Mis on arvutilingvistika?

Lilian Ariva
Liina Eskor
Tartu ülikooli üldkeeleteaduse magistrandid

Arvutilingvistika (inglise *computer linguistics*, *computational linguistics*) on teadusharu, mis lihtsalt öeldes tegeleb sellega, kuidas arvutile (inim)keelt selgeks õpetada. Nii seisabki see teadusharu keele- ja arvutiteaduse vahel ning sellega tegelejad peavad hästi tundma nii keelt kui ka arvuteid.

Et teha keelt arvutile arusaadavaks, tuleb talle selgitada keele struktuuri: keel koosneb eri liiki sõnadest (nt nimi-, tegu-, omadus-, määr- ja sidesõnad), millest on reeglite abil võimalik kokku panna lauseid. Lausete moodustamisel mängivad tähtsat rolli sõnade grammatilised vormid, nt käänded ja pöörded. Võiks ju lihtsalt sõnaraamatu koos grammatiliste vormide ja nende tähendustega arvutisse sisestada, kuid sellest siiski ei piisa, et arvuti mõistaks inimkeelt samamoodi nagu meie.

Sama sõna võib eri kontekstides tähistada hoopis erinevaid asju (homonüümid) ja vastupidi, eri sõnad võivad tähendada samu asju (sünonüümid). Peale selle on kirjalikes tekstides alati mingi osa infot puudu – kui kirjutaja ja lugeja on ühesuguse tausta ja teadmistega, ei ole elementaarseid asju vaja kirja panna, vaid puuduv info mõeldakse lugedes alateadlikult tekstile juurde. Arvutile keelt õpetades tuleks aga kogu selline maailmateadmus alguses ikkagi selgeks teha, sest temal elukogemus ja taustateadmised puuduvad.

Keeletehnoloogia

Arvutilingvistika ajalugu ulatub 1950. aastatesse, kui hakati tõsist huvi tundma masintõlke võimalikkuse vastu. Alates sellest ajast on arvutilingvistika areng läbi elanud nii tõuse kui ka mõõnu, kuid mida aeg edasi, seda kindlamaks muutub arusaam, et arvutite kasutamine keele töötlemisel on õige ja vajalik.

Arvutilingvistikal on nii teoreetiline kui ka rakenduslik külg. Teoreetilise komponendi moodustab inimese keeleliste võimete teooriate loomine ja kontrollimine. Rakendusliku osa eesmärk on luua ja arendada keele automaattöötluseks vajalikku tarkvara. Rakenduslik pool hõlmab nii keeleressursse (konkreetse keele materjal, mida saab kasutada keelest sõltuva arvutitarkvara loomiseks) kui ka konkreetseid keeletöötlusvahendeid (arvutitarkvara). Rakendusteks on näiteks masintõlge, suhtlus arvutiga loomulikus keeles, kõneanalüüs, -süntees. Kõik rakendused saab kokku võtta terminiga *keeletehnoloogia*. Konkreetse keele jaoks loodavad keeletehnoloogilised ressursid ja vahendid peaksid tagama keele konkurentsivõimelisuse infotehnoloogilises keskkonnas. Lisaks on keeletehnoloogia abil võimalik kaasa aidata puuetega inimeste suhtlus- ja tööhõiveprobleemide lahendamisele.

Korpus

Keeleressursside alla kuuluvad (teksti)korpused ja leksikonid. Korpuseks nimetatakse arvutisse viidud tekstikogumit, mis on kindlal viisil märgendatud ehk siis tekstidele on lisatud mingit tüüpi infot.

Eesti keele baasil on loodud järgmised korpused:

- tänapäeva eesti kirjakeele korpus jt kirjakeele korpused,
- suulise kõne korpus ning selle põhjal loodud dialoogikorpus,
- vana kirjakeele korpus,
- eesti murrete korpus,
- püsiühendite andmebaas.

Tallinna tehnikaülikooli küberneetika instituudi foneetika ja kõnetehnoloogia laboris koostatakse eestikeelse kõne andmebaasi, mille eesmärk on saada nii kõnekorpus kui ka selle põhjal tehtud tekstikorpus.

