TUNGUTÆKNI
SKÝRSLA STARFSHÓPS

Menntamálaráðuneytið
1999

Tungutækni
Skýrsla starfshóps

Menntamálaráðuneytið
Apríl 1999

Menntamálaráðuneytið : Skýrslur og álitsgerðir 9

Apríl 1999

Útgefandi: Menntamálaráðuneytið
Sölvhólsgötu 4
150 Reykjavík
Sími: 5609500
Bréfasími: 5623068
Netfang: postur@mrn.stjr.is
Veffang: www.mrn.stjr.is

Hönnun: XYZETA ehf.
Mynd á forsíðu Pálmi Guðmundsson
Prentun: Svansprent ehf.

ISBN 9979-882-22-0

Efnisyfirlit

Viðaukar

Formáli

Í september 1998 fól menntamálaráðherra, Björn Bjarnason, undirrituðum að kanna stöðu og möguleika tungutækni á Íslandi. Þar skyldi fjallað um: Hvað er tungutækni? Stöðu mála hér á landi, hið ritaða mál, tölvulestur, staðla og letur, þýðingar, leit í erlendum gagnabönkum, tölvutal og tölvuheyrn. Einnig skyldi fjallað um kostnað við að gera íslenskt mál meðfærilegt í tölvum, nauðsynlegar aðgerðir stjórnvalda og hvernig best verði staðið að þeirri vinnu. Við þessa vinnu skyldi tekið mið af þörfum samfélagsins í heild. Niðurstöðum skyldi skilað í skýrslu og var undirrituðum falið að velja með sér til þess verks þá sérfræðinga sem honum þætti hæfastir.

Eins og fram kemur í skýrslunni er tungutækni blanda af málvísindum og ýmiss konar tölvutækni. Hér á landi er lítil þekking á þessu sviði. Undirritaður var svo lánsamur að fá til liðs við sig tvo ágæta menn með reynslu af ákveðnum sviðum tungutækni. Þeir koma annars vegar úr málvísindum og hins vegar úr tölvuheiminum en hafa báðir þekkingu og reynslu af hinu sviðinu. Þetta eru Eiríkur Rögnvaldsson, prófessor í íslensku við Háskóla Íslands, og Þorgeir Sigurðsson, rafmagnsverkfræðingur og íslenskufræðingur, starfsmaður Staðlaráðs Íslands. Eiríkur hefur lengi haft áhuga á íslensku og tölvum og kennir námskeið á því sviði. Þorgeir hefur hannað talgervil og starfar nú við að búa til staðla, m.a. fyrir upplýsingatækni. Sigurður H. Pálsson, B.A. í almennum málvísindum og nemi í tölvudeild Verslunarháskólans, vann ýmis störf fyrir hópinn. Þótt allir í hópnum hafi á einn eða annan hátt reynslu af því sem skýrslan fjallar um getur enginn þeirra talist sérfræðingur í tungutækni. Því miður er þá ekki að finna hér á landi.

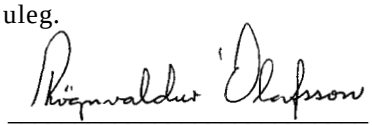
Hópurinn, sem hér er ýmist kallaður starfshópur eða nefnd, hefur eftir bestu getu reynt að kanna þau atriði sem menntamálaráðherra óskaði eftir. Sumt var erfiðara en búist var við, en annað auðveldara, einkum vegna þess hve lítið er til af tungutækni hér á landi. Undirrituðum er ríkt í huga hve mikið verk er hér óunnið og hve skammt á veg við erum komin í því að gera íslensku hæfa til þess að taka við upplýsingatækninni. Þetta er í ósamræmi við hinn almenna áhuga á tungunni hér á landi og það starf sem unnið hefur verið á ýmsum sviðum, til dæmis í nýyrðasmíð.

Hópurinn álítur að stöðu tungutækni hafi hrakað hér á landi síðan upplýsingatækniöld hófst hér fyrir rúmum tveimur áratugum. Nú hefur hins vegar verið blásið til sóknar og nýlega var gerður samningur við Microsoft um að þýða stýrikerfið Windows á íslensku. Hópurinn álítur að næsta skref hljóti að vera að búa til tól til þess að leiðrétta ritað íslenskt mál og að því verki ætti að hraða. Hann vill vara við því að grunnurinn er ótraustari en margir kunna að álíta að óathuguðu máli og því er hér er mikið verk óunnið. Vegna fámennis er markaður hér lítill og því skipta aðgerðir og skilningur stjórnvalda meiru hér á landi en í öðrum löndum.

Hópurinn vill þakka þeim fjölmörgu sem sýndu starfinu áhuga og aðstoðuðu hann á margvíslegan hátt. Greinilegt er að mikill áhugi er á því að geta notað íslensku sem víðast og að eignast tól til að vinna með íslenskt mál í tölvum.

