EESTI KEEL TEHNOLOOGIATE MÕJUTUSES

MEELIS MIHKLA

Eesti Keele Instituudi vanemteadur

Rääkivatest ja mõtlevatest masinatest on inimkond unistanud juba aegade hämarusest peale. „Seesam, avane!"-fenomen on ilmselt meie kõigi alateadvuses vähemal või suuremal määral peidus. Teadaolevalt esimesena formuleeris teaduslikult sellise unistuse Šveitsi matemaatik ja insener Leonhard Euler 1761 aastal: „Muidugi oleks märkimisväärne, kui leiutataks masin, mis on võimeline järele tegema meie kõnet koos heli ja hääldusega. Ma arvan, et see polegi võimatu."[1] Eestis nii pikka teadusmõtte ajalugu kõnetehnoloogias pole ette näidata. Küllap on ka eestlased unistanud nii rääkivatest masinatest kui ka intellektuaalsetest robotitest pikka aega, paraku pole neid varasemaid unistusi fikseeritud. Kõnesünteesist ja -tuvastusest kui keele iseäralikust võlust saame siinmail rääkida pisut vähem kui poole sajandi ulatuses. Kui XX sajandi viimasel veerandil realiseerusid rääkivad ja kõnest arusaavad masinad näidiseksemplaride ja üksikute rakenduste näol, siis XXI sajandi esimesel veerandil tungivad need kõnetehnoloogilise arengu produktid meie igapäevaellu. Klaviatuurita arvutid, kõnest arusaavad ja rääkivad kodumasinad muutuvad lähikümnendi jooksul meie paratamatuteks kaaslasteks. Kas ja kuivõrd on eesti ühiskond nendeks muutusteks valmis? Milliseks kujuneb eesti keele staatus infoühiskonnas ja kuidas mõjutada tehnoloogia arengust tingitud muutusi keelekasutuses?

Praegune seis

Maailmas on praegu umbes 6000 keelt. Kõnelejate arvu poolest on eesti keel esimese veerandt[1]uhande piirimail, kuid kasutuse ulatuselt tunduvalt eespool. Eesti keelel on maailma *ca* 200 riigikeele hulgas ka riigikeele staatus. Me peame ennast haritud rahvaks ja seda õigustatult. Kõrgharidust saab maailmas umbes 100 keeles, sh Euroopas ligi 30 keeles ja meie keel on üks nendest. Veelgi rõõmustavam pilt avaneb siis, kui vaatleme keeli, millel on keeletehnoloogiline tugi. Omakeelset Windowsi ja Microsofti laiatarbetarkvara saavad kasutada maailmas vaid 37 rahvast, eestlased seal hulgas. Toimiv kõnesüntees on olemas umbes 50 keele jaoks, Eestis on tekst-kõne süntees vabavarana kasutatav aastast 2002[2]. Kõnetuvastus on olemas või valmimas vaid 25 rahval, ka Eestis on olemas omakeelne kõnetuvastussüsteem.[3] Seega, kui eesti keel on kõnelejate arvult maailmas esimese kolmesaja hulgas, hariduse valdkonnas esimese saja hulgas, siis infotehnoloogia arengus oleme maailma esimese 30 arenenuma keele hulgas.

Tundub, nagu oleks kõik hästi ja eesti keele tulevik kindlustatud. Tegelikult asi nii lihtne ei ole ja selleks et ajaga kaasas käia, peab pidevalt tehnoloogiliselt võidurelvastuma ja aktiivselt tegutsema. Pessimistlikumate ennustuste kohaselt jääb XXI sajandi lõpuks alles mõnikümmend keelt ja nimelt just need keeled, millel on olemas korralik keeletehnoloogiline tugi. Hoiakutes eesti keele suhtes on täheldatavad ka teatud ohumärgid, mis näitavad, et me ei saa loorberitele puhkama jääda, et tuleviku nimel tuleb pingutada. Meil on küll olemas eestikeelne tarkvara, kuid tavaliste koolinoorte hulgas on selle kasutajaid vaid 10–15%. Eesti koolides ja ülikoolides on tarkvara valdavalt ingliskeelne. Innovaatilise Tiigrihüppe programmi sära on pisut tuhmumas ning e-riigi ja e-ühiskonna areng pole olnud alati küllalt