Olemasolevad eesti keele korpused on korpuslingvistika mõistes väikesed, nt eesti kirjakeele 1980. aastate korpus sisaldab 1 miljon sõna. Praegune suundumus on suurte korpuste loomine ning seda tehakse ka Eestis. Tartu ülikoolis on käsil projekt „Eesti keele segakorpus", mille eesmärk on luua piisavalt suur, umbes 200 miljoni sõnaga tekstihulk.

Märgendamine on vajalik, et teha korpus paremini kasutatavaks. Kuigi on olemas ka märgendamata tekstikogusid (tekstoteegid), on nende kasutusvõimalused piiratud just märgenduse puudumise tõttu.

Märgendatud ja esindusliku korpuse põhjal saab uurida erisuguseid keelenähtusi. Kuna analüüsitava materjali maht on suur ja tasakaalustatud, aitab see kaasa objektiivsete uurimistulemuste saavutamisele. Korpuses olevaid tekste märgendatakse käsitsi, automaatselt arvutiprogrammi abil või n-ö segatehnikas – automaatse märgenduse kontrollib üle inimene. Tekste saab märgendada ühel või mitmel järgnevaist tasanditest:

- tehniline – eraldatakse tekstiosad: pealkirjad, laused, lõigud, fraasid jms ning lisaks nähtused, mis võivad käituda tavalistest sõnadest erinevalt (pärisnimed, lühendid, numbrid jms);
- ortograafiline – määratakse kindlaks nt punkti funktsioon (lause lõpus, lühendites);
- foneetiline – kasutatakse spetsiaalseid tähestikke ja transkriptsioonireegleid suulise kõne korpustes kõneuuringute ja kõnetehnoloogia arendamiseks;
- prosoodiline – kasutatakse suulise kõne korpustes rõhkude, intonatsiooni, pauside jms märgendamiseks;
- morfoloogiline – märgendatakse iga sõna sõnaliik ja sõnavorm. Morfoloogiline märgendamine on korpustes kõige levinum märgendustasand;
- süntaktiline – tekstile lisatakse süntaktiline info;
- semantiline – märgendatakse semantilisi suhteid või sõnade tähenduslikku kuuluvust;
- diskursuslik – igasugune märgendamine, mis tegeleb lause tasandist kõrgemate nähtustega.

Eestis tegeldakse praegu palju korpuste märgendamisega, loodud on mitmesuguseid programme.

Suulise keele töötlus

Kuna inimkeel on suures osas suuline, siis moodustab suulise keele töötlus keeletehnoloogias väga olulise osa ning pakub ka palju tehnoloogilisi rakendusi. Eestis tegeldakse suulise kõne uurimise ja töötlemisega Tallinna tehnikaülikooli küberneetika instituudi foneetika ja kõnetehnoloogia laboris.

Suulise keele töötluse saab jagada kolmeks suuremaks valdkonnaks:

- kõnesüntees,
- kõnetuvastus,
- kõnelejatuvastus.

Kõnesünteesi eesmärk on teisendada ortograafiline tekst (nt tekstifail arvutis) loomuliku kõlaga kõneks. Kõnesüntesaatori loomiseks tuleb uurida tavalist suulist suhtlust ning teha kõnelemise eripärad selgeks ka arvutile, vajalik on meloodiakontuuri ja kõnesignaali genereerimine. Eesti keele jaoks on loodud kõnesüntesaator, mille üks eesmärke on aidata pimedatel arvutifaile lugeda (vt http://www.phon.ioc.ee/access/). 2003. aastal said Meelis Mihkla, Arvo Eek, Einar Meister ja Heiki-Jaan Kaalep riigi teaduspreemia tehnikateaduste alal töö „Eesti keele tekst-kõne süntees" eest.

