**Eiríkur Rögnvaldsson**

# Máltækni

**Erindi flutt í Ríkisútvarpinu 2. nóvember 2010**

Ég geri ráð fyrir því að flestir hlustendur hafi einhvern tíma hringt í þjónustuver, t.d. hjá símafyrirtæki, þar sem þeim er boðið að velja milli nokkurra kosta með því að ýta á mismunandi takka á símanum – 1 fyrir sölu áskrifta, 2 fyrir reikninga, 3 fyrir tæknilega aðstoð, o.s.frv. Stundum eru möguleikarnir svo margir að notandinn tapar þræðinum og man ekki að upptalningu lokinni fyrir hvað 1 stóð eða hvernig átti að velja tæknilega aðstoð. Þegar notandanum hefur svo tekist að velja rétta kostinn er eins víst að hann sé númer 29 í röðinni og þurfi að bíða óratíma eftir að fá svar við fyrirspurn sem jafnvel er hægt að svara með einu orði eða einni setningu.

Auðvitað væri þægilegra ef notandinn gæti bara borið upp erindi sitt umsvifalaust og fengið svar við því þegar í stað, í stað þess að byrja á að velja milli margra möguleika og bíða síðan eftir að þjónustufulltrúi losnaði til að greiða úr málum hans. En þetta strandar á því að þjónustufulltrúarnir eru ekki nógu margir – álagið er misjafnt, og það væri allt of dýrt fyrir fyrirtækið að ráða svo marga starfsmenn að aldrei væri bið eftir þjónustu, jafnvel ekki á mestu álagstímum.

En þetta þarf ekki að vera svona. Í mörgum þjónustuverum þarf ekki að bíða eftir því að þjónustufulltrúi losni, heldur er það tölva sem hlustar á erindi notandans og greinir merkingu þess. Sú greining er síðan send til gagnabanka þar sem er að finna svör við margvíslegum fyrirspurnum, og viðeigandi svar sótt í bankann. Því svari er svo breytt í eðlilega setningu og hún send til tölvubúnaðar sem les hana fyrir notandann. Hugsanlega er svarið fullnægjandi þannig að notandinn geti þakkað fyrir sig og kvatt, en að öðrum kosti spyr hann áfram og sama ferlið endurtekur sig – spurningin er greind, svar sótt í gagnabanka o.s.frv.

Tölvan getur annað miklum fjölda fyrirspurna í einu þannig að biðtími nánast hverfur. Vissulega eru fyrirspurnir stundum of flóknar til að hægt sé að afgreiða þær á þann hátt, og þá á notandinn þess alltaf kost að biðja um samband við mannlegan þjónustufulltrúa – eða tölvan gefur samband við þjónustufulltrúa ef hún skilur ekki fyrirspurnina eða getur ekki svarað henni.

Þjónustuver af þessu tagi eru þó ekki til á Íslandi – ekki enn a.m.k. Forsenda þeirra er nefnilega flókin málfræðileg greining og þróaður hugbúnaður sem talsvert vantar á að til sé fyrir íslensku. Til að tölva geti greint fyrirspurn notandans þarf að vera til hugbúnaður sem greinir einstök hljóð og orð í tali notandans. Það dugir þó ekki til – það þarf einnig að greina setningagerð þess sem notandinn segir til að átta sig á merkingu þess. Tölvan þarf því að kunna reglur um íslenska setningagreiningu, og hún þarf að hafa aðgang að góðu orðasafni til að geta flett einstökum orðum upp. Síðan þarf að vera til viðamikill gagnabanki þar sem spurningar og svör eru tengd saman. Þá þarf tölvan að búa yfir reglum sem gera henni kleift að orða svörin á eðlilegri íslensku; og að lokum þarf að vera til búnaður sem getur borið svarið fram með eðlilegum íslenskum framburði.

Allt þetta – vélræn talgreining, vélræn setningagreining, vélræn merkingargreining, vélræn setningamyndun og tölvutal – fellur undir það sem heitir á ensku 'language technology' og kallað hefur verið **máltækni** eða tungutækni á íslensku. Það er mjög víðfeðmt og fjölbreytt svið, en í stuttu máli má segja að með **máltækni** sé átt við hvers kyns samvinnu tungumáls og tölvutækni sem hefur einhvern hagnýtan tilgang; beinist að því að hanna eða útbúa einhvern hugbúnað eða tæki sem nýtist mönnum í starfi eða leik. Þessi samvinna er tvenns konar, og felst annars vegar í notkun tölvutækninnar í þágu tungumálsins; hins vegar í notkun tungumálsins í þágu tölvutækninnar.