kiire. Kui Microsofti tarkvara on ligi neljakümnes keeles, siis programmides kasutatavad kõnetehnoloogilised vahendid, nagu tekst-kõne-süntesaatorid ja kõnetuvastajad, on kättesaadavad vaid üksikutes suurtes keeltes: inglise, hispaania, hiina, jaapani, saksa ja prantsuse keeles. Sellised suulise keele suhtlusvahendid nagu automaatsed diktofonid, kõnesüntees ja seadmete hääljuhtimise võimalus loovad eeldused selleks, et mõni suurem keel – tõenäoliselt inglise keel – võib kergesti sattuda suhtluskeele tasandile. Väikesed riigid peavad oma piiratud ressursside ja võimaluste kiuste väga targalt tegutsema: ühelt poolt mõjutama protsesse seadusandlikult, teisalt vaatamata kitsastele oludele suunama piisava hulga vahendeid asjakohaste väljatöötluste edasiarendamiseks.

Eesti keele keeletehnoloogiline tugi

Eesti keel oma pisut alla miljonilise kõnelejaskonnaga pole paraku keeletehnoloogiliste rakenduste jaoks kuigi rentaabel. Riigi toetus on vältimatu. 2006. aastal käivitus riiklik programm „Eesti keele keeletehnoloogiline tugi 2006–2010“, mille eesmärk on arendada keeletehnoloogiline tugi tasemele, mis võimaldaks eesti keelel tänapäevases infoühiskonnas edukalt toimida ja areneda.

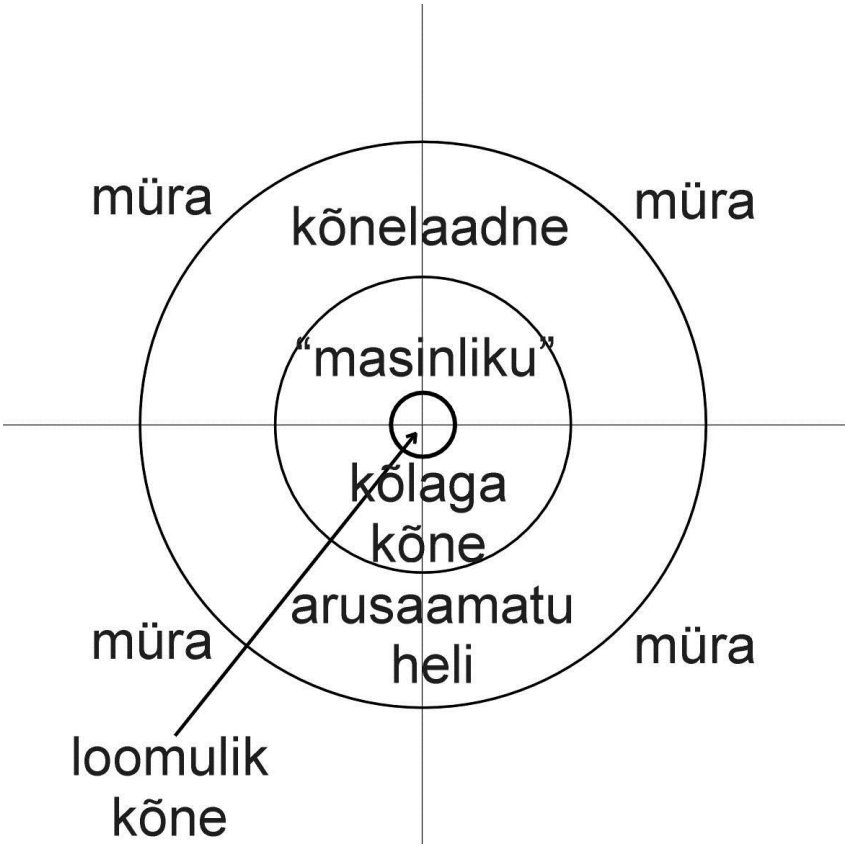
Programmi prioriteetideks kirjaliku teksti töötluse alal on masintõlge ja infodialoog. Masintõlge on teksti tõlkimine ühest keelest teise ilma inimese osaluseta. Probleem on eriti aktuaalne Euroopa Liidu tõlkevajadusi silmas pidades. Teatavasti on Euroopa Liidus 23 ametlikku keelt. Selleks et tagada eri keeli kõnelevate inimeste suhtlemine ja teha võimalikuks dokumentide tõlkimine igasse keelde, on Euroopa Liidu peakorteris tööle rakendatud tohutu tõlkijate armee. Kui tõlkimine õnnestuks automatiseerida, säästaks see palju raha ja inimressurssi.

