülikooli eesti kõnekeele audio- ja videokorpuse kogumine ja otsingutarkvara loomine (2009-2010).

Mõlema projekti keskne ülesanne oli eesti suulise keele korpuse tegemine. Esimene projekt keskendus ainult audiokorpuse kogumisele, teise projekti lisaeesmärkideks olid videokorpuse kogumise alustamine ja otsingutarkvara väljatöötamine.

Kõnekeele korpus (KK) on tegelike spontaansete dialoogide ja monoloogide suhtluskeelekorpus. Sellesse kuuluvad argi- ja institutsionaalsed (avalikud) suhtlused, monoloogid ja dialoogid, silmast-silma, telefoni- ja meediasuhtlus. Korpuse alaosa on Dialoogikorpus ehk institutsionaalsete infodialoogide korpus inimese-arvuti suhtluse modelleerimiseks (Mare Koidu projekt "Intelligentne kasutajaliides andmebaasidele").

KK osad

1. Põhiosas audiosalvestused, kokku ca 400 tundi, on alustatud videosalvestuste tegemist.

2. Verbaalse suhtluse transkriptsioonid rahvusvaheliselt kõige levinumas vestlusanalüüsi transkriptsioonis, kokku ca 1 600 000 sõna. Mitteverbaalse osa tarvis on töötatud välja omapoolne põhiskeem, kuna rahvusvaheline standard puudub.

3. Taustakirjeldused salvestatud suhtlussituatsioonide ja suhtlejate sotsiaalsete omaduste kohta. Aluseks on rühma enda väljatöötatud süsteem.

4. Päringusüsteem ehk tarkvara, mis võimaldab otsida korpusest materjali. Aluseks Jaak Vilo rühma ligikaudse otsimise süsteem. Võimaldab otsida sama sõnavormi erinevaid variante ning erinevaid situatsioonitüüpe ja erinevate sotsiaalsete tunnustega inimeste tekste.

Projektide muud tööd

- 1997-2004 kogutud analoogformaadis korpuseosa digitaliseerimine ja transkriptsioonide täpsustamine, viimaks need vastavusse transkriptsiooni praeguse seisuga.

- Täppistranskriptsiooni (maksimaalse põhjalikkusega tehtud transkriptsiooni) valdavate transkribeerijate koolitamine.
- Korpuse kogumise ja kasutamisega seotud juriidiliste probleemide lahendamine vastavuses Eesti ja Euroopa Liidu areneva seadusandlusega.
- Korpus on kasutatav teadus- ja õppe-eesmärkidel. Ta jaguneb kasutajate jaoks eri piirangutasemetega alaosadeks, aluseks tekstide ja suhtlussituatsioonide eetiline tundlikkus.
- Kasutajad peavad allkirjastama konfidentsiaalsuskohustuse ja piirama avalikult esitatavad tsitaadid kõnelejate identifitseerimist mittevõimaldava mahuni.

Eesti keele spontaanse kõne foneetiline korpus

Vastutav täitja	Pire Teras
Teised põhitäitjad	Pärtel Lippus, Tuuli Tuisk, Nele Salveste, Liis Raasik
Finantseerimine	3 miljonit krooni
Kestus	2006-2010

Eesmärgid ja tähtsus

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

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

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

Põhitulemused

Korpuses on lindistused 35 keelejuhilt (mõni keelejuht osaleb mitmes lindistuses) kogukestusega 28 tundi. Keelejuhtidest 12 on 20. aastates, 10 on 30. aastates, 8 on 40. aastates ja 5 on 50. aastates ja vanemad. Lindistustest 23 on dialoogid ja 7 monoloogid.

Lindistused märgendatakse eri märgenduskihtidel, millest peamised on järgmised: sõnad (sellel kihil leidub infot ka häälelaadi ja paralingvistiliste nähtuste kohta), häälikud, häälikustruktuurid, silbid, lausungid. Praeguse seisuga on korpuses sõna- ja häälikutasandil kokku 700 756 segmenti (sõnatasandil 177 456 ja häälikutasandil 523 300 segmenti). Muudel tasanditel on kokku 276 160 segmenti. Kõiki segmente on korpuses kokku 976 916. Põhitasanditel (st sõna- ja häälikutasandil) on märgendatud kokku umbes 26 tundi kõnet.

