Ülevaade

Keeletehnoloogia rakendustest eesti keeles

Haldur Õim
Tartu ülikooli emeriitprofessor
Neeme Kahusk
Tartu ülikooli teadur

Termin *keeletehnoloogia* (ingl k *language technology*) võeti kasutusele Euroopa Liidus 1990ndate alguses; varem oli üldmõisteks *arvutilingvistika* (*computational linguistics*), ingliskeelses maailmas kasutatakse suurel määral senini rakenduste puhul terminit *natural language processing* (*NLP*). Lähteks oli ELi põhimõte, et liikmesriikidel on õigus säilitada oma riigikeel kogu asjaajamises (igal juhul riigisiseselt) ehk et EL on ja jääb oma asjaajamises mitmekeelseks. See on praegusenigi nii: ükski ELi liikmesriik pole loobunud oma riigikeelest. Samas tõdeti aga, et reaalselt ei ole see võimalik, kui asjaajamine toimub vaid paberil. Igal liikmesriigi keelel peab olema keeletehnoloogiline tugi, mis võimaldab infotehnoloogilistele vahenditele tuginevat infovahetust. Mõeldud pole kaugeltki ainult masintõlget tavalises mõttes, ennekõike on vaja, et ka riigisisene keelelises vormis esitatud info oleks elektrooniliselt kättesaadav ja töödeldav ning ka vahetatav. Selleks algatati mitu europrogrammi (Danzin 1992). Eesti arvutilingvistid liitusid kogu üritusega programmi Copernicus kaudu 1994.–1995. aastal[1] ning võiks öelda, et päris jõuliselt: alates kõnetehnoloogiast, korpustest jt keeleressurssidest morfoloogia, süntaksi, semantika ja pragmaatika jm tavalise teksti töötluse aladeni.[2]

---

1 Siis ei olnud Eesti veel ELis.

2 Populaarset ajaloolist ülevaadet vt nt Õim 2009; detailsemalt on asjade kulgu kirjeldatud artiklis Koit jt 2006.

Edasise arengu mõjutajatest tuleb kindlasti märkida kaht. Esiteks, aastal 2004 käivitati haridus- ja teadusministeeriumi algatusel „Eesti keele arendamise strateegia 2004–2010" ja seejärel „Eesti keele arengukava 2011–2017", kus mõlemas on peatükk „Eesti keele keeletehnoloogiline tugi". Neis sõnastatud töid koordineerib Eesti keelenõukogu ja selle kodulehelt (ekn.hm.ee) võib leida ülevaate nii tehtust kui ka kavandatust. Teiseks tuleb mainida 2006. aastal sündinud riiklikku programmi „Eesti keele keeletehnoloogiline tugi 2006–2010", mis sai jätkuprogrammi „Eesti keeletehnoloogia 2011–2017" (www.keeletehnoloogia.ee). Muidugi ei ole keeletehnoloogilisi rakendusi tehtud ainult nende programmide raames.

Keeletehnoloogia rakendused on väga kirju ala – juba sellepärast, et keelekasutus, selle valdkonnad ja vormid (kirjalikud tekstid ja suuline suhtlus, ametlikud, ajakirjandus-, ilukirjandus- ja muud tekstid ning igasugune vahetu suhtlus) on väga mitmetahulised.

Tekstitöötlus ja rakendused

Mõttekas on teha vahet ühelt poolt lõppkasutajale mõeldud süsteemidel (olgu neiks kasutajaiks masstarbijad või mingi kitsa ala spetsialistid) ja teiselt poolt vastava tarkvara arendajatele endile mõeldud vahenditel. Viimastel on palju pikem ajalugu ja selle arengu (lühi)ülevaate kaudu saab anda mingi raampildi keeletehnoloogia arengukäigust uusimate (lõppkasutajale mõeldud) rakenduste juurde. Siinkohal peame eelkõige silmas eesti keele töötlemiseks loodud analüüsi-sünteesiprogramme: morfoloogia, süntaks, semantika (pluss leksikoloogia ja leksikograafia) ning pragmaatika.