Masintõlke täiusliku tasemeni jõudmiseks tuleb veel lahendada rohkesti keerukaid probleeme. Niisuguse taseme saavutamine ei ole veel selle programmi eesmärk. Sageli võib aga juba abi olla ka arvutiga tehtud toortõlkest. Suvalisest võõrkeelsest tekstist eestikeelse toortõlke saamiseks on juba praegu võimalik kasutada Google'i automaattõlkijat.[4] Viimane sisaldab kokku 51 keelt (sh eesti keelt) ning võimaldab aru saada, millest tekstis räägitakse. Sellise masintõlkijaga saab ka lihtsaid eestikeelseid lauseid teistesse keeltesse ümber panna. Kvaliteet on muidugi kehv, mistõttu suurt täpsust nõudvate tõlgete tegemiseks, nt seaduste või lepingute tõlkimiseks, seda kasutada ei saa. Ometi võib sellisestki abivahendist mingi kasu olla.

Riikliku programmi raames keskendutakse põhiliselt eesti keelest inglise keelde ning inglise keelest eesti keelde tõlkimisele. Tõlkeprogrammid on vajalikud ka mitmekeelses infootsingus, tõlkimaks eestikeelsed päringud võõrkeeltesse ja otsingutulemused tagasi eesti keelde.

Inimene-masin-inimene-suhtluses muutub järjest olulisemaks suulise kõne roll. Kogu maailmas käib selles valdkonnas intensiivne uurimis- ja arendustöö. Riikliku programmi põhisuunad suulise keele alal on kõnesüntees, kõnetuvastus ja inimene-masin-dialoogsüsteemid. Kõnesüntees teisendab ortograafilise teksti tehiskõneks, mida kasutatakse dialoogsüsteemides, nägemis- ja kõnepuudega inimestele mõeldud abivahendites jm. Kasutajale on oluline eelkõige väljundkõne kvaliteet, s.o kõnest arusaadavus ja loomulikkus. Loomuliku väljundkõne tekitamine ei ole lihtne ülesanne. Erinevaid inimhääli konkreetses keeles võib olla miljoneid või isegi sadu miljoneid, ometigi suudame pea ilmeksimatult eristada inimhäält sünteeskõnest. Meie kõrv on üldjuhul nii tundlik, et mingis nüansis ei suuda kõnesüntesaator „inimlikuks“ jääda ja kaldub loomulikkuse alast välja. Kui püüda näitlikult kujutada inimhäälele lähedasi helisid mingis kahemõõtmelises tunnusruumis, siis

loomulik kõne võtab sellest vaid väga väikese osa, n-ö märklaua kümne või keskosa (joonis 1).



Joonis 1. Loomulik kõne inimhäälele lähedaste helide kahemõõtmelises tunnusruumis.

Kõnetuvastus on vastupidiselt kõnesünteesile inimkõne teisendus kõne sisule vastavaks tekstiks. Kõnetuvastussüsteem on oluliseks mootoriks teaduse-tehnika revolutsioonis. Hiljuti veel muinasjutulise „Seesam, avane!“ fenomeni vilju saaksime hakata kasutama teksti sisestamisel arvutisse, seadmete, süsteemide ja robotite hääljuhtimisel jm. Eestikeelse kõnetuvastuse aktuaalseim ülesanne on „automaatne diktofon“, mille sisendiks on eestikeelne kõne ja väljundiks ortograafiline tekst. Probleemiks on see, et selliseid diktofone on võimalik küllalt hästi treenida konkreetse inimese häälele vastavaks, mingi teise inimese hääle tuvastamiseks tuleb süsteem ümber häälestada. Paljude kasutajatega süsteemid on valdavalt piiratud sõnavaraga (kindlad käsud, numbrid, tähed jms).

