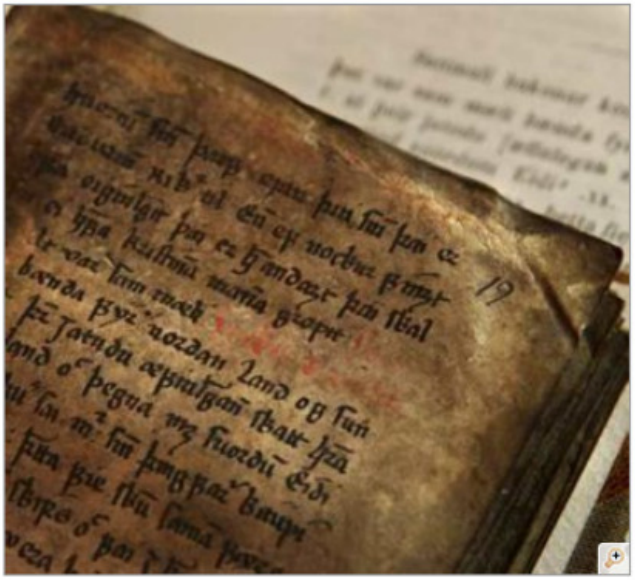
Eiríkur Rögnvaldsson
Háskóla Íslands

Mennta- og menningarmálaráðherra – góðir gestir!

Íslenskan er verkfærið sem við notum til allra samskipta – við fjölskyldu okkar, samstarfsfólk og aðra samtímamenn. En hún er ekki bara samskiptatæki heldur einnig dýrmætasti menningararfur okkar því að hún kemur okkur í beint samband við Íslendinga fyrri alda. Máltækni snýst um sambúð tölvutækninnar við þennan dýrmæta menningararf. Við notum tölvurnar til að styðja við tungumálið á ýmsan hátt – leiðrétta stafsetningu og málfar, þýða texta af einu tungumáli á annað, breyta rituðum texta í talað mál, og svo mætti lengi telja. En tungumálið gagnast ekki síður tölvutækninni – talað mál eða ritað er nú notað til að stjórna ýmsum tölvustýrðum tækjum og gera þau þannig aðgengilegri og notendavænni. Ef vel tekst til lifir hvorugur aðilinn sníkjulífi á hinum, heldur verður til samlífi sem kemur báðum vel, rétt eins og sveppar og þörungar eiga farsælt samlífi í fjallagrösum sem við sjáum hér á myndinni.

Innlent | Morgunblaðið | 17.5.2011 | 5:48

Tæknin les fornritin líkt og nútímamál



„Ég hugsa að það séu ekki mörg tungumál sem væri hægt að fara svona með, að nota forrit sem er lagað að einu málstigi og beita því svona auðveldlega á 700 til 800 árum eldri texta," segir Eiríkur Rögnvaldsson prófessor í íslenskri málfræði við Háskóla Íslands.

Forrit sem þróað var innan Árnastofnunar til málfræðigreiningar á nútímatexta reyndist einnig geta greint texta Íslendingasagnanna og Sturlungu með viðunandi hætti. Sú fullyrðing að Íslendingar geti lesið fornrit sín með auðveldari hætti en aðrar þjóðir virðist því hreint ekki svo fjarri lagi.

Gott dæmi um þetta samlífi tungumáls og tækni er að finna í þessari frétt sem birtist í Morgunblaðinu fyrir réttu ári, 17. maí í fyrra. Blaðamaður hafði frétt það, ég held reyndar á ársfundi Árnastofnunar, að hugbúnaður sem var gerður til að greina nútímaíslensku málfræðilega hefði verið reyndur á miðaldatextum, Íslendingasögum og Sturlungu, og hefði gefið álíka góða raun á þeim og nútímamálinu. Þetta þótti fréttaefni – og er það. Sá menningararfur okkar sem felst í tungumálinu er einstakur, og stórmerkilegt að hægt skuli að nota greiningarbúnað fyrir nútímamál óbreyttan á fornmálið. Þessu samhengi viljum við ekki glata.