Morfoloogiast alustades: siin pole sündinud palju uut, kuid nt praegune sõnastikupõhine morfoloogiline analüsaator-süntesaator (ESTMORF), mis on kasutusel eesti keele spelleris, poolitajas, liitsõnade analüüsis ja teistes rakendustes (vt www.filosoft.ee), aga mis on vajalik ja kasutusel ka kõigi kõrgemate keeletasandite analüüsiks, loodi juba üle kümne aasta tagasi (TÜ ja OÜ Filosoft koostöös). Seda on aga vaja pidevalt ümber teha, mitte ainult täiendamiseks ja vigade parandamiseks, vaid ka selleks, et seda kasutavad programmid üha uuenevatel riist- ja tarkvaraplatvormidel töötada saaksid. Sama kehtib eesti keele instituudis arendatava nn reeglipõhise morfoloogiaprogrammi kohta.

Süntaksianalüsaator (sellega tegeldakse TÜs) on läbi teinud keerulisema arengu. Algvariandi aluseks oli kitsenduste grammatika (töötati

välja Soomes 1980/90ndate vahetusel), mis lause morfoloogiliselt analüüsitud sõnavormide järjendit sisendina kasutades määrab lausliikmed (predikaat, subjekt, objekt, adverbiaalid).[3] See on lineaarne esitus, milles ei kajastu lausesisene fraaside hierarhia. Järgmiseks sammuks oli programm, mis teisendab selle lineaarse esituse hierarhiliseks ehk puukujuliseks[4] struktuuriks. Selliste puukujuliste struktuuridena analüüsitud lausekorpused on tuntud puudepankadena (ingl k *treebank*). Eesti keele süntaksipuude panka on arendatud koostöös Põhjamaade keeletehnoloogidega. Väärib märkimist, et 2010. a toimus Tartus üheksas rahvusvaheline puudepankade ja lingvistiliste teooriate konverents[5]. Olemasolevatest süntaksianalüüsi vahenditest saab üldise ülevaate veebilehel http://vvv.cs.ut.ee/~kaili/grammatika/.

Süntaktilise analüüsi väljund on nii lausete semantilise kui ka pragmaatilise analüüsi sisendiks, aga vajalik ka kõnesünteesis, sest sünteeskõne loomulikkus sõltub suurel määral lauseintonatsioonist, see on aga otseselt seotud lause süntaktilise liigendusega. Selle ülesande lahendamisega on eestikeelse kõnesünteesi arendajad just viimastel aastatel hoolega tegelenud. Kuid lisaks tugineb süntaksianalüsaatorile selline oluline rakendus nagu grammatikakorrektor (praegu olemasolev grammatikakorrektor tunneb ära koma- , ühildumis- ja rektsioonivigu, kuid töö on pooleli). Samuti on see kasutatav infootsingus, nt nimisõnafraaside tuvastaja aitab leida mitmesõnalisi termineid, ja tekstide sisukokkuvõtete tegemise programmi prototüübis, mis teksti (lausete) struktuuri ja sõnasageduste põhjal leiab tekstis kõige informatiivsemad laused ja esitab need sisukokkuvõttena (vt http//math.ut.ee/kaili/estsum2009/). Kõnealust süntaksianalüsaatorit on kohaldatud ka eesti suulise kõne ja murdetekstide analüüsimiseks.

Semantikaalased tööd jagunevad leksikaalse ja lausesemantika vahel. Lausesemantikaga on tegeldud mõned aastad ja olemas on küll kontseptuaalne mudel, kuid automaatset lauseanalüüsi ennast on vaid katsetatud. Põhiidee on selles, et süntaksianalüüsi väljundina saadud struktuuris tuleb kõigepealt süntaktilised kategooriad, nagu subjekt, objekt, adverbiaal, asendada semantiliste rollimõistetega, nagu agent

---

3 Aastal 2000 kaitses Kaili Müürisep TÜs doktoritöö „Eesti keele arvutigrammatika: süntaks".

4 *Puu* on graafiteooriast üle võetud süntaksitermin.