Kõnetuvastuse ülesanne on teisendada tekstiks kõnesignaal, mis on mikrofoni kaudu arvutisse sisestatud. Lõpptulemuseks võivadki olla tuvastatud sõnad (nt programmides, millele saab käske anda suuliselt), kuid saadud sõnad võivad olla ka sisendiks programmile, mille ülesanne on kõne mõistmine. Kõnetuvastust saab rakendada kontoritarkvaras (nt programmide käivitamine ja juhtimine suuliste käskudega), infootsingul (nt mobiiltelefoni telefoniraamatust nime otsimine ja numbri valimine), dikteerimisel (kirjade jt dokumentide loomine) jms.

Eesti keele jaoks on seni loodud tuvastussüsteeme, mis sisaldavad väikesemahulisi sõnastikke (nt arvude tuvastamine).

Kõnelejatuvastuse käigus tehakse kindlaks kõneleja isik. Kõnelejatuvastuses on kaks põhilist ülesannet:

- kõneleja identifitseerimine – tundmatule kõnenäitele (nt telefonikõne) otsitakse võrdlusmaterjali hulgast kõige sobivam vaste. Sellise tööga tegeleb nt politsei hääleekspert;
- kõneleja verifitseerimine – registreeritud isiku kõnenäidet võrreldakse tema varem salvestatud mudeliga. Verifitseerimist saab kasutada turvameetmena andmetele juurdepääsu reguleerimisel, pangatoimingute tegemisel jm.

Suulise kõne töötluse ja tehnoloogiliste rakenduste loomise teeb keeruliseks see, et keel varieerub olenevalt keelekasutajast – kõnet mõjutavad kõneleja vanus ja sugu, emotsioonid, tervislik ja sotsiaalne seisund, elukoht jne. Nii et peale kõikvõimalike ebaselguste ja aru-

saamatuste semantilisel ja pragmaatilisel tasandil tuleb siin arvestada ka kõneleja foneetiliste ja prosoodiliste eripäradega (kirjalike tekstide puhul seda probleemi pole).

Morfoloogiline analüüs ja süntees

Morfoloogilise analüüsi ja sünteesiga tegeldakse nii Tartu ülikoolis kui ka eesti keele instituudis (EKI). EKI uurimistöö käsitleb avatud morfoloogiamudelit: kõik produktiivsed ja regulaarsed nähtused kirjeldatakse ja lahendatakse aktiivse morfoloogia reeglite abil, sõnastikus esitatakse vaid erandid. Tundmatud sõnad püütakse analüüsida tüübituvastusreeglite abil.

Tartu ülikoolis on välja töötatud morfoloogilise analüüsi programm ESTMORF, mis on sõnastikupõhine (kasutatav Internetis aadressil http://www.filosoft.ee/html_morf_et/). Peale selle on Tartu ülikoolis loodud kahetasemelise mudeli rakendus eesti keele jaoks. Mudelit saab eesti keele morfoloogia analüüsimisel kasutada, kuid kirjeldus läheb kohati väga keeruliseks. Kahetasemelise mudeli erinevus kahest eelnevast seisneb selles, et paralleelselt on vaatluse all sõnavormi kaks esitust: pindesitus (ehk lihtsalt kirjapilt) ja süvaesitus (ehk sõnastikuesitus).

Näiteks sõnavormi *künkail* süva- ja pindesitus:

```
süvaesitus: k ü n K a S + i + l #
pindesitus: k ü n k a 0 0 i 0 l 0
```

Näites viitab **+** käändelõpu või tunnuse järgnemisele ning **#** tähistab sõnavormi lõppu. Pindesituses on seda tüüpi sümbolite kohal 0, et säilitada teineteisele vastavate süva- ja pindsümbolite kohakuti asetsemine.

ESTMORFis antakse sõnale analüüsi käigus kõik tema võimalikud tõlgendused kujul

```
<sõna>
<tüvi>+<lõpp> //<sõnaliik> <vormi nimetus> //
```

Näiteks sõna *mees*

```
mees
mees+0 //_S_ sg n, //
mesi+s //_S_ sg in, //
```

Selles näites on sõnal *mees* kaks tõlgendust. Seda saab välja lugeda sõnatüvedest ja sõnade järel olevatest märgenditest: *S* tähistab nimisõna, *sg* singulari (ainsus), *n* nominatiivi (nimetav kääne) ja *in* inessiivi (seesütlev kääne).

