Íslenska, upplýsingatækni og máltækni – fortíð og framtíð

Eiríkur Rögnvaldsson
Háskóla Íslands
eirikur@hi.is

Útdráttur

Hér verður gerð grein fyrir uppbyggingu íslenskrar máltækni[1] undanfarinn áratug og helstu verkefnum sem unnið hefur verið að, svo og þeim sem eru á döfinni.[2] Einnig er fjallað um meginmarkmið og aðgerðir íslenskrar málstefnu á sviði máltækni og hugbúnaðarþýðinga, og rætt um forsendur þess að íslenska eigi sér framtíð innan upplýsingatækninnar.

1 Inngangur

Í stefnuskrá Íslenskrar málnefndar 2006-2010 segir: „Upplýsingatækni verður æ mikilvægari þáttur í daglegu lífi Íslendinga og er afar brýnt að þeir geti notað móðurmálið á þeim vettvangi. Til að mynda er áríðandi að ýmiss konar grundvallarhugbúnaður (til dæmis stýrikerfi, ritvinnsluforrit, póstforrit og vafrar) verði fáanlegur með íslensku notendaviðmóti og að í algengum ritvinnsluforritum verði hjálparbúnaður fyrir íslensku, svo sem forrit sem leiðrétta stafsetningu, málfar og skipta orðum rétt á milli lína. Þá hafa einnig orðið örar framfarir í vélrænum þýðingum á undanförnum árum og er því nauðsynlegt að íslenska verði hluti af þeirri þróun. Til þess að staða íslenskrar tungu í upplýsingatækni sé tryggð þarf að eiga sér stað kröftugt rannsóknar-, þróunar- og uppbyggingarstarf í íslenskri tungutækni“ (http://www.islenskan.is/Stefnuskra_2006-2010.htm).

Þarna er talað bæði um máltækni (tungutækni) og hugbúnaðarþýðingar. Þetta tvennt er oft tengt saman og mörkin þarna á milli eru óskýr og í hugum margra jafnvel ekki til, en í raun og veru er hér þó um tvö sjálfstæð viðfangsefni að ræða. Það er alveg hægt að hugsa sér íslenskan hugbúnað án nokkurrar máltækni – þ.e. án íslenskra leiðréttingarforrita, án vélrænna þýðinga milli íslensku og annarra mála, án íslenskra talgervla og talgreina, o.s.frv. Á sama hátt má vel hugsa sér íslenska máltækni þótt allt viðmót hugbúnaðarins sé á ensku, og það er reyndar staðan hjá mörgum notendum – algengt er að menn séu með villuleitarforrit eins og *Púkann* en séu svo með Windows á ensku.

En það er eðlilegt að almennir notendur dragi ekki skil milli máltækni og hugbúnaðarþýðinga. Hvorttveggja varðar tölvur og notkun tungumáls í tengslum við þær. Þetta snýst sem sé um hugtak sem hefur verið áberandi í málræktarumræðu undanfarin ár – *umdæmi* tungumálsins, sem í þessu tilviki er þá tölvuheimurinn og upplýsingatæknin. Verður íslenska ríkjandi í þessu umdæmi í framtíðinni – og hvað þarf til að svo verði? Í þessari grein ætla ég fyrst að rekja í stuttu máli þær breytingar sem hafa orðið á þessu sviði undanfarinn áratug, og velta síðan fyrir mér möguleikum íslenskrar tungu innan upplýsingatækninnar á næstu árum.