Valminud on veebipõhine otsingumootor1. Otsida saab sõna ortograafilist kuju (nt midagi – 237 rida); häädlust, mis on märgitud SAMPA transkriptsioonis (nt mit_vAk_vi – 59 rida); häälikustruktuure, kus konsonandile vastab "C" ja vokaalile "V" (nt CVCVCV). Päringule antakse korpusest leitud vastetena kahesekundiline lõik, milles otsitav sõna esineb. Lisaks veebis kuvatavale on võimalik vastav helilõik (wav-fail) ja/või TextGrid alla laadida

Korpuse põhjal on tehtud mitmeid uurimusi, kus on uuritud kõne struktuuri, kvantiteedi ja intonantsiooni vastastikust mõju, sõnaalgulise h, vokaalidevaheliste lühikeste klusiilide, järgsilpide e ja sagedate sõnade häädlust. Ilmunud on

kolm artiklit, valminud on kolm bakalaureuse- ja kaks magistritööd.

Lihtlause semantiline analüüs I ja II

Vastutav täitja	Haldur Õim
Teised põhitäitjad	Erki Luuk, Siim Orasmaa, Pille Taremaa, Karol Toompuu, Neeme Kahusk, Heili Orav
Finantseerimine	2,5 miljonit krooni
Kestus	2006-2010

Eesmärgiks oli luua automaatse semantilise analüüsi programm, mis antud projekti puhul analüüsib ainult lihtlauseid, s.o lauseid, kus pole alistavaid sidesõnu ega alistusseoses olevaid osalauseid. Esialgu piirasime semantilist analüüsi teatud kitsama ontoloogilise valdkonnaga. Selle projekti puhul on selleks valitud liikumisega (k.a liigutamine, asetamine jne) seotud situatsioonid. Valdkond on tähelepanu keskmes olevaid alasid teoreetilises semantikas, aga ka mitmesugustes rakendustes (nt suhtlemine robotitega). Liikumisega on vältimatult seotud ruumisuhted: füüsiline liikumine toimub alati teatud viisil füüsilises ruumis ja liikumise erinevaid viise iseloomustatakse ruumisuhete kaudu (kus, kust, kuhu, mis suunas jne).

Projekti käigus on tegeletud lause predikaadist sõltuvate argumentide semantiliste rollide määratlemisega. Loodud on nö testkorpus, mille peal on treenitud arvutiprogrammi prototüüpi – semantiliste rollide märgendaja ja järelduste tegemise programm. Et valitud valdkonnaks on liikumis- ja liigutamissündmused, siis tuvastab praegune järeldusprogramm lauses need entiteedid, mis liiguvad (erinevate predikaatide puhul on need erinevad) ja fikseerib liikuva entiteedi asukoha pärast liikumissündmust.

Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp

Vastutav täitja	Peep Küngas
Teised põhitäitjad	
Finantseerimine	
Kestus	

Projekti eesmärgiks on uurida võimalusi keeletehnoloogia kasutamiseks veebi ja süvaveebi otsingute kombineerimisel. Ühelt poolt töötatakse välja nimega üksuste ja ajaväljendite tuvastamise lahendus eesti keele jaoks ning teiselt poolt uuritakse süvaveebi allikatele ligipääsu ohjavate andmeteenuste semantiliste kirjelduste kasutamist täiendavateks päringuteks sobilike andmeteenuste tuvastamiseks ning nende tulemuste agregeerimiseks veebiotsingute tulemustega. Projekti otseseks tulemuseks on keele-, semantika- ja veebitehnoloogiatel baseeruv infootsisüsteemi prototüüp, mida saab laiendada lisaks eesti keelele ka otsingutele teistes keeltes. Projekti kaugemaks eesmärgiks on luua platvorm mitmekeelseid otsinguid toetavate infootsisüsteemide loomise lihtsustamiseks nii veebi kui süvaveebi jaoks.