Það er hægt að nýta tölvu- og upplýsingatækni á ýmsan hátt til þess að auðvelda mönnum að nota tungumálið. Þar má nefna ýmiss konar hugbúnað til að leiðrétta og leiðbeina um stafsetningu og málfar. Slíkur búnaður fylgir t.d. algengum forritapökkum eins og *Microsoft Office* á ýmsum tungumálum. Íslensk stafsetningarleiðréttingarforrit eru til, einkum *Púki* Friðriks Skúlasonar, en ekkert málfræðileiðréttingarforrit er til fyrir íslensku. Vélrænar þýðingar af einu máli á annað falla undir þetta. Nokkrar tilraunir hafa verið gerðar til að láta tölvur þýða texta á og úr íslensku, og má þar vísa á vefsetur Stefáns Briem, `tungutorg.is`. Ljóst er þó að langt er í land með að til verði fullkomin þýðingaforrit milli íslensku og annarra mála.

Hér má einnig telja ýmiss konar hjálpartæki handa þeim sem eiga erfitt með mál eða lestur sökum einhvers konar fötlunar. Áður var minnst á búnað sem gerir tölvu kleift að lesa upp ritaðan texta. Slíkur búnaður nefnist talgervill og hefur verið til fyrir íslensku frá því um 1990. Íslenskir talgervlar hafa nýst blindum og sjónskertum mjög vel, en talsvert vantar þó á lýtalausan íslenskan framburð hjá þeim.

En tungumálið er ekki bara þiggjandi í þessari samvinnu við tölvutæknina. Það er líka notað á margvíslegan hátt til að gera tæknina aðgengilegri og auðvelda mönnum að nýta sér hana. Dæmi um það var nefnt hér á undan, þar sem tungumálið er notað til að auðvelda mönnum að fá svör við fyrirspurnum í þjónustuverum og afla sér upplýsinga úr gagnabönkum.

Í öðru lagi má nefna notkun málsins við stjórn tölva og ýmiss konar tölvustýrðra tækja. Það fer mjög í vöxt að slíkum tækjum sé stjórnað með venjulegu máli, annaðhvort rituðu eða töluðu. Skipanir eru þá ýmist slegnar inn á lyklaborð eða talaðar í hljóðnema, í stað þess að ýtt sé á þartilgerða takka. Þetta mun á næstunni taka til sífellt fjölbreyttari tækja, s.s. ýmiss konar framleiðslutækja, heimilistækja og bíla. En slík tæki skilja ekki íslensku – a.m.k. enn sem komið er.

Til að koma íslenskri máltækni á laggirnar þarf að byggja upp viðamikil málleg gagnasöfn – orðasöfn, textasöfn, hljóðsöfn o.fl. Slík söfn eru síðan notuð til að afla margvíslegra og nákvæmra upplýsinga um tungumálið. Til að hægt sé að þróa forrit til málfarsleiðréttingar þarf t.d. að liggja fyrir nákvæm og ítarleg greining á íslenskri setningagerð – mun nákvæmari og ítarlegri en finna má í handbókum og kennslubókum. Það er ekki hægt að útbúa leiðréttingarforrit nema skrá nákvæmlega hvaða setningagerðir eru leyfilegar í málinu og hverjar ekki, og jafnframt semja lýsingu á því hvernig eigi að lagfæra það sem betur má fara.

Til að hægt sé að útbúa fullkominn íslenskan talgreini þarf að safna upptökum af framburði mikils fjölda Íslendinga, greina þessar upptökur og skrá þau tilbrigði sem geta komið fyrir og talgreinirinn þarf að ráða við. Hann þarf að geta greint raddir bæði karla og kvenna, ungra og aldinna, hratt og hægt tal, skýrt og óskýrt, norðlensku og sunnlensku, o.s.frv. En það er ekki nóg að byggja upp gagnasöfn – það þarf líka að skrifa hugbúnað sem vinnur með þessi gagnasöfn. Þar má nefna forrit til málfarsleiðréttinga, þýðingaforrit, talgervla o.s.frv.