2 Sagan

Fyrir tíu árum var staða íslenskrar tungu innan upplýsingatækninnar ekki álitleg, hvort sem litið var til hugbúnaðarþýðinga eða máltækni. Staðan í hugbúnaðarþýðingum hafði þá raunar farið versnandi frá níunda áratugnum. Þegar frumstæð ritvinnslukerfi fóru að koma á markaðinn hér á landi upp úr 1980 voru þau undantekningarlítið á íslensku. Stýrikerfi og notendahugbúnaður Macintosh-tölva hefur verið á íslensku fram undir þetta, og einnig stóð IBM á Íslandi fyrir umfangsmiklum hugbúnaðarþýðingum í samstarfi við Orðabók Háskólans. En fyrsta alvöru

[1] *Máltækni* er þýðing á því sem nefnist á ensku *language technology* og hefur verið kallað *tungutækni* á íslensku undanfarin ár. Það orð er á ýmsan hátt óheppilegt og margir þeirra sem vinna mest á þessu sviði hafa því tekið upp orðið *máltækni* í staðinn.

[2] Greinin er byggð á erindunum „Framtíð íslensku innan upplýsingatækninnar“ sem haldið var á málþinginu *Á íslenska sér framtíð innan upplýsingatækninnar?* á vegum Tungutækniseturs og Íslenskrar málnefndar í Háskólanum í Reykjavík 7. mars 2008, og „Íslensk máltækni – fortíð og framtíð“ sem haldið var á Hugvísindaþingi í Háskóla Íslands 14. mars 2009, svo og á grein í *Skímu*, tímariti Samtaka móðurmálskennara (Eiríkur Rögnvaldsson 2009).

ritvinnslukerfið sem var þýtt fyrir PC-tölvur var WordPerfect (*WordPerfect handbók* 1986). Sú þýðing var gerð árið 1986 og náði um skeið yfirburðastöðu á markaðnum – án efa að verulegu leyti vegna tungumálsins.

En með útbreiðslu Windows upp úr 1990 náði Word yfirhöndinni og var ekki íslenskað. Um sama leyti fóru svo að koma tölvupóstkerfi eins og Outlook og Eudora, vafrar eins og Netscape og Mosaic, og síðar Internet Explorer. Þessi forrit voru öll á ensku, og með sprengingu í netnotkun á síðasta áratug vandist almenningur því að tölvur og allt sem þeim tengdist væri á ensku. Staðan í máltækni var síst betri – íslensk máltækni var varla til fyrir áratug. Við áttum að vísu ágætan stafrýni, *Púka* Friðriks Skúlasonar, og talgervil sem tæpast var boðlegur til almennrar notkunar þótt hann nýttist blindum og sjónskertum vel. En þar með var það upp talið.

Fyrir tíu árum komst svo hreyfing á málin, bæði á sviði hugbúnaðarþýðinga og máltækni, að frumkvæði Björns Bjarnasonar, þáverandi menntamálaráðherra. Sumarið 1998 hóf hann viðræður við Microsoft sem leiddu til þess að gerður var samningur menntamálaráðuneytisins og Microsoft um íslenskun á Windows 98 stýrikerfinu (sjá http://www.bjorn.is/greinar/1999/01/20). Sama haust, 1998, fékk Björn Rögnvald Ólafsson dósent í eðlisfræði til að gera skýrslu um ástand og horfur í íslenskri máltækni. Rögnvaldur fékk tvo menn með sér í skýrslugerðina, Eirík Rögnvaldsson prófessor í íslenskri málfræði og Þorgeir Sigurðsson, rafmagnsverkfræðing og íslenskufræðing. Þessi starfshópur skilaði af sér fyrir tíu árum, vorið 1999, og í skýrslu hans var eftirfarandi slegið föstu:

> Meginmarkmið Íslendinga hlýtur að verða að unnt verði að nota íslenska tungu, ritaða með réttum táknum, sem víðast innan tölvu- og fjarskiptatækninnar. Þar verður þó að sjálfsögðu að sníða sér stakk eftir vexti. Það er mikið verkefni að gera íslensku gjaldgenga á öllum sviðum, við allar aðstæður. Því verður að leggja megináherslu á þá þætti sem varða daglegt líf og starf alls almennings, eða munu gera það á næstu árum (Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson 1999:34).