Undirritaður vill þakka Eiríki, Þorgeiri og Sigurði fyrir ánægjulegt og fróðlegt samstarf. Í starfinu mættust oft mismunandi heimar og sjónarmið og skoðanaskipti urðu oft fjörleg, en alltaf ánægjuleg.

Reykjavík, 24. febrúar 1999,



Rögnvaldur Ólafsson

Ágrip og niðurstöður í stuttu máli

Ekki fer á milli mála að fáir tala íslensku. Hins vegar er íslenska virk þjóðtunga sem notuð er í öllum samskiptum og viðskiptum þjóðarinnar. Þýðing hennar er því mun meiri en tungumála sem fleiri nota, en eru aukatungur þjóða, eða tungur þjóðflokka sem eru minnihlutahópar í stærri þjóðfélögum. Þá er íslensk upplýsingatækni vel þróuð miðað við það sem gerist á öðrum málsvæðum. Hér eru fleiri tölvur á hvern íbúa en í flestum löndum og fleiri tengingar við Netið en víðast hvar. Því skiptir meiru fyrir íslensku að ráða við upplýsingatækni en ætla mætti af fjölda Íslendinga.

Íslensk stjórnvöld hafa áttað sig á þessu. Í bæklingi menntamálaráðuneytisins *Í krafti upplýsinga*, sem gefinn var út 1996, segir: „Styrkja þarf notkun íslenskrar tungu í upplýsingatækni og stuðla að nægu framboði efnis þannig að hægt sé að nálgast sem fjölbreyttast efni á íslensku. Framleiðendur hér á landi verða að geta nýtt sér nýja tækni og stuðlað að góðu framboði á íslensku efni á geisladiskum og Interneti á komandi árum.” Í samræmi við þessa stefnu var nýlega samið um þýðingu á Microsoft Windows stýrikerfinu á íslensku.

Hópurinn álítur að næsta skref á þessari braut eigi að vera að hvetja til þess að útbúin verði ýmiss konar tungutæknitól sem vinni með íslenskan texta og auðveldi notkun íslensku í upplýsingaþjóðfélaginu. Með þessu er m.a. átt við að gerð verði tól til að leiðrétta stafsetningu og málfræði, skipta orðum milli lína og svo framvegis; að samin verði rafræn íslensk orðabók og samheitaorðabók sem séu öllum aðgengilegar; að upplýsingar um beygingar orða verði í hverri tölvu; og fleira slíkt. Stígi þjóðin ekki þetta skref er hætt við að erfitt verði að nota íslenska tungu í upplýsingaþjóðfélaginu.

Sum vandamál tungutækninnar munu væntanlega leysast sjálfkrafa vegna öflugri tækni og breyttrar stefnu framleiðenda gagnvart erlendum mörkuðum, en önnur verða Íslendingar að leysa sjálfir. Hér skiptir höfuðmáli að reyna að tryggja að á öllum sviðum sé tekið tillit til íslenskrar tungu og sérkenna hennar strax við framleiðslu búnaðar. Einnig þarf að ganga hart fram í að koma íslensku inn í alþjóðlega staðla. Almennt þarf að nota altækar lausnir í stað sértækra. Þetta er eina stefnan sem getur tryggt að íslenska sé nothæf í upplýsingatækni í framtíðinni. Sérlausnir eru dýrar, þær hafa stuttan endingartíma og eru mjög erfiðar og mannfrekar í viðhaldi og þeim ætti ekki að beita nema í brýnustu neyð.

Sem stendur er markaður fyrir tungutækni á Íslandi ekki nægilega stór til þess að hann geti staðið undir þeirri þróunarvinnu sem þarf til þess að tryggja stöðu íslenskrar tungu í upplýsingasamfélaginu. Þetta er skýrt frekar í skýrslunni. Ekki er víst að þannig þurfi þetta að vera til frambúðar. Íslendingar hafa hingað til greitt fyrir sína íslensku ef svo má segja, útgáfa er mikil af bókum og blöðum og þjóðin greiðir fyrir það efni hærra verð vegna þess að efnið er á íslensku og markaðurinn er lítill. Á sama hátt mun þjóðin væntanlega smátt og smátt greiða þann kostnað sem hlýst af því að íslenska upplýsingatæknina.

Nefndinni virðist samt sem áður að átak þurfi að gera til þess að koma tungutækninni á fæturna og það verði ekki gert án stuðnings hins opinbera. Nefndin álítur að slíkt átak muni borga sig til lengri tíma litið. Markmiðið með átakinu ætti að vera að styrkja sameiginlegan grunn tungutækninnar og söfnun hráefnis fyrir tungutæknitólin og að hvetja fyrirtæki til að þróa tólin, meðal annars með því að nýta hráefnissafnið.
Á þennan hátt gæti skapast nýr iðnaður í tungutækni og sá sem þegar er fyrir hendi mundi styrkjast. Með þessu er átt við ýmsan iðnað tengdan útgáfu og meðferð tungumálsins, svo sem útgáfu á orðabókum og orðasöfnum, hugbúnað til leiðréttinga á stafsetningu og málfari, ýmis hjálparforrit við textasmíð, talgervla og hljóðtól. Vænta má, og ýta ætti undir, að slíkur iðnaður á Íslandi mundi nýta þekkingu sína og færni til þess að sækja inn á erlenda markaði, en þar munu vafalaust bjóðast ýmis tækifæri á næstu árum og áratugum.