Automaatne parafraaside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpuste abil

Vastutav täitja	Maarika Traat
Teised põhitäitjad	Raul Sirel, Mark Tehver
Finantseerimine	348 200 EEK
Kestus	2008-2010

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

Kirjeldatud tööriista töö põhineb joondatud paralleelkorpuste kasutamisel. Masintõlkes on selliste korpuste kasutamine väga levinud, nende kasutamine parafraaside leidmiseks on aga kaunis värske idee. Idee pärineb Chris Callison-Burchilt, kes on seda meetodit üksikasjalikult kirjeldanud oma doktoritöös (2007). Antud projekt erineb masintõlkeprojektist, kuna sisendfraasidele vastusena väljastatavaid üksikuid tõlkefraase ei kombineerita kokku erinevatest allikatest, vaid alati on tegu inimtõlkide poolt mingis projektis kasutatud tõlkevastetega. (Vigu tekib siiski päris palju, peamiseks süüdlaseks on siin korpuste automaatsel sõnatasandil joondamisel tehtud vead.)

Andmetest on meie põhiallikas olnud algusest peale Acquis Communautaire tõlkemälu DGT-TM (http://langtech.jrc.it/DGT-TM.html), aga samas oleme ka ise aktiivselt materjali juurde muretsemisega tegelenud. Tööd raskendab asjaolu, et tõlkebürood ei taha oma (ka meile väga) väärtuslikku tõlkemäludes talletatud materjali välja anda.

Projekti jooksul on arendatud tööriista koodi (algse koodi saime Chris Callison-Burchilt), seda täiendatud ning paremate tulemuste saamise eesmärgil ning eesti keele spetsiifikat arvestavalt modifitseeritud. Leitud parafraaside ja tõlgete kvaliteet paranes oluliselt, kui hakkasime kasutama sisendfraaside lemmatiseerimist ning lemmatiseeritud korpusi. Tööriistale on loodud kasutajaliides, mis asub aadressil http://ats.cs.ut.ee/parafraasid/. Hetkel on tähelepanu keskmes väljundi parandamine morfosüntaktilise informatsiooni kasutamise abil.

Projekti raames on järjekindlalt tegeletud andmete kogumise, nende sobivale kujule teisendamise ja ühtseks mitmekeelseks paralleelkorpuseks ühendamisega. Lisaks Acquis Communautaire'le oleme andmeid saanud Eesti Pangalt, tõlkebü-

roolt A-Script, Siseministeeriumilt, Kultuuriministeeriumilt, Sotsiaalministeeriumilt, Haridus- ja Teadusministeeriumilt, Keskkonnaministeeriumilt, Põllumajandusministeeriumilt, EASi Turismiarenduskeskuselt, SA Lõuna-Eesti Turismilt, Statistikaametilt, vabakutselistelt tõlkidelt ja veebilehtedelt. Lisasime ka KDE Linuxi töölaua eesti keelde tõlkimise tulemusena tekkinud inglise-eesti paralleelkorpuse. Acquis'le lisaks kogutud andmetest on hetkel nii sobivale kujule teisendatud, kui ühendkorpusele juurde lisatud 26 742 847-sõnaline korpus. Põhiliselt lisandus materjali inglise-eesti, soome-eesti, mõnevõrra ka rootsi-eesti ja saksa-eesti korpusesse ning algust sai tehtud vene-eesti korpusega, mida Acquis's ei olnud.

Elektrooniliste teatmeteoste kasutajasõbralikud päringusüsteemid

Vastutav täitja	Jaak Vilo
Teised põhitäitjad	
Finantseerimine	
Kestus	

Eesti- ja laiemalt mitmekeelse info-otsingu üks keerulisi probleeme on kirjutamisel ja kirjaviisidel esinevad variandid ja vead. Kui eeldada, et sõna otsimisel kasutatakse õiget kirjaviisi ja kõigis dokumentides on just sama esitus, siis taandub info otsimine kõige relevantsemate dokumentide leidmisele kus sees vastavad sõnad esinevad. Selliseid päringuid saab teha nn pöörd-indeksite kaudu, kus indekseeritakse kõik sõnad ning iga sõna juurde antakse dokumentide loetelu kus vastav sõna esines, koos võimalike kaaludega tähistamaks relevantsust või ka väljamärgendeid.

Paraku ei ole kõik sõnad alati ühtmoodi kirjutatud. Suuremad või väiksemad erinevused võimalikud nii kirjavigadest, alternatiivsetest kirjapiltidest, sõna keerulisest morfoloogiast, tähestike ja kooditabelite probleemidest, erinevatest kasutaja töövahenditest ja terminalidest.