Hópurinn taldi að til að byggja upp íslenska máltækni þyrfti þrjár meginstoðir; menntað fólk, málsöfn, og málgreiningarforrit. Áhugi fyrirtækja þyrfti að vera fyrir hendi en einnig stuðningur hins opinbera. Íslensk máltækni sprytti ekki af sjálfu sér vegna smæðar málsamfélagsins og markaðarins. Því væri nauðsynlegt að hefja sem fyrst átak til að skjóta stoðum undir íslenska máltækni. Ríkið yrði að hafa forgöngu um þetta átak og bera meginkostnaðinn á fyrstu stigum þess. Æskilegast væri að markaðurinn tæki síðan við, en hann gæti ekki borið þróunarkostnaðinn í upphafi.

Í framhaldi af þessu lagði starfshópurinn til að stjórnvöld beittu sér fyrir átaki á fjórum sviðum til eflingar íslenskri máltækni næstu fimm árin (Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson 1999:28):

1. Byggð verði upp sameiginleg gagnasöfn, málsöfn, sem geti nýst fyrirtækjum sem hráefni í afurðir.
2. Fé verði veitt til að styrkja hagnýtar rannsóknir á sviði máltækni.
3. Fyrirtæki verði styrkt til þess að þróa afurðir máltækni.
4. Menntun á sviði máltækni og málvísinda verði efld.

Í tillögum starfshópsins var gert ráð fyrir að máltækniátakið stæði í a.m.k. fjögur ár og heildarkostnaður á ári yrði 225-250 milljónir króna (Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson 1999:30). Eftir það vonuðust menn til að íslensk máltækni yrði orðin sjálfbær og þarfnaðist ekki meiri opinberra framlaga.

3 Staðan

Í framhaldi af skýrslunni setti menntamálaráðuneytið af stað tungutækniáætlun til að styrkja stofnanir og fyrirtæki til að byggja upp grunngögn og búnað fyrir íslenska máltækni. Á fjárlögum áranna 2000-2004 voru alls veittar 133 milljónir króna til máltækni (sjá http://hamar.stjr.is/). Það er verulegt fé, en þó aðeins u.þ.b. 1/8 þess sem starfshópurinn taldi að þyrfti til að ná tilætluðum árangri. Það er því ekki von að öllum verkefnum sem talin voru brýn í skýrslunni hafi verið gerð skil.

Þó er óhætt að segja að furðu mikið hafi áunnist miðað við tilkostnað, og máltækniáætlunin hafi skilað heilmiklu. Þannig hafa mikilvæg gagnasöfn verið byggð upp, og ýmsum rannsóknar- og þróunarverkefnum hefur verið ýtt af stað. Ég skal nefna nokkur þau helstu – öll nema það síðasttalda voru að talsverðu eða öllu leyti kostuð af máltækniáætlun menntamálaráðuneytisins.

- Íslenskur talgreinir (stakorðagreinir; sjá Eiríkur Rögnvaldsson 2004; Helga Waage 2004)
- Íslenskur talgervill (Ragga – vefþulan; sjá Eiríkur Rögnvaldsson, Björn Kristinsson og Sæmundur Þorsteinsson 2006)
- Endurbætt ritvilluleit (Púki; sjá Friðrik Skúlason 2004)
- Beygingarlýsing íslensks nútímamáls (hátt í 300 þúsund orð og hátt á sjöttu milljón beygingarmynda; sjá Kristín Bjarnadóttir 2004, 2005)
- Mörkuð málheild (25 milljón orða safn fjölbreyttra texta, málfræðilega greint; sjá Sigrún Helgadóttir 2004)
- Málfræðilegur gagnamarki (forrit sem greinir texta málfræðilega; sjá Sigrún Helgadóttir 2005, 2007)
- Beygingar- og málfræðigreinikerfi (lauk ekki; sjá Maren Albertsdóttir og Stefán Stefánsson 2004)
- Leitarvél (Embla og leitarvél Já; sjá Hjálmar Gíslason 2006)