Lagt er til að átakið verði á fjórum sviðum:

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 tungutækni.
3. Fyrirtæki verði styrkt til þess að þróa afurðir tungutækni.
4. Menntun á sviði tungutækni og málvísinda verði efld.

Í þessu felst að komið verði upp þróunarmiðstöð í tungutækni sem verði falið að vinna með útgefendum og öðrum við að koma upp þeim grunnsöfnum tungumálsins sem þarf. Nauðsynlegt er að auk ríkisvaldsins standi hagsmunaaðilar eins og tölvufyrirtæki, útgefendur, þýðendur og aðrir að þróunarmiðstöðinni.

Í öðru lagi leggur nefndin til að fé verði lagt í rannsóknasjóð sem styrki rannsóknir og þróun á sviði tungutækni. Þar gæti hvort sem er verið um að ræða sérstakan sjóð, eða sjóðir Rannsóknarráðs Íslands yrðu styrktir með fé eyrnamerktu til þessa iðnaðar. Sjóðurinn verði tvískiptur eins og Rannsóknasjóður Rannsóknarráðs er nú, og veiti annars vegar fé til hagnýtra grunnrannsókna, sem gagnist iðnaðinum til lengri tíma litið, og hins vegar til þróunarverkefna fyrirtækja, einkum til þess að smíða tungutól.

Nauðsynlegt er að fjárstuðningurinn geti nýst sem mótframlag á móti styrkjum frá Evrópusambandinu þar sem í Evrópu er ein helsta uppspretta þekkingar á sviðinu og verkefni fimmtu rammaáætlunar Evrópusambandsins gætu skipt miklu fyrir þróunina hér á landi, bæði hvað varðar sambönd og fé.

Þá telur nefndin nauðsynlegt að menntun á þessu sviði verði efld og leggur til að komið verði upp stuttu hagnýtu námi í máltækni og meistaranámi í tungutækni.

Heildarkostnaður á ári, við átakið sem lagt er til, yrði því:

Þróunarmiðstöð	25 til 50 MKR
Rannsókna- og þróunarsjóður	150 MKR
Sérstakur styrkur til stærri alþjóðlegra verkefna	30 MKR
Stutt hagnýtt nám í máltækni	10 MKR
Meistaranám í máltölvun	10 MKR
Alls	**225 til 250 MKR á ári**

Þetta kann að þykja allmikið fé og er það vissulega, en mat nefndarinnar er að áætlunin sé mjög hófleg og raunhæf og sé mikið úr henni dregið muni hún ekki ná tilætluðum árangri. Það hefur sem sagt ekki verið gert ráð fyrir órökstuddum niðurskurði í þessari áætlun.

Áríðandi er að starfsemin fari fljótt í gang. Stefna ber að því að verkefnið sé tímabundið og starfsemin verði sjálfbær á fimm til tíu árum.

Staða íslenskunnar

Íslenska og alþjóðleg upplýsingatækni

Miklar breytingar hafa orðið í samskiptum þjóða undanfarna hálfa öld. Þjóðríkin sem urðu til í Evrópu á fyrri hluta nítjándu aldar með samruna smáríkja og greifadæma eru nú of lítil fyrir þau samskipti sem eðlileg teljast. Þessi ríki eru því að sameinast í Evrópusambandinu. Samruninn er afleiðing þess að flutningur fólks og varnings, frétta og menningarstrauma er orðinn greiður um miklu stærra svæði en áður var. Ástæðan er að sjálfsögðu bættar samgöngur á landi og í lofti og gífurleg aukning fjarskipta og fjölmiðlunar.

Menning ræðst mikið af sjónvarpi og öðrum alþjóðlegum fjölmiðlum og verður sífellt einsleitari um allan heim

Smátt og smátt er að byggjast upp alþjóðlegt samfélag á flestum sviðum. Menning ræðst mikið af sjónvarpi og öðrum alþjóðlegum fjölmiðlum og verður sífellt einsleitari um allan heim. Viðskipti færast í auknum mæli á hendur stórra alþjóðlegra fyrirtækja sem starfa í fjölda landa. Samskiptakerfin sem fylgja verða sífellt alþjóðlegri. Tungumálin fylgja í kjölfarið, enska verður æ sterkari í öllum samskiptum. Nýlegt dæmi um það er að stóru norsku fyrirtækin Norsk Hydro og Statoil ákváðu að enska, en ekki norska, yrði samskiptatunga fyrirtækjanna.