Innlent | Morgunblaðið | 25.4.2012 | 5:30

Íslensk tunga er að missa af talandi tölvum

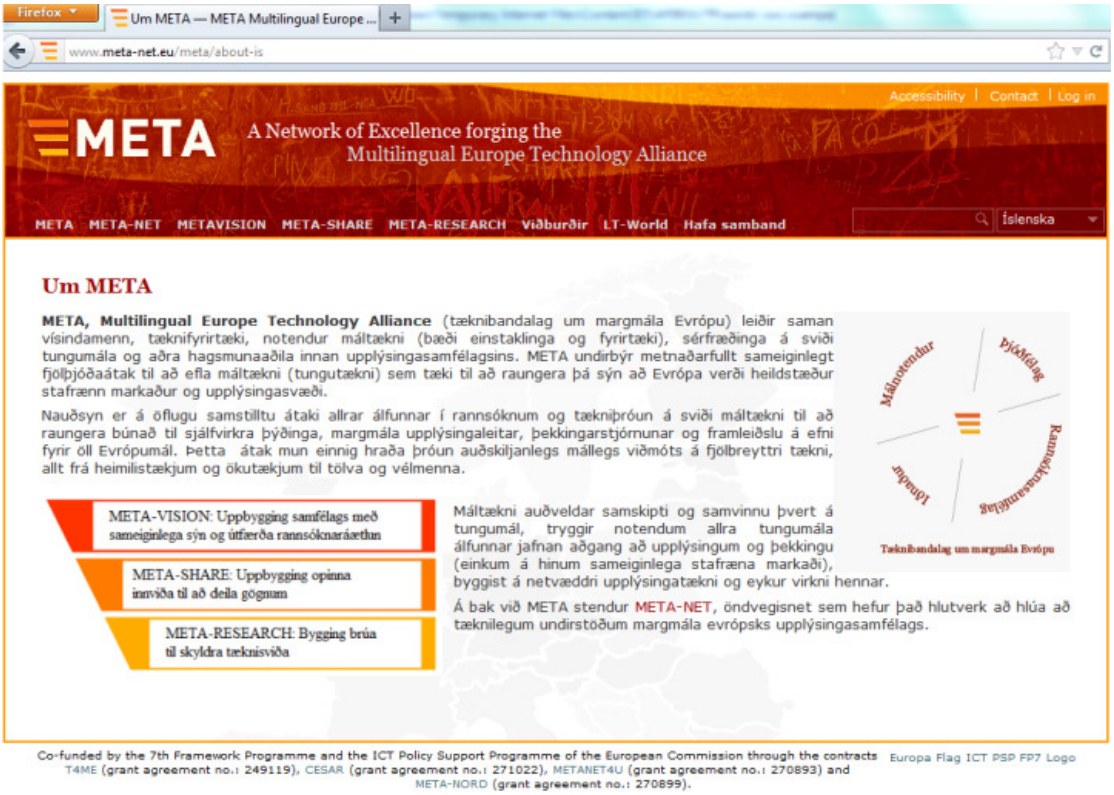
Tölvur eru komnar í síma sem geta t.d. grófþýtt texta. mbl.is/reuters

„Íslenska er illa búin undir framfarir í upplýsingatækni. Notkun tungunnar í upplýsingatækni er að aukast. Þeim fjölgar alltaf tölvunum sem hægt er að tala við og gefa ýmis fyrirmæli.

Slík samskipti við tölvur munu fara fram á ensku í framtíðinni ef okkur tekst ekki að bæta stöðu íslenskunnar að þessu leyti," segir Eiríkur Rögnvaldsson, prófessor í íslensku við Háskóla Íslands, um þetta áhyggjuefni í Morgunblaðinu í dag.

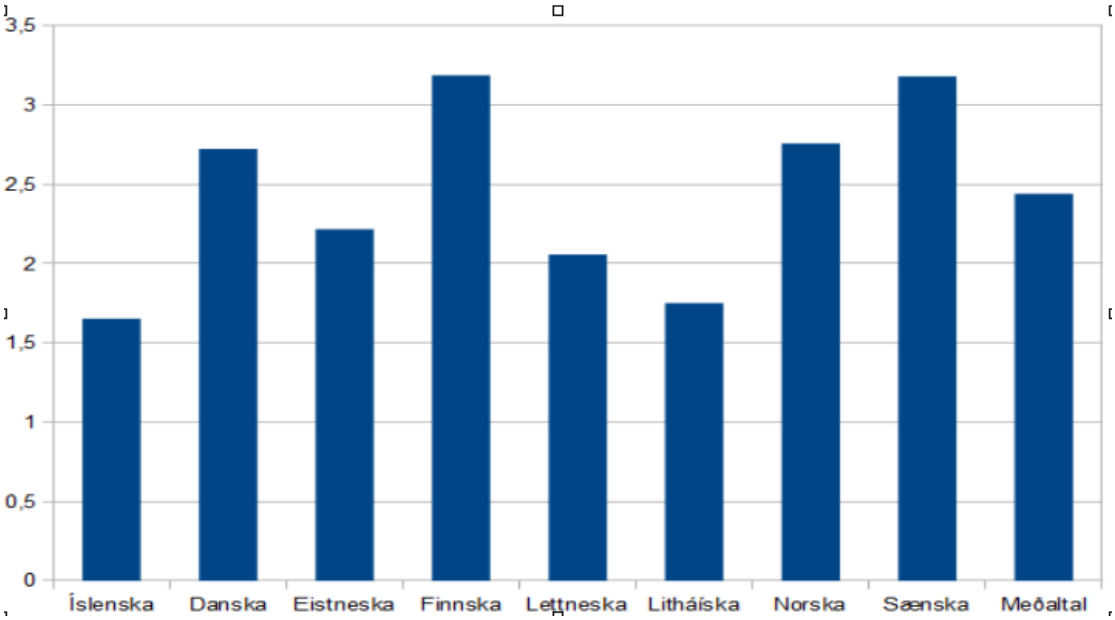
Tilefnið er slæm útkoma íslensku í skýrslu um stöðu tungunnar gagnvart nýrri upplýsingatækni í 30 Evrópulöndum. Niðurstöðurnar eru flokkaðar í nokkra flokka og fær íslenska iðulega lökustu einkunn. Meðal þess sem horft er til eru orða- og textasöfn og hugbúnaður til talvinnslu (talgervlar og talgreinar), vélþýðinga og málfræðilegrar greiningar, svo sem málfarsleiðréttingar.