Margt af þessum gagnasöfnum og búnaði nýtist nemendum, skólakerfinu og almenningi beint. *Púkinn* er mjög naskur við að vinna stafsetningar- og innsláttarvillur. *Beygingarlýsingin* (http://bin.arnastofnun.is/) er t.d. ómetanlegt hjálpargagn í kennslu og verkefnavinnu. Talgervillinn *Ragga* nýtist blindum og sjónskertum nemendum mjög vel, og fjölmargir, m.a. lesblindir, nýta sér einnig þjónustu *Vefþulunnar* (http://www.vefthulan.is/lesa-texta/) til að lesa texta upphátt.

Þverfaglegt meistaranám í máltækni var í boði við Háskóla Íslands á árunum 2002-2004, en 2007 var námið endurskipulagt og er nú rekið í samvinnu Háskóla Íslands og Háskólans í Reykjavík. Nemendur í þessu námi hafa átt þess kost að taka námskeið í norræna máltækniháskólanum, Nordic Graduate School of Language Technology (NGSLT, http://www.ngslt.org) og hefur það verið ómetanlegt. Nú er framtíð námsins að vísu ótrygg þar eð starfstíma NGSLT er lokið og ekki lengur hægt að sækja námskeið þangað.

Auk aðildar að NGSLT á árunum 2004-2009 hafa Íslendingar tekið þátt í margvíslegu norrænu samstarfi á undanförnum árum. Þar má nefna norrænu rannsóknaráætlunina í máltækni (Nordic Language Technology Research Programme) og ýmis net á vegum hennar árin 2001-2004, og Norræna máltæknifélagið (Northern European Association for Language Technology, NEALT) stofnað 2006. Enn fremur hafa íslenskir fræðimenn í máltækni tekið þátt í ýmsum norrænum umsóknum um rannsóknar- og þróunarstyrki, sem fæstar hafa raunar hlotið brautargengi.

Árið 2005 var komið á fót formlegum samstarfsvettvangi þriggja stofnana þar sem helst hefur verið unnið að íslenskri máltækni. Þessi vettvangur nefnist Tungutæknisetur (á ensku Icelandic Center for Language Technology, ICLT). Aðstandendur setursins eru Málvísindastofnun Háskóla Íslands, tölvunarfræðideild Háskólans í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum. Því er ætlað að vera samstarfsvettvangur um rannsóknir, þróun og kennslu í máltækni og gegnir hlutverki sínu m.a. með því að:

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

Frá því að máltækniáætlun menntamálaráðuneytisins lauk í árslok 2004 hafa aðstandendur Tungutækniseturs staðið fyrir eða tekið þátt í þróun ýmissa afurða sem flestar hafa verið styrktar af Rannsóknasjóði. Þar má helst nefna:

- Málfræðilegur reglumarkari, *IceTagger* (sjá Hrafn Loftsson 2006, 2007)
- Hlutaþáttari, *IceParser* (forrit sem greinir helstu setningarliði og setningarleg hlutverk; sjá Hrafn Loftsson og Eiríkur Rögnvaldsson 2007)
- Lemmunarforrit, *Lemmald* (forrit sem greinir uppflettimynd orða; sjá Anton Karl Ingason o.fl. 2008)
- Samhengisháð ritvilluleit (búnaður sem skoðar samhengi orðanna í stað þess að líta bara á einstakar orðmyndir; sjá Anton Karl Ingason o.fl. 2009)

IceTagger, *IceParser* og *Lemmald* er aðgengilegt á netinu (http://nlp.ru.is:8080/IceNLPWeb/icenlp_isl.html) og geta nemendur og aðrir nýtt sér það við greiningu orða og setninga. Frumkóti allra forritanna er nú opinn á http://sourceforge.net/projects/icenlp/. Tilraunaútgáfa af samhengisháðri ritvilluleit hefur verið felld inn í LanguageTool (http://www.languagetool.org/) sem vinnur með OpenOffice (http://www.openoffice.org/).