Svipuð staða hefur oft komið upp áður. Latína var heimsmál á meðan styrkur Rómaveldis og síðar katólsku kirkjunnar var mikill. Alþjóðleg póstkerfi voru til skamms tíma á frönsku þar sem auðveldast var að hafa eitt tungumál fyrir alþjóðlegt kerfi póstsins. Evrópusambandið reynir hins vegar að nota tungur þjóðríkjanna. Það er erfitt og kostnaðarsamt og gæti vel orðið sá þáttur sem frekari stækkun svæðisins strandar á.

Hví skyldi fámenn þjóð hafa fyrir því og leggja í það ærinn kostnað að gera tungumál sitt hæft til notkunar í alþjóðlegu upplýsingaþjóðfélagi?

Við þessar aðstæður vakna eðlilega spurningar um hagkvæmni þess að nota íslenska tungu í alþjóðlegu umhverfi. Hví skyldi fámenn þjóð hafa fyrir því og leggja í það ærinn kostnað að gera tungumál sitt hæft til notkunar í alþjóðlegu upplýsingaþjóðfélagi? Hví notar hún ekki alþjóðlega málið ensku og kemur sér þar með hjá kostnaði og umstangi? Hví skyldi samhæfing tungumála heimsins ekki fylgja auknum samskiptum og meiri samhæfingu menningar?

Auðvelt er að afgreiða þessar spurningar á þjóðernislegum forsendum og tala um forna menningu og frægð og nauðsyn þess að geta lesið Njálu og Eglu á því máli sem þær voru skrifaðar á og vitna því til stuðnings í Jón Sigurðsson og Fjölnismenn. Þótt slík rök séu góð og gild og hafi gagnast Íslendingum ágætlega í sjálfstæðisbaráttu sinni á síðustu öld verður málið samt rætt hér á öðrum forsendum. Sumir kunna að kalla þær forsendur kaldar og efnahagslegar, en einnig má segja að þær sé raunhæfar og nútímalegar.

Ástæða þykir til að taka þetta fyrir hér þar sem þegar hafa komið upp ýmis dæmi um að einstaklingar og fyrirtæki hér á landi hafi ákveðið að vinna frekar á alþjóðlega málinu en þjóðtungunni til þess að spara sér fyrirhöfn og kostnað og í mörgum tilfellum tryggja betri og öruggari afgreiðslu mála. Þeir líta því á ensku sem „raunhæfan valkost” eins og það er stundum nefnt. Um slíkar ákvarðanir hefur gjarna orðið ágreiningur.

Stærsta fjarskiptafyrirtæki landsins tók upp nýtt viðskiptakerfi og þýddi það ekki á íslensku, heldur bauð starfsmönnum námskeið í ensku svo þeir réðu við að vinna við kerfið.

Nefna má af handahófi nokkur nýleg dæmi. Stærsta fjarskiptafyrirtæki landsins tók upp nýtt viðskiptakerfi og þýddi það ekki á íslensku, heldur bauð starfsmönnum námskeið í ensku svo þeir réðu við að vinna við kerfið. Háskóli Íslands krefst þess að umsóknir um stöður séu á ensku til þess að hægt sé að fá erlenda menn til þess að meta þær. Símsvarar í íslenskum stofnunum eru á ensku og virka þar með jafnt fyrir innlenda menn og erlenda.

Slík notkun ensku getur verið mikill kostur í fyrirtækjum og stofnunum þar sem samskipti eru mikil við útlönd. Í öllum tilfellum má til sanns vegar færa að hagkvæmara sé að nota ensku en íslensku. Þegar ákveða skal hvort velja eigi allmiklum fjárhæðum til þess að laga tól upplýsingatækninnar að íslensku kemur óhjákvæmilega upp spurningin um hvort ekki sé hagkvæmara að upplýsingatækni á Íslandi verði á ensku.

Nefndin hefur komist að þeirri niðurstöðu að allar líkur séu á að samskiptakerfi upplýsingaþjóðfélagsins verði fjöltyngd á næstu árum og áratug en verði ekki eingöngu á ensku

Nefndin hefur íhugað þetta mál og komist að þeirri niðurstöðu að allar líkur séu á að samskiptakerfi upplýsingaþjóðfélagsins verði fjöltyngd á næstu árum og áratug en verði ekki eingöngu á ensku. Þessu ræður m.a. sterk staða stærri málsvæða í Evrópu, svo sem franska og þýska málsvæðisins. Á þessum svæðum er mjög ákveðin stefna að nota þjóðtunguna í hugbúnaði og öðrum verkfærum upplýsingatækninnar. Ólíklegt verður að teljast að þessi málsvæði hætti að nota tungumál sitt á næstu árum.

Stefna Evrópusambandsins á þessu sviði er einnig sú að ýta undir fjöltyngi. Í bæklingnum *Language and Technology* sem ESB gaf út 1996 segir um þetta: „We should see Europe's linguistic diversity not as a weakness, however, but as one of its great strengths. National and regional differences — which reflect and are reflected by language — lead to a rich diversity of attitudes and approaches to solving problems and creating solutions. This leads to the creation of a wider variety of products, in many languages."