Ent kui sõna on lauses, on tal tavaliselt vaid üks õige tõlgendus. Selle tähenduse leidmist nimetatakse ühestamiseks. Ühestada saab nii käsitsi, automaatselt kui ka mõlemat varianti kombineerides. Vaatame näiteks lauset *Mees läks metsa pesuväel*.

```
Mees
        mees+0 //_S_ sg n, //
        mesi+s //_S_ sg in, //
läks
        mine+s //_V_ s, //
metsa
        mets+0 //_S_ adt, sg p, //
        mets+0 //_S_ sg g, //
pesuväel
        pesu_vägi+l //_S_ sg ad, //
```

Ühestatuna:

```
Mees
        mees+0 //_S_ sg n, //
läks
        mine+s //_V_ s, //
metsa
        mets+0 //_S_ adt, sg p, //
pesuväel
        pesu_väel+0 //_D_ //
```

Nagu näha, on viimane sõna saanud hoopis uued märgendid, mille on lisanud inimene. Põhjuseks see, et programm on eestikeelsest sõnast *pesuväel* valesti aru saanud. Seepärast peavadki inimesed arvuti poolt ühestatud teksti kontrollima. Kui sõna on saanud vaid ühe tähenduse, ei tähenda see veel, et ühestamine on korralikult tehtud. Sageli tuleb (nagu siinses näites) märgendid hoopis eemaldada ning lisada täiesti uued.

Süntaktiline analüüs

Keeleanalüüsi järgmine etapp on süntaktiline analüüs. Kasutades morfoloogilise ühestaja väljundit, hakatakse määrama sõnade süntaktilisi funktsioone (st missuguse lauseliikmena esineb sõna lauses). Süntaktiliselt on analüüsitud osalause *kus kasteheinas põlvini me lapsed jooksime* Lydia Koidula luuletusest „Kodu".

```
kus
   kus+0 //_D_ // **CLB  @ADVL
kasteheinas
   kaste_hein+s //_S_ com sg in  //  @ADVL @NN>
põlvini
   põlv+ini //_S_ com pl term  //  @ADVL @<NN
me
   mina+0 //_P_ pers ps1 pl gen //  @NN>
lapsed
   laps+d //_S_ com pl nom  //  @SUBJ
jooksime
   jooks+ime //_V_ main indic impf ps1 pl ps af  #FinV //  @+FMV
   joo+ksime //_V_ main cond pres ps1 pl ps af  #FinV //  @+FMV
```

Siin tähistab CLB osalause piiri. Pärast @-märki esitatakse sõna süntaktiline funktsioon: ADVL on määrus (adverbiaal), <NN viitab järeltäiendile (põhisõna on eespool), NN> eestäiendile, FMV on öeldis (finiitne põhiverb), SUBJ on lause alus. Morfoloogilised märgendid pole hetkel olulised.

Jooksime on saanud kaks tõlgendust. Esimene tähendus on meile tuttav *jooksime* – mitmuse esimese pöörde lihtminevik sõnast *jooksma*. Teiseks tähenduseks on aga tingiva kõneviisi olevik sõnast *jooma*, mis enamikul lugejail jääb märkamata just kontekstitingimustega arvestamise ning maailmateadmuse olemasolu tõttu. Üks põhjusi, miks teine tähendus märkamata jääb, on kindlasti ka Lydia Koidula kui autori isiksus – mõne tänapäeva popluuletaja puhul oleks *jooksime* verbi *jooma* vormiks tõlgendamine võib-olla mõeldav.

Tegemist on mitmesusega, mis tuleks lahendada morfoloogilise ühestamise etapis. Mitmesus on aga alles jäänud semantilise ja pragmaatilise info puudumise tõttu. Kui süntaktiline ühestaja väljastaks vaid süntaktilisi märgendeid ning jätaks välja toomata morfoloogilise informatsiooni, mille põhjal süntaktiline väljund on saadud,

saaks sõna *jooksime* siin märgendi @+FMV ning mitmesust ei esineks.