Einnig má nefna að íslenskir fræðimenn hafa á þessum árum flutt erindi og birt veggspjöld á fjölda norrænna og alþjóðlegra máltækniráðstefna, og birt greinar um máltækni í ráðstefnuritum og ritrýndum alþjóðlegum tímaritum.

Það er því óhætt að segja að íslensk máltækni hafi orðið til á undanförnum áratug. Það má marka af því að

- menntun á sviði máltækni er hafin
- þátttaka í norrænni samvinnu hefur verið veruleg
- mikilvæg gagnasöfn hafa verið byggð upp
- ýmis grundvallarhugbúnaður hefur verið þróaður
- máltæknirannsóknir eru komnar af stað

Þrátt fyrir þetta fer því fjarri að íslensk máltækni sé orðin sjálfbær, eins og stefnt var að í tungutækniskýrslunni fyrir áratug – og það er engin furða. Það kostar jafnmikið að byggja upp málleg gagnasöfn og máltæknibúnað fyrir tungumál sem 300 þúsund manns tala og fyrir tungumál milljónaþjóða. Það er ekki von að fyrirtæki sjái sér hag í því að leggja í mikinn kostnað við að þróa og aðlaga máltæknibúnað fyrir íslensku – sú fjárfesting skilar tæpast hagnaði. Á mörgum sviðum máltækninnar er nánast allt óunnið. Það hefur t.d. sáralítið verið gert á sviði vélrænna þýðinga, svo að nefnt sé eitt verkefni sem tilgreint er í stefnuskrá Málnefndarinnar sem vitnað var til í upphafi.

Þótt illa hafi gengið að auka áhuga fyrirtækja á gerð máltæknibúnaðar hafa rannsóknir á sviði máltækni eflst verulega undanfarin ár. Rannsóknasjóður hefur veitt nokkra styrki til máltækniverkefna, m.a. til gerðar hlutaþáttara og samhengisháðrar ritvilluleitar sem nefnd eru hér að framan. Í ársbyrjun 2009 veitti Rannsóknasjóður svo öndvegisstyrk, u.þ.b. 14,5 milljónir á ári í þrjú ár, til nýs máltækniverkefnis sem fræðimenn og framhaldsnemar við Háskóla Íslands, Háskólann í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum standa að, í samstarfi við fræðimenn í University of Pennsylvania í Philadelphia í Bandaríkjunum og Universitat d'Alacant á Spáni. Verkefnið nefnist *Hagkvæm máltækni utan ensku – íslenska tilraunin* (*Viable Language Technology Beyond English – Icelandic as a Test Case*). Þetta er þverfaglegt rannsóknarverkefni sem hefur að meginmarkmiði að þróa vísindalegar máltækniaðferðir sem henta auðlindalitlum tungumálum, einkum beygingamálum. Að því verður unnið með því að

- endurbæta rannsóknaraðferðir og laga að íslensku
- nýta sérkenni íslenskunnar til að þróa nýjar hagkvæmar aðferðir sem gera kleift að byggja upp tól og gögn á einfaldari hátt en áður
- nýta þverfaglega þekkingu rannsóknarhópsins, reynslu hans úr fyrri verkefnum og samstarf við framúrskarandi erlenda vísindamenn til að tengja á frjóan hátt aðferðir ólíkra fræðigreina

Málvísindalegum og tölfræðilegum aðferðum verður stefnt saman og látnar vinna í sameiningu til að skapa nýja þekkingu og opna nýja möguleika.

Verkefnið skiptist í þrjá meginþætti. Sá fyrsti er vélræn merkingargreining texta; annar verkþáttur er vélræn grófþýðing milli ensku og íslensku; og þriðji verkþátturinn felst í vélrænni setningagreiningu og gerð íslensks trjábanka, en trjábanki er texti sem hefur verið greindur setningafræðilega. Afurðir allra þessara verkþátta munu nýtast í margvíslegum búnaði, s.s. við þýðingar, leit á netinu og í gagnabönkum, vélræna málfarsleiðréttingu o.m.fl.