Því má búast við að samskiptakerfi upplýsingaþjóðfélagsins verði margtyngd. Séu þau á annað borð hönnuð til þess að nota fleiri en eitt tungumál ætti í framtíðinni fremur að verða auðveldara en erfiðara að laga íslensku að hinni alþjóðlegu upplýsingatækni. Verði því stefna stjórnvalda áfram að nota íslensku í upplýsingatækni ætti það að vera mögulegt. Það verður þó ekki gert án verulegs átaks til að styrkja stöðu tungutækni á Íslandi og það mun kosta fé og vinnu.

Staða íslenskrar tungu á alþjóðlegum markaði

Íslenska er virk þjóðtunga sem notuð er í öllum samskiptum og viðskiptum þjóðarinnar

Ekki fer á milli mála að fáir tala íslensku. Við samanburð við aðrar tungur þarf hins vegar að taka til greina að íslenska er virk þjóðtunga sem notuð er í öllum samskiptum og viðskiptum þjóðarinnar. Þýðing hennar er því mun meiri en tungumála sem fleiri nota en eru aukatungur þjóða, eða tungur þjóðflokka sem eru minnihlutahópar í stærri þjóðfélögum.

Hér eru fleiri tölvur á hvern íbúa en í flestum öðrum löndum og fleiri tengingar við Netið en víðast hvar

Í sambandi við tungutækni þarf líka að taka með í reikninginn að íslensk upplýsingatækni er vel þróuð miðað við það sem gerist á öðrum málsvæðum. Hér eru fleiri tölvur á hvern íbúa en í flestum öðrum löndum og fleiri tengingar við Netið en víðast hvar. Sé miðað við aðrar þjóðir ætti því tungutækni að vera meira notuð hér á landi en fjöldi landsmanna gefur ástæðu til að ætla.

Því miður er það svo að á Íslandi hefur virðing fyrir hugverkum verið fremur lítil og viðgengist hefur að hugbúnaður sé tekinn ófrjálsri hendi. Þetta minnkar að sjálfsögðu markað fyrir hugbúnað og veikir stöðu íslensku þegar semja skal við erlenda framleiðendur um þýðingu á forritum og aðlögun þeirra að íslensku.

Þetta er verulegt vandamál sem þarf að ráða bót á. Það er ekki við hæfi að þjóð sem að verulegu leyti hyggst byggja framtíð sína á hugverkum virði ekki rétt annarra sem hafa tekjur sínar af hugverkum. Nýlega gerði Björn Bjarnason menntamálaráðherra samning við Microsoft um þýðingu á Windows hugbúnaði þar sem lofað er átaki til að hindra stuld hugverka hér á landi.

Markaður fyrir vörur sem eru tengdar tungumáli er því lítill á Íslandi

Þegar á allt er litið er samt erfitt að komast fram hjá þeirri staðreynd að þjóðtungur Evrópu tala tíu til hundrað sinnum fleiri en tala íslensku. Markaður fyrir vörur sem eru tengdar tungumáli er því lítill á Íslandi miðað við það sem gerist hjá öðrum þjóðum sem nú eru að laga sig að tungutækni. Því er hætt við að framleiðendur reyni að komast hjá aðlögun að íslensku.

Afstaða Íslendinga skiptir þó miklu máli í þessu sambandi. Sé það ljóst að hér á landi er þess krafist að tæki og forrit séu löguð að tungunni, og að án þess að það sé gert sé íslenskur markaður að einhverju eða öllu leyti tapaður, mun það ýta undir að framleiðendur taki íslenskar sérþarfir með í vörur sínar. Fyrrnefndur samningur við Microsoft sýnir skýrt stefnu íslenskra stjórnvalda og styrkir kröfuna um að tæki og hugbúnaður falli að íslenskri tungu.

Notkun íslensku

Það dugir ekki að orðin séu til ef tungumálið er ekki gjaldgengt

Undanfarna áratugi hefur verið unnið mikið starf við að íslenska orðaforða ýmissa fræðigreina. Þar hefur oft tekist vel til og starfið borið árangur. En nú hafa aðstæður gerbreyst. Málið snýst ekki lengur um það að ekki séu til íslensk fag- og fræðiorð á tilteknum sviðum. Það snýst þess í stað um það að ekki er lengur hægt að nota neina íslensku, ekki heldur algengu orðin sem eru til í málinu, við ýmsar aðstæður í tölvuheiminum; forritin eru erlend og skilja ekki málið. Það dugir ekki að orðin séu til ef tungumálið er ekki gjaldgengt.