Eesti keele puhul tehakse nii süntaksil põhinevat morfoloogilist ühestamist kui ka süntaktilist analüüsi ja ühestamist kitsenduste grammatika abil. Kitsenduste grammatika idee on selles, et mitmesuse puhul hakatakse kõrvaldama mittesobivaid struktuurseid ühendeid, kitsendused esitatakse vastavate reeglitega. Näiteks: kui lause öeldis on umbisikulises tegumoes, siis võib sellest lausest kõrvaldada kõik @+SUBJ (aluse) märgendid. Analüsaatoris on reegleid vormistatud programmina, mida iseloomustabki see, et algul lisatakse sõnale kõik võimalikud märgendid, millest osa hiljem reeglite abil kustutatakse (ära võetakse kõik vasted, mis sellesse lausekonteksti ei sobi, ning alles jääv tähendus peaks olema õige).

Morfoloogiliseks ühestamiseks kasutatakse ka statistilisi meetodeid. Eesti keele jaoks on loodud ka Markovi peitmudelil põhinev süsteem, mille idee on, et alles jäetakse statistiliselt kõige tõenäolisem vaste.

Semantika

Süntaktiline ühestamine on seotud **semantikaga** (tähendusõpetus). Eesti keele süntaktilist ühestamist teeb kitsenduste grammatika programm ESTKG. Programmi õigsus on 90%, semantika kaasamine parandaks olukorda. Semantikaga tegeldakse Tartu ülikoolis samuti: loomisel on eesti keele semantiline andmebaas ehk tesaurus (http://www.cl.ut.ee/ee/ressursid/teksaurus.html), millesse koondatakse eesti keele sõnade ja väljendite tähendused ja tähendussuhted (hüpo- ja hüperonüümia[1], antonüümia, osa-terviku suhted, põhjuslikkus- ja rollisuhted, tuletus- ja gradatsioonisuhted jms, suhteid on kokku umbes 60). Praegu on tesaurusesse koondatud umbes 15 000 mõistet (tähendust). Tesauruses saab esitada veebipäringuid, mille vastuseks tuuakse sõna hüperonüümid (kõrgema, üldisema tasandi mõisted), sünohulk (sünonüümirida, mille moodustavad ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid), seletus (ja näide). Nt sõna *mets* saab ühe tähenduse:

[1] Hüponüümiaga tähistatakse tähenduste hierarhilisi alistussuhteid. Alammõiste sõna on oma ülemmõiste suhtes hüponüüm, ülemmõiste sõna on oma alammõiste suhtes hüperonüüm.

Hüperonüüm(id)	Sünohulk	Seletus
taimkate 1, floora 1, taimestik 1	**mets 1**	maastiku osa ja taimekooslus, mis on kujunenud puude koos kasvades

ent sõnal *minema* on üheksa tähendust.

Tesaurus on vajalik ka teiste keeletöötlusvahendite loomiseks. Nii kasutatakse tesaurust sõnade semantiliseks ühestamiseks tekstis (programm *semyhe*): nt *pank* kui rahandusasutus ja kui looduslik maastikumoodustis.

Pragmaatika

Et arvuti inimkeelt täielikult mõistaks, tuleks talle selgeks õpetada ka **pragmaatika** – keele(nähtuste) kasutamine ja tõlgendamine olenevalt kontekstist, keelevälistest situatiivsetest teguritest. Ent see on juba üliraske, kui mitte võimatu ülesanne. Esiteks puudub arvutil kogemus kui selline. Kui mingil müstilisel moel õnnestuks seegi probleem lahendada, jääks alles veel teinegi – nimelt on inimesele omane mingi n-ö seitsmes meel, mida ei suudeta defineeridagi, rääkimata kirjeldamisest, mis oleks vajalik selle arvutil modelleerimiseks.