Projekti peamiste uudsete tulemustena arendati välja metoodikad kuidas läbi viia ligikaudsel otsingul põhinevaid infopäringuid.

Üks meetod põhines sõna algvormi põhjal genereeritud eri käände- ja pöördekujudel, mida saab esitada reguaaravaldisena millele omakorda vastab deterministlik lõplik automaat. Sellise kõiki sõnavorme sisaldava regulaaravaldise/automaadi sobitamiseks eelnevalt mitte indekseeritud tekstile arendasime välja ligikaudse sobitamisalgoritmi mis võtab arvesse standardseid teisendusi (tähe lisamine, kustutamine, muutmine). Prototüüpi katsetati OpenOffice tarkvara pistikprogrammina.

Teine meetod ligikaudse sobitamise jaoks põhineb üldistatud teisenduskaugusel. Sel juhul antakse kõikvõimalike tähtede ja täheühendite asenduste loetelud ette koos eraldi kaaludega. Rakendus "katsetab" kõiki võimalikke lubatud teisendusi, kombineerides neid tavaliste asendustega, otsides kõige "odavamat" teisenduste hulka.

Vana kirjakeele jaoks sobivad näiteks h->hh, v->w, (vahva->vahhwa), ning venekeelse päringu jaoks ladina ja kirillitsa asendusreeglid. Inglise sõnastikust otsides häälduse järgi 'grenits' saab aga esimese vastena 'Greenwich' ning kirjapildi järgi 'greenish'.

Välja töötatud programme saab katsetada:

- http://biit.cs.ut.ee/software/
- https://biit-dev.cs.ut.ee/~orasmaa/gen_ed_test/
- https://biit-dev.cs.ut.ee/~orasmaa/ing_ligikaudne/

Raaltöödeldav eesti keel: piiratud loomuliku eesti keele mooduli prototüüp teadmushaldusplatvormidele

Vastutav täitja	Martin Luts
Teised põhitäitjad	Marius Kutateladze, Monika Saarmann, Daniel Tikkerbär
Finantseerimine	
Kestus	2010

Projekti taust

Piiratud loomulikud keeled (ingl.k. Controlled Natural Languages) on tehiskeeled millised kasutavad eesmärgipäraselt valitud alamhulka loomulikest keeltest, sealhulgas:

- sõnavarast
- morfoloogiast
- grammatikakonstruktsioonidest
- tähenduse interpretatsioonivõimalustest
- jm

Piiratud loomulikud keeled hõlbustavad:

- inimestevahelist suhtlust, nt
  - tõlkimine
  - dokumenteerimine
- inimene–masin kommunikatsiooni, nt
  - päringuid andmebaasidest
  - semantiliste vikide masinarusaadavat sisutootmist
  - veebiressursside — nt blogide — täislausetega sildistamist (tagging)
  - dokumendihaldussüsteemides dokumentide kokkuvõtete kirjutamist

Teadaolevalt on käesolevaks aastaks loodud piiratud loomulikud keeled vähemalt inglise, esperanto, prantsuse, saksa, kreeka, jaapani, mandariini, hispaania ja rootsi keelte baasil.

Projekti kaugemaleulatuv eesmärk

Projekti peamine eesmärk on luua raaltöödeldava piiratud eesti keele (RTEK) raamistik, sh järgmised raamistiku komponendid:

1. RTEK disainiprintsiibid
2. RTEK lihtsustatud grammatika ja (esimest järku loogikal põhinev) esitusviis (seotud riikliku programmi p3.2.4.1)
3. RTEK piiratud sõnavara eesti keele üldontoloogia baasil (korduvkasutatakse ja arendatakse edasi EKKTT 2009 aasta projekti http://ats.cs.ut.ee/semantika/wiki/index.php/%C3%9Cldontoloogia tulemeid)
4. semantilise vikimootori prototüüp järgmise funktsionaalsusega:
   - **4.1. viki artiklite RTEKis koostamine, so RTEK editor ja** parser (seotud riikliku programmi p2.11, p3.1.10)
   - **4.2. info otsimiseks RTEKis koostatud artiklitest, kusjuures** infopäring esitatakse RTEKis (seotud riikliku programmi p2.5, p3.1.5)
   - **4.3. RTEKis koostatud viki artiklite masintõlge inglise keelde** ja vastassuunas (seotud riikliku programmi p2.6, p3.1.6)