5 Ettekanded avaldati Põhja-Euroopa Keeletehnoloogia Assotsiatsiooni NEALT toimetiste sarjas.

e tegija, kogeja, vahend, põhjustaja jne; ja lisada lausest tulenevad järeldused, mis fikseerivad selle, mis lauses kirjeldatud tegevuse või sündmuse järel maailmas on teisiti. Ehk siis lausesemantikat ei huvita niivõrd lause kui keeleline struktuur, vaid lausetega kirjeldatavate tegevuste, sündmuste jne struktuur (vt nt Õim jt 2009).

Leksikaalne semantika tegeleb sõnade, aga täpsemini leksikaalsete üksuste (ja need võivad olla mitmesõnalised – vrd kas või ühend- ja väljendverbe) tähendustega. Eesti keele puhul on ilmselt tuntuim keeletehnoloogiline väljund eesti keele tesaurus EstWN (www.cl.ut.ee/ressursid/teksaurus), semantiline andmebaas, mis on üles ehitatud mõistepõhiselt: selle üksusteks on sünohulgad (samatähenduslikud sõnad), mis on omavahel süstemaatiliselt seotud semantiliste seostega, nii hierarhilistega (nt hüponüümia-hüperonüümia: *vesi – vedelik*) kui ka funktsionaalsetega (nt põhjus-tagajärg: *tapma – surema*).

Selle poolest erinebki see näiteks Eesti keele seletussõnaraamatust (www.eki.ee/).

Tesauruse koostamine oli üks esimesi töid, kus liitusime vastava europrojektiga EuroWordNet (1997) – projektiga, mille raames töötati välja mitmekeelne tesaurus, kus üksikute keelte tesaurused on seotud nn *Interlingual Index*'i (ILI) kaudu. EstWN koostamine kestab senini. *Wordnet*-tüüpi andmebaaside olulisust iseloomustab ehk kõige selgemini fakt, et neid on praeguseks üle maailma loodud kümnete keelte jaoks ning eksisteerib ka *Global WordNet Association* (http://globalwordnet.org/), mis neid ühendab. Olulisus tuleneb mitte ainult võimalusest tesaurusi juba praegu kasutada keeletehnoloogilistes rakendustes (nt lausete-tekstide semantilises analüüsis ja tõlgendamises, kus ka meie seda teeme), vaid asjaolus, et niimoodi ehitatud andmebaasis sisaldub teatud mõttes vastava keele, selle kõnelejate maailma mudel, mida saab teiste keelte omaga võrrelda. Samahästi võime samade põhimõtete järgi luua andmebaase kitsamate valdkondade mudelitena alates nt mingist looduse- või kultuurivaldkonnast ja lõpetades õlletootmistehnoloogiatega (seda on tehtud). Siis ei ole see enam keeletehnoloogia, sest keel on vaid üks osa, vaid keelevahendeid kasutav *ontoloogia*[6] ja sellega tegeleb nt *semantiline veeb*.

On ka hoopis teist tüüpi, kuid siiski sõnavaraga tegelejatele mõeldud rakendusi: eesti keele instituudis leksikograafidele loodud sõnastike koostamist ja haldamist abistavad tarkvarasüsteemid. EELex on üldine

---

[6] Infotehnoloogias on sel mõistel teine sisu kui filosoofias.

sõnastike veebipõhine haldussüsteem, mis ühendab sõnastike koostajatele ja toimetajatele vajaliku tarkvara ja keeleressursid. EELexiga on ühendatud eesti keele reeglipõhine morfoloogiatarkvara (see on välja töötatud EKIs) ja EELexi avalikus laiatarbeversioonis kakskeelsete sõnastike jaoks kasutatav kakskeelsete sõnastike põhi — Eesti-X sõnastiku (EXS) andmebaas, kus on juba olemas eesti märksõna kohta käivad andmed, nt sõnaliik, muutevormid, tähendusjaotus, näitelaused jm. Sihtkeele (tõlkevastete) info lisab uue kakskeelse sõnastiku koostaja (vt http://eelex.eki.ee).