Þetta uppbyggingarstarf er dýrt. Það kostar jafnmikið koma upp máltækni fyrir íslensku og fyrir tungumál milljónaþjóða. Margs konar máltæknibúnaður er hins vegar góð markaðsvara og skilar miklum tekjum sem standa undir háum þróunarkostnaði – ef markaðurinn er nógu stór. En því er ekki að heilsa á Íslandi. Vegna smæðar markaðarins er ljóst að það verður seint arðvænlegt að þróa dýran máltæknibúnað fyrir íslensku. Ef við viljum að íslenska sé nothæf innan tölvu- og upplýsingatækninnar þarf opinber stuðningur við þróunarstarf að koma til.

Um síðustu aldamót setti menntamálaráðuneytið af stað sérstakt átak til að byggja upp íslenska máltækni. Til þeirrar uppbyggingar var varið talsverðu fé á árunum 2000-2004 og fyrir það fé tókst að byggja upp ýmis gagnasöfn og hugbúnað, svo sem íslenska beygingarlýsingu sem er aðgengileg á vefsetri Árnastofnunar, `arnastofnun.is`; talgervilinn Röggu, sem m.a. les upp fréttir

mbl.is og hægt er að láta lesa fyrir sig á vefthulan.is; íslenskan talgreini, og svokallaða markaða íslenska málheild, sem er málfræðilega greint safn 25 milljón orða af fjölbreyttum textategundum.

Eftir að átaki ráðuneytisins lauk hefur verið haldið áfram að byggja upp hugbúnað á þessu sviði, einkum fyrir styrki frá Rannsóknasjóði. Þar má nefna íslenskan markara, sem er forrit sem greinir orðflokk og beygingarlega þætti orða; þáttara, sem greinir texta setningafræðilega; og lemmald, sem greinir uppflettimynd orða. Þessi forrit eru öll aðgengileg á nlp.ru.is og þar er hægt að prófa að láta þau greina texta. Einnig er unnið að hugbúnaði sem þýðir úr íslensku á ensku, vélrænni merkingargreiningu, og smíði svokallaðs trjábanka sem er textasafn með setningafræðilegri greiningu.

Talgreinirinn og talgervillinn voru gerðir í samvinnu Háskóla Íslands og fyrirtækja, einkum Hex og Símans, og í samstarfi við alþjóðlega fyrirtækið Nuance sem sérhæfir sig á þessu sviði. En að framantöldum verkefnum hafa að öðru leyti einkum unnið fræðimenn frá Háskóla Íslands, Háskólanum í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum. Þeir hafa komið sér upp samstarfsvettvangi sem nefnist Máltæknisetur. Því er m.a. ætlað

- að vera upplýsingaveita um íslenska máltækni og reka vefsetur í því skyni
- að stuðla að samstarfi háskóla, stofnana og fyrirtækja um máltækniverkefni
- að skipuleggja og samhæfa háskólakennslu á sviði máltækni
- að taka þátt í norrænu, evrópsku og alþjóðlegu samstarfi á sviði máltækni
- að eiga frumkvæði að og taka þátt í rannsóknarverkefnum og hagnýtum verkefnum á sviði máltækni
- að halda utan um ýmiss konar hráefni og afurðir á sviði máltækni
- að halda ráðstefnur með þátttöku fræðimanna, fyrirtækja og almennings
- og að beita sér fyrir eflingu íslenskrar máltækni á öllum sviðum

Vefsíða setursins er maltaeknisetur.is, og þar má finna margvíslegar upplýsingar um íslenska máltækni.

En hverjar eru framtíðarhorfur íslenskrar máltækni – munum við í framtíðinni geta notað íslensku innan upplýsingatækninnar? Ég er sannfærður um að það veltur algerlega á íslenskum almenningi. Ef almennir málnotendur vilja hafa upplýsingatækni á íslensku og sætta sig ekki við annað þá verður hún á íslensku. Til þess þarf að halda áfram uppbyggingu íslenskrar máltækni, og á því eru engar tæknilegar hindranir. Hindranirnar eru fyrst og fremst fjárhagslegar, en það er samt smámál miðað við jarðgöng, virkjanir, mislæg gatnamót og annað sem Íslendingar hafa lagt fé í á undanförnum árum. Og þá má spyrja hvers virði tungumálið sé okkur.