Aðstandendur verkefnisins líta á það sem lið í uppbyggingu íslensks BLARK (Basic LAnguage Resource Kit, sjá Krauwer 2003) en með því er átt við tilteknar lágmarksforsendur (gögn og hugbúnað) sem þurfa að vera til fyrir hvert tungumál eigi málið að vera nothæft í upplýsingatækni. Ýmsar þjóðir vinna að uppbyggingu BLARK fyrir tungumál sín, t.d. Eistar sem hafa gert um það metnaðarfulla áætlun (sjá Meister og Vilo 2008).

Árið 2006 beitti Norræna ráðherranefndin sér fyrir gerð skýrslu um forsendur þess að Norðurlöndin yrðu leiðandi á sviði máltækni árið 2016 (Koskenniemi o.fl. 2007). Í skýrslunni var lögð áhersla á

- stofnun NEALT (norræna máltæknifélagsins) og vinnuhópa á vegum þess
- samningu BLARK-skýrslna fyrir einstök ríki

- ráðstöfun norræns fjár í samvinnu um menntun og þjálfun
- að einstök ríki styrki hagnýt rannsóknarverkefni með þátttöku háskóla og fyrirtækja

Þegar BLARK-skýrslur lægju fyrir yrði norrænu fé veitt til gerðar máltæknibúnaðar, og norrænu og innlendu fé í hverju landi fyrir sig veitt til uppbyggingar málheilda, trjábanka og orðasafna. Ekkert hefur þó verið gert með skýrsluna; sótt hefur verið til Norrænu ráðherranefndarinnar og NordForsk um fé til norræns meistaranáms og til uppbyggingar rannsóknarinnviða en ekkert fengist.

4 Framtíðin

En hverjar eru þá framtíðarhorfur íslenskrar máltækni, og íslensku innan upplýsingatæknninnar? Ég er sannfærður um að þær velta 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, en 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 við leggjum fé í.

Það er 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. Við þurfum að sannfæra fólk um að það er engin ástæða til lítilþægni – upplýsingatæknin getur verið á íslensku og á að vera það. 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 upplýsingatæknninnar. Ef við missum það svið algerlega til enskunnar náum við því aldrei til baka.

Í nýrri íslenskri málstefnu, sem Alþingi samþykkti 12. mars sl., er sett fram eftirfarandi meginmarkmið um notkun íslensku innan upplýsingatækninnar (*Íslenska til alls* 2008:49):

- Að íslensk tunga verði nothæf – og notuð – á öllum þeim sviðum innan tölvu- og upplýsingatækninnar sem varða daglegt líf alls almennings.

  Þetta merkir í fyrsta lagi að viðmót algengs hugbúnaðar (valmyndir, hjálpartextar o.s.frv.) þarf að vera íslenskt; í öðru lagi að til þarf að vera ýmiss konar hugbúnaður sem liðsinnir og leiðbeinir notendum við notkun íslensks máls (leiðréttingarforrit, þýðingarforrit, hjálparforrit fyrir fatlaða); og í þriðja lagi að unnt á að vera að nota íslensku sem samskiptamál við ýmiss konar tölvu- og tæknibúnað (upplýsingakerfi, þjónustuver, tölvustýrð tæki af ýmsu tagi) (*Íslenska til alls* 2008:49).

Í framhaldi af þessu er svo sett fram aðgerðaáætlun í níu liðum, m.a. eftirfarandi (*Íslenska til alls* 2008:50):

- Að allur almennur notendahugbúnaður í íslensku skólakerfi, frá leikskólum til háskóla, verði á íslensku innan þriggja ára.
- 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.

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 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.