Projekti teisesed, projekti peamise eesmärgi elujõulisust ja jätkusuutlikkust toetavad eesmärgid on:

1. aidata kaasa RTEK kasutajate ja arendajate kogukonna tekkele
2. tõsta Eesti valmisolekut piiratud loomulike keelte valdkonnas rahvusvahelistes projektides osalemiseks
3. arendada piiratud loomulike keelte teooriat, osaleda vastavates kogukondades (nt http://cnl.wikia.com/)

Projekti 2010. aasta tegevused ja tähtajad

1. Piiratud loomuliku eesti keele kasutajate sihtrühmade ja rakendusvaldkondade kaardistamine, mille tulemusena on kirjeldatud vähemalt 3 piiratud eesti keele kasutuslugu OpenUP metoodikast lähtuvalt (tegevus 1.1, 01.06.2010).
2. Piiratud loomuliku eesti keele omaduste määratlemine ca 40-s piiratud loomulike keelte ontoloogia dimensioonis, lähtudes Wyner et al 2010 publitseeritud nn piiratud loomulike keelte manifestist (tegevus 1.2, 01.09.2010).
3. Piiratud loomulikus eesti keeles korduvkasutatavate ressursside kaardistamine, mille tulemusena tekib loetelu projektidest, teadusartiklitest, tarkvaraplatvormidest, keeleressurssidest ja piiratud loomulike keelte esitlusviisidest koos hinnanguga nende kasutatavusest piiratud loomuliku eesti keele loomiseks (tegevus 1.3, 01.12.2010).
4. Projekti käiku ja tulemeid kajastab avalik veebisait leheküljel http://www.keeletehnoloogia.ee/

Veebipõhine interaktiivne keeleõpe ja selleks vajalikud ressursid

Vastutav täitja	Kristiina Praakli
Teised põhitäitjad	Neeme Kahusk, Kadri Sõrmus, Maria Loginova, Helin Roosileht
Finantseerimine	740 000 krooni
Kestus	2008–2010

Eesmärgid ja tähtsus

Tartu Ülikooli eesti keele (võõrkeelena) osakonna õppijakeele korpus on õppijakeele kirjali-ke tekstide elektrooniline kogu, mis sisaldab Tartu Ülikoolis eesti keelt (võõrkeelena) õppivate üliõpilaste loodud eri liiki kirjalikke tekste. Õppijakeele all mõistetakse keeleõppija keelekuju sihtkeelest. Õppijate emakeelt, mille baasil teist keelt omandatakse, nimetatakse esimeseks keeleks või lähtekeeleks, õpitavat keelt sihtkeeleks (lähemalt Pool 2007: 13).

Õppijakeele korpuse loomise eesmärgiks on olnud luua andmebaas, mis pakub autentset keelematerjali õppijakeele uurimiseks ning õppematerjalide koostamiseks. Sellega võimaldab korpus uurida muu emakeelega (vene, soome, saksa, läti, hispaania, inglise) üliõpilaste kirjalikku keelekasutust eesti keeles ning tuua välja need erijooned, mis eristavad õppijakeelt emakeelena kõnelejate keelest.

Õppijakeele korpuse struktuur

Õppijakeele korpus koosneb kahest elektroonilisest andmebaasist:

paralleelkorpus 2006-2007

(põhitöörühm Raili Pool, Elle Vaimann, Ingrid Rummo). 2006. aastal alustati B1 ja B2 kirjaliku keeleoskustasemega üliõpilaste kirjalike tööde (kodukirjandid) vigaste lausete sisestamist. Paralleelkorpus koosneb juhuslikest, üksikutest lausetest. Iga normidele mittevastava lause juurde on paralleelselt sisestatud parandustega lause (kas üks või mitu parandusversiooni) ning vajadusel kommentaarid.