Þegar mikilvægi íslenskrar máltækni er metið verður að líta til þess að upplýsingatæknin er orðin mikilvægur þáttur í daglegu lífi alls almennings í landinu. Ef ekki verður hægt að nota íslensku innan hennar kemur upp splunkuný staða, sem ekki á sér hliðstæðu fyrr í málsögunni. Þá verður orðið til mikilvægt svið í daglegu lífi venjulegs fólks, þar sem móðurmálið er gagnslítið eða ónothæft. Hvaða áhrif hefði slíkt umdæmistap á málnotendur og málsamfélagið? Hvað gerist ef móðurmálið er ekki lengur nothæft í nýrri tækni og öðru sem er nýtt og spennandi; á sviðum þar sem nýsköpun af ýmsu tagi á sér stað; og á sviðum þar sem ný atvinnutækifæri bjóðast? Menn þurfa varla að velta þessu lengi fyrir sér til að sjá hættumerkin.

En það er ástæðulaust og rangt að meta þörf á íslenskri máltækni eingöngu út frá sjónarmiði tungumálsins og varðveislu þess. Við eigum einnig og ekki síður að líta á þetta út frá þörfum og hagsmunum okkar, almennra málnotenda. Við eigum kröfu á því að geta notað móðurmálið hvar sem er í íslensku málsamfélagi – líka innan upplýsingatækninnar. Við eigum að krefjast þess að hugbúnaðurinn sem við notum sé á íslensku, að við fáum leiðréttingarhugbúnað fyrir íslenskan

texta, að við getum talað við ýmis tölvustýrð tæki á íslensku, að við fáum þýðingarforrit sem geti þýtt milli íslensku og annarra mála, að við getum unnið flóknar upplýsingar úr íslenskum texta- og gagnasöfnum og leitað í þeim á margvíslegan hátt, o.s.frv. Við eigum þetta skilið – og íslenskan á það skilið.

Það er hins vegar ekki hægt að búast við því að stjórnmálamenn eða fyrirtæki vilji leggja fé í búnað sem enginn hefur áhuga á eða vill nota. Ég held því að við þurfum á vitundarvakningu meðal almennings að halda. Málnotendur þurfa að átta sig á að það er engin ástæða til lítilþægni – upplýsingatæknin getur verið á íslensku og á að vera það. En það gerist ekkert nema almennir málnotendur vilji geta notað íslensku innan upplýsingatækninnar og sýni þann vilja í verki. Og þetta er ekki bara okkar mál – það er beinlínis skylda okkar við komandi kynslóðir að gera tungumálið gjaldgengt innan upplýsingatækninnar. Ef við missum það svið algerlega til enskunnar náum við því aldrei til baka.
Í nýrri íslenskri málstefnu, sem Alþingi samþykkti 12. mars 2009, er sett fram það meginmarkmið um notkun íslensku innan upplýsingatækninnar "Að íslensk tunga verði nothæf – og notuð – á öllum þeim sviðum innan tölvu- og upplýsingatækninnar sem varða daglegt líf alls almennings". Þetta merkir m.a. að til þarf að vera ýmiss konar hugbúnaður sem liðsinnir og leiðbeinir notendum við notkun íslensks máls (leiðréttingarforrit, þýðingarforrit, hjálparforrit fyrir fatlaða); og að unnt á að vera að nota íslensku sem samskiptamál við ýmiss konar tölvu- og tæknibúnað (upplýsingakerfi, þjónustuver, tölvustýrð tæki af ýmsu tagi). Í framhaldi af þessu er svo sett fram aðgerðaáætlun í níu liðum, m.a. eftirfarandi:

- Að stöðugt verði unnið að uppbyggingu og eflingu mállegra gagnasafna sem eru forsenda fyrir þróun og smíði margs kyns máltæknibúnaðar.
- Að hugbúnaður til að lagfæra og leiðrétta íslenskt málfar verði gerður og kominn í notkun innan þriggja ára.
- Að nothæf þýðingarforrit milli íslensku og valinna erlendra mála, a.m.k. ensku, verði gerð innan fimm ára.
- Að íslenskur talgervill og talgreinir sem gerðir voru á vegum tungutækniátaks menntamálaráðuneytisins verði endurbættir og lagaðir að nýjustu tækni.

Þessi stefnumörkun er mikilsverð þótt óljóst sé um efndir og ekki sé líklegt að mikið fé verði lagt í aðgerðir á þessu sviði á næstunni. En forsenda þess að stefna af þessu tagi skili árangri er sú að hún njóti almenns stuðnings. Fátt væri verra fyrir íslenska tungu en opinber málstefna sem væri aðeins fögur orð en hefði ekki tengsl við almenna málnotendur og stuðning þeirra.