Hér er svo önnur frétt úr Morgunblaðinu, þriggja vikna gömul. Þar er bent á þá ógn sem íslenskunni stendur af örri þróun í upplýsinga- og samskiptatækni. Allar horfur eru á því að í framtíðinni munum við aðallega nota tungumálið í samskiptum við tölvurnar, í stað þess að ýta á takka eða velja milli kosta í valmynd. Þetta á ekki eingöngu við tölvur í venjulegum skilningi, því að skilin milli tölvu og síma eru að hverfa og sífellt fleiri tæki eru nú tölvustýrð – framleiðslutæki, heimilistæki, bílar. En tölvurnar skilja ekki hvaða tungumál sem er, frekar en mannfólkið – móðurmál þeirra er oftast enska og ef þær eiga að geta haft samskipti við fólk á öðrum tungumálum þarf að kenna þeim það með ærinni fyrirhöfn og kostnaði.



Í fréttinni er vísað til skýrslu um íslenskt málsamfélag og íslenska máltækni sem er að koma út þessa dagana hjá Springer-forlaginu í Þýskalandi. Sambærilegar skýrslur voru gerðar á síðasta ári um 30 Evrópumál. Tilgangurinn var að fá yfirsýn yfir stöðu þessara tungumála gagnvart nýrri tækni og auðvelda stefnumótun til að styrkja vígstöðu málanna gagnvart enskunni. Á bak við þessar skýrslur stendur META-NET, sem er stórt Evrópuverkefni sem tekur til flestra ríkja álfunnar. Tilgangur verkefnisins er að kynna og efla evrópska máltækni, með það að markmiði að Evrópa verði margmála upplýsingasamfélag þar sem allir geti notað móðurmál sitt við öflun og úrvinnslu hvers kyns upplýsinga. Undir hatti META-NET eru þrjú svæðisbundin verkefni, í Suður-, Austur- og Norður-Evrópu. Norður-evrópska verkefnið nefnist META-NORD og tekur til allra Norðurlanda og Eystrasaltslanda.

Samanburður META-NORD mála



Hér má sjá samanburð sem byggist á stöðu META-NORD málanna á fjórum meginsviðum; talvinnslu, þ.e. talgreiningu og talgervingu; vélrænum þýðingum; textagreiningu, þ.e. málfræðilegri og setningafræðilegri greiningu; og mállegum gagnasöfnum, s.s. orða- og textasöfnum af ýmsu tagi. Í stöðumatinu er svo miðað við ýmsa þætti, s.s. magn eða fjölda gagnasafna og tóla, gæði þeirra, hversu opin og aðgengileg þau eru, hvort þau eru þróuð áfram, o.fl. Niðurstaðan er sú að íslenska er í neðsta sæti af META-NORD málunum – og raunar í næstneðsta sæti af öllum málunum 30, aðeins maltneska stendur verr að vígi.

Þótt staða íslenskunnar í alþjóðlegum samanburði mætti vissulega vera betri, og þurfi nauðsynlega að batna, er ekki eins og ekkert hafi verið gert í íslenskri máltækni. Sem betur fer hafa einstaklingar og stofnanir sem vinna á þessu sviði átt einkar náið og gott samstarf á undanförnum árum. Þetta samstarf hófst innan tungutækniverkefnis menntamálaráðuneytisins á árunum 2000-2004, en þegar því verkefni lauk ákváðu fræðimenn frá Háskóla Íslands, Háskólanum í Reykjavík og Orðabók Háskólans að taka höndum saman í Tungutæknisetri, sem nú nefnist Máltæknisetur. Þegar Stofnun Árna Magnússonar í íslenskum fræðum varð til gekk hún inn í samstarfið í stað Orðabókarinnar.