Pragmaatika tegeleb suhtlusega ning fookuses ei ole keele vormilised struktuurid (nt laused), vaid suhtlusüksused ja suhtlust korraldavad seaduspärasused. Ehkki keeleline suhtlus toimub keeleväljendite abil ja nii tuleb ka suhtluse arvutimudelite puhul tehnilises mõttes neist alustada, piirdume siin siiski suhtlusmudeli enda ja selle mõne rakendusega. Pragmaatika on keeletehnoloogia aspektist meil Eestis keskendunud dialoogi modelleerimisele (suhtlus üldises mõttes on ju palju laiem). Sellega on tegeletud peamiselt TÜs alates vähemalt 1980. aastatest, kuid esimesed arvestatavad rakendused (nende prototüübid) on pärit selle sajandi algusest. Dialoogsüsteemi töö alus on suhtlusaktide tuvastamine (küsimus, vastus, täpsustav küsimus, selgitus jne – kokku üle saja erineva akti) ja laias mõttes suhtlusreeglid: dialoog on kahe partneri interaktsioon ja kui üks partner on kasutanud teatud suhtlusakti, siis ootab ta sellele teiselt kindlat tüüpi reaktsiooni. Mõeldud on muidugi inimese-arvuti dialooge. Dialoogi teoreetilises mudelis huvitas meid n-ö mitteühtivate huvidega partnerite dialoog, kus üks partner püüab teist panna tegema midagi, mis teda huvitab (palumine, veenmine, käskimine, ähvardamine), sest see võimaldas modelleerida ka partnerite arutluskäike. Praegu on reaalsetes rakendustes mõeldavad siiski palju piiratumate suhtluseesmärkidega dialoogid, nt infohankimisdialoogid, kus arvuti tuvastab kasutaja suhtluseesmärgi ja annab vastuseks tema andmebaasis oleva info, vajadusel esitades ise täpsustavaid küsimusi. Nii ongi loodud nt dialoogsüsteemide prototüübid Reisiagent, mis annab infot lennukite väljumisaegade kohta Tallinna lennujaamast, või Teatriagent, mis annab infot Eesti teatrite mängukavade kohta. Uuematest rakendustest võib viidata hambaravi demosüsteemile www.dialoogid.ee/hambahaldjas / (tasub täiesti proovida, sest vastaja teatab isegi igaks juhuks, et ta on arvuti) ja veel konkreetsemalt dialoogsüsteemile Zelda: www.igemeravi.ee/. Nende autoriks on TÜ arvutiteaduse instituudis doktoritöö kaitsnud Margus Treumuth.

Tekstitöötluse osa lõpetuseks lühidalt ka selle ehk kõige enam räägitud rakendusest – masintõlkest ja selle seisust eesti keeletehnoloogias. Masintõlge oli teatavasti üks esimesi rakendusi, millega hakati tegelema kohe pärast arvutite kasutuseletulekut. Peagi selgus aga, et probleem on palju keerulisem, kui algul arvati (suur kakskeelne sõnastik ja formaalsed ülekandereeglid grammatikas). Tõsisem töö soikus pikaks ajaks. Uus tõus algas 1990. aastatel, kuid juba hoopis teiste meetoditega, nn statistilise masintõlkena. Selle aluseks on paralleelkorpused, s.t korpused, mis sisaldavad kahe huvipakkuva keele tõlgitud tekste ja kus ühe keele teksti iga lause puhul on viide selle vastele (tõlkele) teises keeles (selliste vastavuste kehtestamist nimetatakse joondamiseks). Siit edasi on võimalik liikuda joondamisele fraaside ja ka sõnade tasemel. Statistiline on tõlkimine selles mõttes, et lähtekeele lauset tõlkima hakates leitakse statistiliselt selle osade (fraaside) kõige tõenäolisemad vasted sihtkeeles.[7]

Eestis (TÜs) on tegeldud statistilise eesti-inglise masintõlkega alates 2000. aastate keskelt. 2009. lülitas teatavasti ka Google eesti keele oma tõlketeenusele (http://translate.google.com). TÜ eesti-inglise masintõlke süsteemi demoversiooni (pluss võrdlust Google'iga) võib vaadata http://masintolge.ut.ee/ ja statistilise masintõlke printsiibi demo aadressil http://masintolge.ut.ee/mt-for-kids.