Hér er því komin upp ný staða sem ekki á sér hliðstæðu fyrr í málsögunni. Þarna er kominn til mikilvægur þáttur í daglegu lífi venjulegs fólks, þar sem móðurmálið er ónothæft. Þarna er þrennt sem spilar saman, og það skapar hættuna. Um er að ræða mikilvægan þátt, en ekki eitthvert aukaatriði; þessi þáttur snertir daglegt líf, en kemur ekki bara fram einstöku sinnum, við einhverjar sérstakar aðstæður; og þetta á við venjulegt fólk, allan almenning, en ekki eingöngu sérfræðinga á einhverju þröngu sviði.

Líklegt er að málið gæti varist samspili tveggja þessara þátta, en þegar allir þrír koma saman kann tungumálið að vera í hættu. Þótt einstöku stéttir, t.d. flugmenn, tali ensku að einhverju marki eða sletti ensku í daglegum störfum hefur það ekki í för með sér verulega hættu fyrir tunguna, því að þar er bæði hópurinn og aðstæðurnar takmarkandi.

Það er alþekkt að dauðastríð tungumála hefst einmitt þegar aðstæður af þessu tagi koma upp.

Það er alþekkt að dauðastríð tungumála hefst einmitt þegar aðstæður af þessu tagi koma upp; þegar mál er ekki lengur nothæft við allar aðstæður í hversdagslegu lífi almennings. Móðurmálið verður þá víkjandi, það er aðeins hæft til heimabrúks en ekki til neinna alvarlegra hluta. Við slíkar aðstæður hrekkur jafnvel ríkulegur bókmenntaarfur og öflugt nýyrðastarf skammt. Unga kynslóðin sér þá ekki lengur tilgang í að læra málið, heldur leggur alla áherslu á að tileinka sér erlent mál, enskuna, sem best.

Málið á sér þá ekki viðreisnar von, og hlýtur að hverfa sem lifandi tungumál almennings á tiltölulega stuttum tíma. Íslenskt málsamfélag er ekki enn komið á þetta stig, en verði ekkert að gert gæti verið stutt í það. Breytingarnar á þessu sviði hafi verið miklu örari nú allra síðustu ár en flestir gera sér grein fyrir.

Markmið

Ekki er ólíklegt að þetta verði til þess að öll forrit sem almenningur notar við störf sín og nám verði þýdd á íslensku

Margt þarf til svo íslensk tunga verði notuð í íslenska upplýsingasamfélaginu. Nýlega var undirritaður samningur við Microsoft fyrirtækið um að það þýði Windows98 stýrikerfið á íslensku. Það er þrep á þeirri leið. Við sama tækifæri rituðu fyrirtækið og menntamálaráðherra undir viljayfirlýsingu um að skrifstofuforrit fyrirtækisins, Office, verði í framtíðinni einnig þýtt á íslensku. Ekki er ólíklegt að þetta verði til þess að öll forrit sem almenningur notar við störf sín og nám verði þýdd á íslensku.

Þetta eru stór skref, en eru þó aðeins fyrstu skrefin á langri leið. Næsta skref þarf að verða að tól tungutækninnar verði fær um að vinna með íslensku og á íslensku. Með því er átt við að til verði tól til þess að vinna með íslenskan texta, tól til að leiðrétta stafsetningu og málfræði, skipta orðum milli lína og svo framvegis; einnig að íslensk orðabók og samheitaorðabók verði öllum aðgengilegar á rafrænu formi, að upplýsingar um beygingar orða verði í hverri tölvu og fleira.

Almennt talað þarf að koma upp þeim tólum sem munu auðvelda Íslendingum að vinna á tölvum á íslensku og vinna með íslenskt mál með tungutæknitólum. Stígi þjóðin ekki þetta skref er hætt við að erfitt verði að nota íslenska tungu í upplýsingaþjóðfélaginu.

Nú þegar stýrikerfi verða þýdd á íslensku mun það verk njóta þess að á undanförnum árum hafa verið smíðuð fjölmörg ný íslensk tækniorð. Því eru til hér á landi viðamikil íðorðasöfn, meðal annars nýtt tölvuorðasafn. Þessi söfn hafa orðið til vegna fórnfúsrar vinnu áhugafólks um varðveislu tungunnar. Þau munu létta starfið við þýðingar stýrikerfa.

Áhugi Íslendinga á tungu sinni hefur skilað sér í málhreinsun og nýyrðasmíð, en ekki í starfsemi á sviði tungutækni og því stendur íslenskan verr en tungur nágrannaþjóðanna

Því miður er staðan öll miklu verri þegar kemur að því að búa til þau tól tungutækninnar sem voru nefnd hér að framan. Áhugi Íslendinga á tungu sinni hefur skilað sér í málhreinsun og nýyrðasmíð, en ekki í starfsemi á sviði tungutækni og því stendur íslenskan verr en tungur nágrannaþjóðanna.