Íslensk málföng síðustu 10 ára

- **Beygingarlýsing íslensks nútímamáls (BÍN):** Beygingardæmi hátt á þriðja hundrað þúsund íslenskra orða
- **Mörkuð íslensk málheild (MÍM):** Málfræðilega greint safn 25 milljón orða úr fjölbreyttum textum
- **Icelandic Parsed Historical Corpus (IcePaHC):** Sögulegur íslenskur trjábanki – milljón setningagreindra orða frá 850 ára tímabili
- **Íslenskt orðanet:** Orðaforðinn í samhengi – 233.000 flettur og margvísleg merkingarvensl þeirra
- **ÍSLEX:** 50.000 íslensk flettiorð og þýðingar þeirra á dönsku, norsku, sænsku (og færeysku)
- **Íslenskur merkingarbrunnur:** Íslenskt orðasafn með merkingarvenslum
- **IceTagger:** Íslenskur markari – hugbúnaður sem greinir texta málfræðilega
- **IceParser:** Íslenskur hlutaþáttari – hugbúnaður sem greinir texta setningafræðilega
- **Apertium is-en:** Vélræn þýðing af íslensku á ensku, byggð á Apertium-kerfinu
- **Almannarómur:** Íslenskur talgreinir, í samvinnu við HR og Google
- **Íslenskur talgervill**, í samvinnu við Blindrafélagið og IVONA í Póllandi

Á undanförnum 10 árum hafa verið byggð upp og smíðuð fjölmörg íslensk málföng, en málföng er nýyrði yfir það sem á ensku nefnist 'language resources' og nær yfir bæði málleg gagnasöfn, svo sem orðasöfn og textasöfn af ýmsu tagi, upptökur af töluðu máli o.fl., og einnig hvers kyns hugbúnað til að vinna með tungumálið. En í raun og veru er íslensk máltækni miklu meira mál en tækni. Aðferðirnar, tæknin, eru að verulegu leyti alþjóðlegar, óháðar tungumálum, og óþarfi að finna upp sérstakar aðferðir fyrir meðhöndlun íslensku þótt vitanlega þurfi oft að laga tiltækar aðferðir að ýmsum sérkennum málsins. En til að hægt sé að beita tækninni á skilvirkan hátt þurfa að liggja fyrir gífurlega miklar og ítarlegar upplýsingar um öll svið tungumálsins – hljóðkerfi, beygingarkerfi, setningagerð, merkingu, málnotkun o.s.frv. Slíkar upplýsingar fást aðeins úr stórum gagnasöfnum sem hafa verið vandlega greind, og uppbygging slíkra safna er mjög tímafrek og kostnaðarsöm. Hér eru talin upp flest helstu málföng sem hafa orðið til undanfarinn áratug – fyrst ýmis gagnasöfn, en síðan hugbúnaður. Á bak við flest þessi málföng standa fræðimenn sem vinna undir merkjum Máltækniseturs eða starfa hjá þeim stofnunum sem standa að því.

Eitt meginmarkmið META-NORD er að safna upplýsingum um hvers kyns málföng og gera þau aðgengileg eftir því sem kostur er. Upplýsingarnar, lýsigögnin, verða settar í gagnabrunninn META-SHARE og þar eiga menn að geta fengið á einum stað upplýsingar um tiltæk málföng fyrir öll Evrópumál, hvernig sé hægt að nálgast þau, og með hvaða skilmálum megi nýta þau. Til hagræðis fyrir íslenska notendur höfum við sett upp sérstaka síðu, www.málföng.is, þar sem öll íslensk málföng verða skráð. Þarna eru nú þegar krækjur á ýmsan hugbúnað, textasöfn, orðasöfn og fleira, en margt á eftir að bætast við á næstunni. Við vonumst til að sem flest íslensk málföng verði aðgengileg þarna án greiðslu til að nýting þeirra verði sem mest, bæði í máltækni og til margvíslegra málfræðirannsókna.

Er menningararfurinn í hættu?

Að lokum má spyrja hvort hin bága staða íslenskunnar í þessum samanburði þýði að við eigum á hættu að glata þeim menningararfi sem felst í tungumálinu. Ályktunarorð skýrslunnar eru þau að íslenska sé ekki í bráðri hættu, þrátt fyrir yfirburði enskunnar í máltækni og tölvumálvísindum. Á hinn bóginn gæti staðan gerbreyst á svipstundu þegar ný kynslóð tækninnar fer fyrir alvöru að ráða við mannlegt mál á skilvirkan hátt. Tungumál sem jafnvel mjög fáir tala geta lifað af, verði fullnægjandi máltæknibúnaður tiltækur. Án slíks búnaðar munu jafnvel stórþjóðatungumál verða í mikilli hættu. Eigi íslenska að vera lífvænleg þjóðtunga í þróuðum heimi verður hún að geta staðið undir kröfum upplýsingatækninnar. Efling máltækni verður því að vera grunnþáttur í framkvæmd íslenskrar málstefnu.