Kõnetehnoloogia rakendused

Kõnetehnoloogiaga tegeldakse TTÜ küberneetika instituudis (kõnesüntees ja -tuvastus) ning eesti keele instituudis (kõnesüntees). Uusi ja uudseid rakendusi on viimastel aastatel tulnud enim just selles valdkonnas, eriti kõnetuvastuses (kõnesünteesis on sisendiks tekst ja väljundiks kõne, kõnetuvastuses on sisendiks kõne ja väljundiks tekst).

Kõnetöötluse teeb keeruliseks esmajoones asjaolu, et erinevalt tekstist ei ole kõne kirjeldatav üksikute (kirja)märkide jadana. Tähtedele vastavad kõnes küll häälikud, kuid häälikute konkreetsed omadused varieeruvad oluliselt sõltuvalt naaberhäälikutest. Näiteks teame, et *m* on heliline häälik, aga sõnas *lehm* hääldub ta helituna. Selliseid detaile tuleb kõnesünteesis arvestada. Kõnesüntees ei saa seisneda ka sõnade üksiktähtede kaupa hääldamises, kõne on pidev. Lisandub ka

---

[7] Vt lähemalt nt Kaalep, Heiki-Jaan; Mare Koit 2010. Kuidas masin tõlgib. – Keel ja Kirjandus, 10.

prosoodia – nt sõnades rõhulised-rõhuta silbid, lausetes intonatsioon, seega peab kõnesüntesaator tundma nii sõnade kui ka lausete struktuuri. Kõnetuvastus on veelgi keerulisem ja tehnilistest lahendustest sõltuvam, sest füüsikaliselt on kõne ju lihtsalt akustiline signaal. Tõsi küll, väga spetsiifilise struktuuriga signaal, mille iseärasusi hakati vastavate aparaatidega uurima juba 1950ndatel.[8] Järgnevalt on tutvustatud ainult viimaste aastate olulisemaid rakendusi.

Kõnesüntees: eestikeelne kõnesüntesaator valmis juba aastal 2002 ja selle põhitegijatele Einar Meistrile (Küberneetika instituut), Meelis Mihklale ja Arvo Eegile (EKI) ning Heiki-Jaan Kaalepile (OÜ Filosoft) anti 2003. a Eesti Vabariigi teaduspreemia. Üks uusimaid rakendusi on eestikeelsete elektrooniliste tekstide lugemissüsteem (http://elte.eki.ee/), mis on mõeldud nägemispuudega inimestele ning mille abil saavad nad lugeda uudiseid, ajalehti, raamatuid ning kuulata heliajakirju ja heliraamatuid (loodud 2009–2010).

Loomisel on subtiitrite ettelugemissüsteem nägemispuudega inimestele, kes telepilti näevad, aga teksti ei näe või ei suuda lugeda. Eesti keele instituudi, rahvusringhäälingu ja pimedate liidu koostöös on käimas esimesed katsetused. Süsteem peaks valmima aastal 2012.

Kõnetuvastus: siin on just viimastel aastatel tulnud hulk uudseid rakendusi, mis on ka arusaadav, sest kõnetuvastussüsteemidena on enamik neist kasutatavad nt mobiiltelefonides.