Með því að líta í Word ritvinnslukerfið á tölvunni sinni geta áhugasamir lesendur séð þetta. Smellið á „Tools". Þar má leita að stafsetningar- og málfræðivillum, samheitaorðabók (thesaurus) er þarna og sjálfvirk skipting orða á milli lína (hyphenation). Forritið býður líka upp á að draga fram aðalatriði textans. Margt af þessu má hvort sem er gera jafnóðum og skrifað er eða eftir

að textinn hefur verið saminn. Öll þessi tól fylgja sjálfkrafa og ókeypis með ritvinnsluforritinu og eru talin sjálfsögð; þau vinna hins vegar aðeins með enska tungu.

Sams konar tól eru til fyrir flestar tungur Evrópu. Þau má kaupa sem aukahluti og margir sem rita mikið á erlendum málum eiga safn slíkra tóla, t.d. fyrir dönsku, þýsku og frönsku. Í nágrannalöndunum þykir sjálfsagt að tól fyrir viðkomandi tungu fylgi með forritum, og ekki aðeins ritvinnsluforritum heldur öllum forritum. Nánast ekkert af þessu er til fyrir íslensku og það sem verra er: það er mjög lítill grunnur til að byggja á. Málsöfnin sem eru hráefnið fyrir þau eru ekki til.

Þetta kann að koma lesandanum á óvart því þjóðin trúir því gjarna að við höfum unnið mikið fyrir tunguna, meira en aðrar þjóðir

Þetta kann að koma lesandanum á óvart því þjóðin trúir því gjarna að við höfum unnið mikið fyrir tunguna, meira en aðrar þjóðir. Að hluta til er það rétt og gott dæmi er nýyrðasmíð. Á öðrum sviðum er þjóðin hins vegar illa stödd. Mjög lítið er til af góðum orðabókum, bæði íslensk-íslenskum og milli íslensku og annarra mála. Viðamikil og flokkuð söfn fjölbreyttra texta eru ekki heldur til. Hér hefur hvorki farið fram kennsla né rannsóknir á sviði máltölvunar eða tölvufræðilegra málvísinda og fáir kunna því þá tækni sem þarf til að búa til tól tungutækninnar. Því verður mikið verk að búa slík tól fyrir íslensku.

Tungutækni

Hvað er tungutækni?

Tungutækni er tæknin við meðferð tungumálsins í tölvum og hugbúnaði. Þar er um að ræða að koma máli inn og út úr tölvum, meðhöndla það á ýmsan hátt í tölvum og hugbúnaði o.s.frv. Til greinarinnar telst líka notkun tungunnar til að hafa samskipti við tæki, stýra þeim til dæmis. Í því gæti falist að segja tölvunni að „opna Word", „kveikja á útvarpinu" í bílnum, hringja í vin sinn með því að segja við símann „hringdu í Jón" og annað slíkt.

Tungutæknin er því það sem kallast þverfagleg grein

Greinin er þannig nátengd tölvutækni og tölvuverkfræði. Hún byggist einnig á þekkingu á málvísindum og þjóðtungunni. Á þetta reynir t.d. mjög í villupúkum. Þá styðst fagið við ýmislegt úr sálfræði, skynjunarfræði og hljóðfræði, eins og hvernig fólk skilur tal, hvernig fólk myndar hljóð og orð. Til dæmis verða talgervlar ekki áheyrilegir án þess að beitt sé þekkingu á hljóðfræði og framburði. Að auki styðst fagið oft mjög við gervigreind, t.d. þegar reynt er að greina á milli orðalags með mismunandi merkingu. Tungutæknin er því það sem kallast þverfagleg grein.

Hagnýting tungutækninnar byggist á viðamiklum málrannsóknum af ýmsu tagi. Þær rannsóknir flokkast einkum undir tölvufræðileg málvísindi eða máltölvun (computational linguistics) og textamálfræði eða gagnamálfræði (corpus linguistics). Hagnýtingin byggist einnig á notkun háþróaðrar aðferðafræði tölvutækni og góðar lausnir munu byggjast á farsælli samtvinnun málvísinda og upplýsinga- og tölvutækni.

Hér á landi er tungumálið tengt þjóðernishyggju og frelsisbaráttu ungs þjóðríkis og er þar með tilfinningamál

Hér á landi er tungumálið tengt þjóðernishyggju og frelsisbaráttu ungs þjóðríkis og er þar með tilfinningamál. Það gerir tungutækninni stundum erfitt fyrir, t.d. hefur nýlega verið fjálglega rætt um hvernig rita eigi stafinn 'ð' og lítur fólk þá á málið ýmist frá tæknilegu eða þjóðernislegu sjónarhorni. Hér á landi hafa einnig margir ákveðnar skoðanir á hvað sé „rétt" í framburði og réttritun. Aðrir horfa meira til þess hvernig tungumálið gagnist sem samskiptamiðill.