„Eesti keele keeletehnoloogiline toe“ eesmärk on keeletehnoloogiliste prototüüpide loomine, et IT-firmadel tekiks huvi eesti keelele sobivate vahendite väljatöötamise vastu ja programmi tulemused jõuaksid e-riigi ja e-ühiskonna kasutusse.

Seadusandlusest ja hoiakutest keelekasutuses

Peale teadus- ja arendustegevuseks ettenähtud rahalise toe peaks Eesti riik infotehnoloogia arengust tingitud protsesse mõjutama ka seadusandlikult. Tuleks hoolikalt jälgida, milliseks kujuneb eesti keele staatus infoühiskonnas või kas ja kuidas mõjutada lähiaastatel keelekasutust kogu meediaruumis. Võib küll kiidelda sellega, et infotehnoloogia vallas on

eesti keel maailmas 30 arenenuma keele hulgas ning et meil on eestikeelne Windowsi keskkond ja Microsofti laiatarbetarkvara, kuid paraku on emakeelne tarkvara kasutusel vaid vähestes üldharidus- ja ülikoolides. Arvutikeskkonna kasutusharjumustele luuakse aga põhi just nooruses. Suur hulk eestlasest arvutikasutajaid leiab, et inglise keel ongi arvuti keel ja eesti keelega pole siin midagi teha. Kui küsida näiteks prantslase, rootslase või soomlase käest, miskeelset tarkvara nad kasutavad, siis esialgu ei saa nad küsimuse püstitusest arugi. Pärast mõningast selgitust tuleb selge vastus, et loomulikult emakeelset tarkvara.

Eestlaste arvutikeele kasutuse hoiakute muutmiseks ei piisa ainult teadvustamisest ja üldsõnalistest programmidest, vaja oleks seadusandlikku tuge. Üks riigipoolne samm võiks olla see, et seadusega muudetaks eestikeelse tarkvara ja omakeelsete keeletehnoloogiavahendite kasutamine üldharidus koolides ning riigi- ja omavalitsusasutustes kohustuslikuks. See ergutaks ühelt poolt IT-firmasid tarkvara eestikeelseks lokaliseerima ning „Eesti keele keeletehnoloogiline toe“ raames loodavaid prototüüpe eri rakendustes kasutama. Et eesti keel pole keeletehnoloogiliste rakenduste jaoks kuigi rentaabel, seetõttu saab riik just seadusandlusega neid arenguid suunata. Teisalt annaks eestikeelne arvutikeskkond täiendava keelekümbluskeskkonna venekeelsetes koolides ja ka mõnes omavalitsuses.

Seaduste täiendamise ja ajakohastamise vajadust keelekasutuse valdkonnas tingib ka see, et meediakeskkond on XXI sajandil oluliselt muutunud. Kahjuks reguleeritakse meie seadustes eesti keele kasutust vaid teatud meediavaldkondades: raadios, televisioonis, audiovisuaalsete teoste esitamisel ja ajakirjandusväljaannetes. Paraku on nende osa marginaliseerumas. Seadustest on täiesti välja jäänud arvutikeskkond, ajaveebid, veebiportaalid, koduleheküljed, arvutimängud, internetifoorumid jms. Julgemate ennustuste kohaselt jääb viie aasta pärast traditsioonilisele raadiole, televisioonile ja trükiajakirjandusele vaid 20% meediaruumist. Milliseks see protsent täpselt kujuneb, näitab tulevik, aga traditsioonilise meedia taandumistendents on noorema põlvkonna eelistustes selgelt märgatav. Raadio, televisioon ja ajalehed on noorte jaoks suures osas Internetti kolinud: raadiot asendab Internetis leviv ja pleierisse allalaaditav muusika, filmid ja uue sarjad levivad Internetis enne, kui nad ametlikesse telekanalitesse jõuavad, trükiajakirjanduse infot saab suurepäraselt kätte neti vahendusel. Ja ehkki muusika ja filmide allalaadimine Internetist pole päris süütu tegevus, ei ole ometi suudetud sellele ka kätt ette panna, sest suletud failivahetusserverite asemele tekivad uued. Emori uuringu järgi kasutas eelmise aasta lõpus Eestis Internetti regulaarselt 806 000 inimest. Noored elavadki tegelikult Internetis, 15–24-aastaste hulgas on igapäevaseid kasutajaid 98%. Seevastu 50–74-aastaste hulgas on netist info hankijaid vaid 30%. Seega elavad eri põlvkonnad juba praegu suhteliselt erinevates meediamaailmades.