Näiteks **Kõnele** (2011) – Reaalajalist kõnetuvastust võimaldav suure sõnavaraga rakendus Androidi platvormil. Saadaval tasuta *Android marketis*: (https://market.android.com/details?id=ee.ioc.phon.android.speak);

Arvuta (2011) – Rakendus Androidi platvormile, millega saab suuliselt suheldes teha aritmeetilisi tehteid, mõõtühikute teisendusi, valuutateisendusi, kaardipäringuid. Saadaval tasuta *Android marketis*: https://market.android.com/details?id=ee.ioc.phon.android.arvutaja;

---

[8] Kõnetehnoloogia tehnilisest poolest või lähemalt lugeda asjatundjate kirjutistest (nt ilmus Küberneetika instituudi 50. juubelile pühendatud Horisondi 2010. aasta 5. numbris Einar Meistri ja Tanel Alumäe pikk kirjutis „Kuidas arvuti kuulab ja kõneleb").

Diktofon (2011) – Rakendus Androidi platvormile, millega saab salvestada pikemaid kõnelõike ning lasta neid kõnetuvastustehnoloogia abil automaatselt transkribeerida, saadaval tasuta *Android marketis*: https://market.android.com/details?id=kaljurand_at_gmail_dot_com.diktofon.

Kõnesalvestuste brauser http://bark.phon.ioc.ee/tsab (2010) – Võimaldab sirvida raadiosaadete transkriptsioone, kuulata vestlussaadete salvestusi koos sünkroonsete transkriptsioonidega, samuti otsida märksõnade abil infot raadiosaadete arhiividest, leida sisult sarnaseid saateid. Lihtsalt rakendatav teistes valdkondades (nt loengud, ettekanded).

Kõnetuvastuse prototüüp radioloogia jaoks (2010; info: TTÜ Küberneetika Instituut ja AS Cybernetica). Võimaldab radioloogil arvutisse dikteerida parajasti vaadatavate röntgenpiltide kirjeldusi. Selline rakendus on olemas umbes 15 keeles, sealhulgas soome keeles, arendajaks on olnud Philips.

Kokkuvõtteks võib öelda, et eesti keele keeletehnoloogiline tugi pole sugugi nõrguke. Nagu leiab „Eesti keele arengukava 2011–2017" leheküljelt 30, võib eesti keele paigutada maailmas esimese 50 kõrgelt arenenud keeletehnoloogiaga keele hulka (ja maailmas on umbkaudu 6000 keelt).

Lõpuks viimane hea uudis: 1. märtsil toimus Eesti Keeleressursside Keskuse pidulik avaüritus (ettevalmistustööd keskuse töö alustamiseks käisid juba pikemat aega, vt www.keeletehnoloogia.ee/projektid/eesti-keeleressursside-keskus). Keskus koondab eesti keele digitaalsed keeleressursid (tarkvara, andmebaasid, teksti- ja kõnekorpused, sõnastikud jne), hakkab tegelema nende dokumenteerimise ja säilitamisega ning pakkuma kasutajatele veebipõhist teenust.

Viiteid ja kirjandust edasilugemiseks

Danzin 1992 = Towards European Language Infrastructure. Report by A. Danzin and the Strategic Planning Study Group for the Commission of the European Communities . 1992.

Koit, Mare; **Tuuli Roosmaa**; **Haldur Õim** 2006. Keeletehnoloogia suundumusi: Eesti kuulub Euroopasse. Keel ja Kirjandus, 12, lk 988–992.

Kaalep, Heiki-Jaan; **Mare Koit** 2010. Kuidas masin tõlgib. Keel ja Kirjandus, 10, lk 726–738.

Õim, Haldur 2009. Filoloogi mälestused sellest, kuidas eesti keel ja arvuti Tartus kokku said. Pool sajandit arvutit Tartu Ülikoolis. Tartu Ülikool, matemaatika-informaatikateaduskond, lk 87–95.
Õim, Haldur; **Heili Orav**; **Piia Taremaa** 2009. Lihtlause semantika: teoreetiline kontseptsioon ja arvutianalüüsi võimalused. Keel ja Kirjandus, 7, lk 489–504.