2008. aasta veebruari seisuga on paralleelkorpuses vigaseid lauseid 9000, neis kokku 128 000 sõna; parandusi lausetena on 9100, sõnadena 129 000. Iga vealause juurest leiab ka lause autori profiili, mis sisaldab veategija kohta peamist infot kodeeritud vormis (sugu, rahvus, emakeel, elukoht, keeleoskuse tase). Paralleelkorpuse laused (märgendamata kujul) on avalikuks tehtud (www.keeletehnoloogia.ee).

tekstikorpus 2008-2010

(põhitöörühm Kristiina Praakli, Kadri Sõrmus, Neeme Kahusk, Maria Loginova, Helin Roosileht). 2008. aastal alustati keeleainestiku kogumist uutest kogumisprintsiipidest lähtuvalt. Kogutava materjali aluseks on mitte-eestlastest

üliõpilaste (emakeel vene, soome, saksa, inglise, läti ja hispaania keel) kirjalikud tööd. Eesmärgiks on koguda terviktekste, mis võimaldavad näha ja analüüsida viga ja vea konteksti tervikuna. Tekstikorpus koosneb kindlate kriteeriumide alusel igapäevastest keeleõppesituatsioonidest kogutud tekstidest. Tekstid jagunevad tekstiliigiti järgmiselt: 1) kodukirjandid (lektüüri kokkuvõtted, analüüsid, aktuaalsed teemad) ; 2) eksamitööd (kirjalikud tekstid, mitte grammatikaülesanded); 3) tunnis kirjutatud tekstid ning 4) praktikapäevikud. Märgendatud tekstide koguarv on 570 (sõnu kokku u 210 000).

Korpusesse lisatakse täiendavalt ka üliõpilaste kirjutatud poolametlikud e-kirjad ning lõputööde sissejuhatused ning kokkuvõtted.

Iga teksti juurde on lisatud metaandmed teksti ja autori kohta (teksti liik, informandi tähis, emakeel, sugu, elukoht, keeleoskustase). Õppijakeele vigade märgendamiseks on välja töötatud detailne veamärgendussüsteem (põhiosas Kadri Sõrmus (magistritöö 2008), seda on täiendanud Kristiina Praakli ning Maria Loginova). Vealiigid märgendatakse kuue põhitasandi lõikes (ortograafia, morfoloogia, süntaks, leksika, stiil, muu). Keelevigade põhitasandite kõrval määratletakse ka keelevigade alltüübid. Vead on märgendatud käsitsi.

Põhitulemused

Valminud on veebipõhine kasutajaliides, mis võimaldab teostada õppijakeelekorpusest vea-otsinguid lähtuvalt välja töötatud kuuetasandilisest vealiigitussüsteemist. Veamärgendus-süsteem katab erinevad vealiigid ning võimaldab ühele sõnale, fraasile, lausele ning lause-osale lisada ka mitu veamärgendit. Töötab sõnade sagedusotsing, mis võimaldab leida kõik otsitava sõna esinemisjuhud ning analüüsida sõna esinemist ja kasutust kontekstis. Korpus võimaldab näha tekste ka märgendamata kujul.

Korpuse materjale on kasutatud mitmetes uurimustes (nt Raili Pooli doktoritöö 2007), artiklites ning bakalaureuse- ja magistritöödes.

VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)

Vastutav täitja	Pille Eslon
Teised põhitäitjad	
Finantseerimine	
Kestus	2008-2010

Projekti põhieesmärgid

1. olemasolevat keeletarkvara arendades luua EVKK automaatseks töötlemiseks sobivad tarkvararakendused, mis võimaldavad korpuse tekstide käsitsimärgendamiselt üle minna poolautomaatsele;