Heimildir

Anton Karl Ingason Ingason, Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). Aarne Ranta og Bengt Nordström (ritstj.): *Advances in Natural Language Processing*. Lecture Notes in Computer Science, Vol. 5221. Springer, Berlín, s. 205-216.

Anton Karl Ingason, Skúli Bernhard Jóhannsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir og Hrafn Loftsson. 2009. Context-Sensitive Spelling Correction and Rich Morphology. Kristiina Jokinen og Eckhard Bick (ritstj.): *Proceedings of NODALIDA 17*, s. 231-234. Tartu.

Eiríkur Rögnvaldsson. 2004. The Icelandic Speech Recognition Project *Hjal*. Henrik Holmboe (ritstj.): *Nordisk Sprogteknologi. Årbog 2003*, s. 239-242. Museum Tusculanums Forlag, Kaupmannahöfn.

Eiríkur Rögnvaldsson, Björn Kristinsson og Sæmundur Þorsteinsson. 2006. Nýr íslenskur þulur að koma á markað. *Morgunblaðið* 20. janúar.

Friðrik Skúlason. 2004. Endurbætt tillögugerðar- og orðskiptiforrit Púka. *Samspil tungu og tækni*, s. 29-31. Menntamálaráðuneytið, Reykjavík.

Helga Waage. 2004. Hjal – gerð íslensks stakorðagreinis. *Samspil tungu og tækni*, s. 49-53. Menntamálaráðuneytið, Reykjavík.

Hjálmar Gíslason. 2006. Resourcebehov i informationssøgning. Erindi flutt á málstefnunni *Språkteknologisk infrastruktur i Norden*, Gautaborg, 26. október 2006. http://www.nordisk-sprakrad.no/Bilaga%20presentationer.pdf

Hrafn Loftsson. 2006. Tagging Icelandic Text: A Linguistic Rule-Based Approach. *Technical Report CS-06-04*, Department of Computer Science, University of Sheffield.

Hrafn Loftsson. 2007. Tagging and Parsing Icelandic Text. Doktorsritgerð, Department of Computer Science, University of Sheffield.

Hrafn Loftsson og Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek og Mare Koit (ritstj.): *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, s. 128-135. Tartu.

Íslenska til alls. 2008. Tillögur Íslenskrar málnefndar að íslenskri málstefnu. Menntamálaráðuneytið, Reykjavík.

Koskenniemi, Kimmo, Krister Lindén og Torbjørn Nordgård. 2007. Expert Panel Report: The Nordic Countries – A Leading Region in Language Technology. Publications No. 40. Department of Linguistics, University of Helsinki.

Krauwer, Steven. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. *Proceedings of SPECOM 2003*, Moskvu, s. 8-15.

Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. *Samspil tungu og tækni*, s. 23-25. Menntamálaráðuneytið, Reykjavík.

Kristín Bjarnadóttir. 2005. Modern Icelandic Inflections. Henrik Holmboe (ritstj.): *Nordisk Sprogteknologi. Årbog 2005*, s. 49-50. Museum Tusculanums Forlag, Kaupmannahöfn.

Meister, Einar, og Jaak Vilo. 2008. Strengthening the Estonian Language Technology. Proceedings of *LREC 2008*, Marrakech, s. 3101-3004.

Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson. 1999. *Tungutækni. Skýrsla starfshóps*. Menntamálaráðuneytið, Reykjavík.

Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild. *Samspil tungu og tækni*, s. 67-71. Menntamálaráðuneytið, Reykjavík.

Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. Henrik Holmboe (ritstj.): *Nordisk Sprogteknologi. Årbog 2004*, s. 257-265. Museum Tusculanums Forlag, Kaupmannahöfn.

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga* 9:75-107.

WordPerfect handbók. 1986. Íslensk þýðing Eiríkur Rögnvaldsson lektor og Vilhjálmur Sigurjónsson kerfisfræðingur. Rafreiknir hf., Reykjavík.