Kuidas meediakeskkonna laiendamist ja ajakohastamist seadusandlikult reguleerida? Vaevalt et paragrahvid suudaksid põlvkondade meediaharjumusi muuta sellega, et panna näiteks internetifoorumitele või jututubadele päitsed pähe ja nõuda neis korrektset eesti keelt. Kaudselt võiks aga keelekasutust reguleerivad seadused mõjutada blogosfääri, sest viimase aja trendina on hakatud ajakirjanduses järjest enam tsiteerima ajaveebis kirja pandud tekste. Korrektse eesti keele nõue ajakirjanduses distsiplineerib ka blogipidajaid, sest lohakalt vormistatud mõtteid ei toodaks ajakirjanduses esile. Küll aga peaks seadustega otseselt kindlustama eesti keele positsioone infoühiskonnas seal, kus see on mõistlik ja võimalik. Näiteks praegune keeleseadus tagab audiovisuaalsete teoste eestikeelse tõlke. Eestlased eelistavad dubleeritud filmidele pigem vaadata originaalkeeles subtiitritega varustatud filme ja saateid. On ka küllalt palju inimesi (vaegnägijad, nägemis- ja lugemishäirega inimesed ning väikelapsed), kes ei ole võimelised teksti teleriekraanilt lugema. Digitaaltelevisiooni ajastul võiks olla keeleseaduses ka nõue, et tõlge peab olema nii teksti kui ka häälena. Soomes on

niisugune teenus juba kättesaadav.

Majanduskriisi ajal loetakse raha, mistõttu on oluline, et seaduste rakendamise kulud oleksid minimaalsed. Eestikeelse arvutikeskkonna nõue ei tohiks riigile lisakulutusi tekitada, sest eestikeelne tarkvara on üldjuhul ingliskeelsega samas hinnas. Ja subtiitrite helindamiseks ei ole vaja palgata diktoreid, sest subtiitrite ettelugemisega tuleb edukalt toime ka kõnesüntesaator.

Kokkuvõte

Keele elujõud XXI sajandi tehnoloogia kiire arengu ajastul sõltub paljudest asjaoludest. Keeletehnoloogilise toe loomine ning omakeelne arvuti- ja meediakeskkond on olulised meetmed eesti keele püsimajäämiseks. Riik saab seadustega sihipäraselt reguleerida tehnoloogia arengust tingituid protsesse ning kujundada hoiakuid tagamaks eesti keelele infoühiskonnas kindel seisund ja igakülgne areng. Et saada suureks vaimult – selle Jakob Hurda poolteist sajandit tagasi seatud eesmärgi saavutamisel on eesti keelel tänapäevalgi väärtusliku tööriista tähtis roll. Vaba vaim saab kasvada ja areneda XXI sajandi infoühiskonnas vaid siis, kui emakeelt on võimalik igakülgselt kasutada kõikides valdkondades.

---

[1] **T. Dutoit**. An introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers. Dordrecht, 1997.
[2] **M. Mihkla**. Kõnesüntees?... See on imelihtne. - Oma Keel 2008, nr 1, lk 5-15.
[3] **T. Alumäe**. Methods for Estonian large vocabulary speech recognition. PhD thesis, Tallinn University of Technology, 2006.
[4] Vt Google Translate – http://translate.google.com.