2. EVKK funktsionaalsuste laiendamine.

Tulemused

1. On loodud sõnajärje vealeidja prototüüp, mis on integreeritud EVKK-sse. Prototüübi graafiline liides on valmimisjärgus. Sõnajärje vealeidja aluseks on eesti keele süntaksianalüsaator, mis on implementeeritud korpusesse. Prototüübi puhul on tegu statistikapõhise programmiga, milles kasutatud programmeerimiskeelt Python. Programm testib sõnajärje seisukohalt oluliste lauseliikmete järgnevusi ehk järjendeid esimeses osalauses ja lihtlauses: verbi märgendid `@FMV` (finiitne verb), `@IMV` (infiniitne verb), `@FCV` (olema liitaegades ning modaalverbid ahelverbides, finiitne vorm), `@ICV` (olema liitaegades ning modaalverbid ahelverbides, infiniitne vorm), `@NEG` (verbi eitus) ja lause põhja märgendid `@SUBJ` (alus ehk subjekt), `@OBJ` (sihitis ehk objekt), `@PRD` (öeldistäide ehk predikatiiv), `@ADVL` (määrus ehk adverbiaal, sh fraasiadverbiaal). Nt: lauseliikmete järjend Internetis (`@ADVL`) on (`@FMV`) võimalik (`@PRD`) kasutada (`@SUBJ`) mitmeid (`@NN>`) teenuseid (`@OBJ`) ja sellele vastav sõnajärjemall [‘`@ADVL`’, ‘`@FMV`’, ‘`@PRD`’, ‘`@SUBJ`’, ‘`@OBJ`’]. Sama algusmärgendiga korduvad sõnajärjemallid moodustavad erineva sagedusega ilmnevaid sõnajärjemustreid, mis paigutatakse andmepuusse. Eesti keelele omaseim andmepuu sisaldab sõnajärjemustreid, mis algavad märgendiga `@SUBJ`: ’`@SUBJ`’ ’`@FMV`’ ’`@ADVL`’ ’`@ADVL` jne. Sõnajärje vealeidja prototüübi töö tulemuslikkus on esialgu 87,82%.

2. Õppijakeele sõnastiku aluseks on oletaja-lemmatiseerija (veakindel lemmatiseerija), mis võimaldab teha vormimoodustus- ja ortograafiavigade analüüsi ning automaatselt määrata õppija keeleoskustaset iseloomustavaid morfoloogilisi jooni. Sisuliselt on see ligikaudne õigekirjakorrektor, mitte interaktiivne, mis pakub kandidaate ja kasutaja peab nende hulgast ise valima. Oletaja-lemmatiseerija teeb valikud otsustuspuude abil kirjavahemärkide, mittesõnade, pärisnimede

ja ühetähenduslike sõnade jaoks. Mitmesuste lahendamisel tugineb tõenäosuste arvutamisele: kaugus on võrdne operatsioonide arvuga, mis on vajalik ühe stringi teiseks teisendamiseks; operatsioonideks on tähe lisamine, kustutamine, asendamine. Nt:

```
kantsid vs kandsid -> t asendada d-ga -> kaugus = 1;
igasugulased vs igasugused -> kustutada l ja a -> kaugus = 2.
```

Aluseks on ESTMORF ja foneetiline algoritm Metaphone, mis moodustab iga sõna jaoks kuju, milles sisaldub vaid kõige olulisem selle häälduse kohta. Sarnaselt kõlavad sõnad taandatakse ühesugusele häälduskujule. Kuna algoritm on inglise keele põhine, siis lisati eesti keele vokaalid Õ, Ä, Ö ja Ü. Eesti Ekspressi korpuse alusel (ligi 7 miljonit sõnakasutust) moodustati referentssõnastik, mida töödeldi foneetilise algoritmiga ja moodustati hulgad sarnase häälduse sõnadest. Õppijakeele vigastele sõnakasutustele leiti samuti hääldus-kuju, arvutati teisenduskaugused vigase sõna ja sama häälduskujuga sõnade vahel referentssõnastikus ning leiti suurima tõenäosusega kandidaat nende sõnade hulgast, mille teisenduskaugus on maksimaalselt 2. Oletaja-lemmatiseerija tööd võrreldi ESTMORF-i abil saadud tulemustega. Kui arvata sisse kirjavahemärkide äratundmine, siis ületas oletaja-lemmatiseerija ESTMORF-i 23,8%, kirjavahemärke arvestamata 3,2%. Suurim raskus on määrsõnade (_D_), pre- ja postpositsioonis kasutatud kaassõnade (_K_) ja verbi juurde kuuluva sõna (_X_) eristamine.

Järgnevad tegevused

Oletaja-lemmatiseerija tehtud vigade parandamine; programmi integreerimine õppijakeele korpusesse ja veebiliidese abil kättesaadavaks tegemine; oletaja-lemmatiseerija integreerimine sõnajärje vealeidjaga, mille tulemusel prototüüp analüüsib kõiki õigekirja- ja vormivigu sisaldavaid lauseid, mis avardab oluliselt prototüübi rakendusvõimalusi: nt aitab tuvastada õppijakeele morfoloogia- ja sõnajärjevigu, määrab keeleoskustasemeid iseloomustavad lingvistilised jooned.