Pragmaatika tasandil tegeldakse Eestis dialoogide uurimisega (Tartu ülikoolis). Ka siin kasutatakse uurimismaterjalina tekstikorpust, sedakorda suulise kõne korpuse baasil loodud dialoogikorpust, mis sisaldab inimestevahelisi infodialooge (nt kõned infotelefonile). Suulise kõne uurimisrühmas analüüsitakse infodialoogide struktuuri, lausungite tähendusi ja ülesandeid ning püütakse luua mudel, mis kirjeldaks tüüpilist suulist dialoogi. Sellist tegevust nimetatakse dialoogi modelleerimiseks. Uurimise kaugem eesmärk on luua arvutisüsteem, mis vastaks loomulikus keeles inimese küsimustele (nt info bussiaegade kohta), kusjuures arvuti peab suutma täpsustada kliendi küsimusi (kui need on liiga üldised), ära tundma ja lahendama suhtluse käigus tekkivaid arusaamatusi jms.

Tõenäoliselt ei saagi arvutile inimkeelt päris selgeks teha. Kuidas olekski see võimalik, kui inimesed ei saa tihtilugu üksteisestki aru. Oluline on, et arvuti suudaks ning inimene tahaks ja oskaks omavahelises suhtluses tekkivaid arusaamatusi lahendada.

Olemasolevate keeletehnoloogiliste vahendite baasil on eesti keele jaoks loodud järgmised vahendid:

- MS Office'i koosseisus speller (õigekirja korrektor), poolitaja, tesaurus;
- optiline tekstituvastus (paberil esitatud tekst teisendatakse automaatselt elektrooniliseks);
- kõnesüntees;
- 2 masintõlkeprogrammi (eesti ja vene keel) ja umbes 10 elektroonilist sõnaraamatut.

Lähituleviku ülesanneteks on järgmiste vahendite loomine:

- automaatne diktofon – suuline tekst esitatakse ortograafilise tekstina;
- masintõlkesüsteemid Internetis;
- infootsiprogrammid, mis vastavad loomulikus eesti keeles esitatud küsimustele;
- eesti keele grammatika korrektor;
- eestikeelse kõnesisendi ja kõneväljundiga automaatsed süsteemid (nt e-postitaja pimedatele).

Arvutilingvistika õppimine

Eestis saab arvutilingvistikat õppida omaette erialana ainult Tartu ülikoolis. Kuna arvutilingvistika on hübriidala, kuuluvad õppekavasse keeleteaduse, matemaatika, informaatika ja arvutilingvistika ained. Uue õppekava (3+2) järgi õpetatakse välja kaht liiki spetsialiste: filosoofiateaduskonnas eesti ja soome-ugri keeleteaduse osakonnas saab spetsialiseeruda arvutilingvistikale ning matemaatika-informaatikateaduskonnas informaatika erialal keeletehnoloogiale.

Eestis kaitstud doktoritööd

Heiki-Jaan Kaalep, Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. Dissertationes Philologiae Estonicae Universitatis Tartuensis 7. Tartu Ülikooli Kirjastus. Tartu, 1999.

Kaili Müürisep, Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu Ülikooli Kirjastus. Tartu, 2000.

Tiina Puolakainen, Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu Ülikooli Kirjastus. Tartu, 2001.

Einar Meister, Promoting Estonian Speech Technology: from Resources to Prototypes. Dissertationes Linguisticae Universitatis Tartuensis 4. Tartu Ülikooli Kirjastus. Tartu, 2003.

Soovitusi edasilugemiseks

A&A 2002, nr 5 (keeletehnoloogia erinumber).
Arvutimaailm 2002, nr 8; 2003, nr 4; 2003, nr 6.
Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toimetaja **T. Hennoste**. Tartu, 2000.
Eesti keeletehnoloogia arenduskava.
http://www.eki.ee/keeletehnoloogia/-tutvustus/arenduskava.html.
Keel ja Kirjandus 1998, nr 1 (arvutilingvistika erinumber).
K. Muischnek, H. Orav, H.-J. Kaalep, H. Õim, Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Haridus- ja Teadusministeerium. Eesti keelenõukogu. Eesti Keele Sihtasutus. Tallinn, 2003.
T. Roosmaa, M. Koit, K. Muischnek, K. Müürisep, T. Puolakainen, H. Uibo, Eesti keele formaalne grammatika. Tartu Ülikooli Kirjastus. Tartu, 2001.
Tartu ülikooli arvutilingvistika uurimisrühma koduleht. www.cl.ut.ee.