Hvað varðar verklag kemur upp skoðanamunur á milli þeirra sem horfa á málið frá sjónarhorni hefðbundinna málvísinda og þeirra sem líta á það frá sjónarhorni tölfræði. Það skiptir miklu að vita af þessum mismun þegar fjallað er um tungutækni. Til dæmis liggur oft beint við fyrir þá sem ekki þekkja til að ætla að faginu sé best fyrir komið hjá málvísindamönnum því þeir þekki tunguna best. Þetta þykir þeim sem koma að málinu frá sjónarhorni tölvu- og verkfræði hins vegar ekki sjálfgefið og benda á að málvísindamenn þekki lítið til þeirra tölfræði- og tölvunarfræðilegu aðferða sem beitt er í tungutækni.

Farsælast er að líta á tungutækni sem nýjan iðnað í þekkingarþjóðfélaginu, iðnað sem þarf á mörgum greinum vísinda og fræða að halda

Þessi inngangur er hér vegna þess að hann skiptir máli þegar huga þarf að því hvar ný starfsemi á sviði tungutækni skuli sett niður. Farsælast er að líta á tungutækni sem nýjan iðnað í þekkingarþjóðfélaginu, iðnað sem þarf á mörgum greinum vísinda og fræða að halda, en er samt fyrst og fremst iðnaður. Þessu má líkja við að í matvælaiðnaði þarf mjög góða þekkingu á gerlafræði, en greinin stjórnast ekki af gerlafræði og þeim aðferðum sem notaðar eru í rannsóknum í gerlafræði.

Margar eða flestar atvinnugreinar hafa rannsóknastofnanir sem sinna fræðilegum verkefnum sem gagnast iðnaðinum í heild til lengri tíma litið. Áríðandi er að hafa í huga að starfsemi þessara stofnana á að gagnast iðnaðinum í heild og að þar á að líta til lengri tíma, lengra en hvert fyrirtæki gerir

daglega. Nefndarmenn álíta að tungutæknin þurfi slíka rannsóknastofnun. Grundvallaratriði er að þar er ekki um að ræða rannsóknastofnun í málvísindum eða íslenskri tungu, heldur í tungutækni.

Hér verður þetta nefnt þróunarmiðstöð til þess að leggja áherslu á að þar á fyrst og fremst að fara fram þróun en ekki vísindarannsóknir

Innan slíkrar stofnunar þyrfti að vera þekking á ýmsum þeim fögum sem nefnd voru hér að framan, eins og t.d. málvísindum og íslensku; ekki djúp fræðileg þekking á hverju fagi, heldur almenn og praktísk þekking. Eins og aðrar rannsóknastofnanir atvinnuvega mundi þessi stofnun leita til sérfræðinga í hverju fagi þegar á dýpri þekkingu þarf að halda. Hér verður þetta nefnt þróunarmiðstöð til þess að leggja áherslu á að þar á fyrst og fremst að fara fram þróun en ekki vísindarannsóknir.

Verklag og aðferðir tungutækni

Viðfangsefnum tungutækni má skipta í tvennt eftir því hvort um er að ræða tól fyrir hið ritaða mál eða hið talaða mál. Vinnuaðferðir eru nokkuð ólíkar sem og hráefnið sem notað er.

Hvað varðar hið ritaða mál og almenna uppbyggingu tungumálsins þarf að greina texta af ýmsu tagi og koma upp málsöfnum og skrám. Forsendur fyrir því að unnt sé að búa til tól sem skili góðum árangri og notendur verði sáttir við er að byggt sé á mjög stóru og fjölbreyttu textasafni. Texta í þetta safn þarf að velja af kostgæfni og gæta þess að þar sé að finna góð dæmi um sem allra flest tilbrigði íslensks máls; formlegt mál og óformlegt, ritgerðir og samtöl, blaðamál og skáldverk, fræðitexta, tölvupóst, lagamál, auglýsingatexta, stjórnmálaumræður o.s.frv.

Orðasöfn og önnur söfn, sem eru hráefni tungutækninnar, eru miklu stærri en hingað til hefur þekkst í íslenskum málfræðirannsóknum

Orðasöfn og önnur söfn, sem eru hráefni tungutækninnar, eru miklu stærri en hingað til hefur þekkst í íslenskum málfræðirannsóknum og því þarf að nýta tölvutækni til hins ýtrasta við gerð þeirra. Söfnin þurfa einnig að miðast við tunguna eins og hún er á hverjum tíma. Orðaforðinn er stöðugt að breytast og aukast, og því eru söfn sem eru eldri en tíu eða tuttugu ára að jafnaði lítils virði. Þetta er vegna þess að tól tungutækninnar þurfa að vinna með hið lifandi mál.

Þessi textasöfn þarf síðan að greina mjög nákvæmlega á ýmsan hátt. Gera þarf nákvæma beygingarlýsingu allra orða, lýsingu á setningafræðilegri stöðu, merkingarlýsingu o.fl. Þessa greiningu þarf að setja fram á ákveðinn staðlaðan hátt þannig að unnt sé að nýta hana í ýmiss konar forritum og tólum, s.s. leiðréttingarforritum, þýðingarforritum, orðabókum o.s.frv. Þetta má gera með því að taka til greiningar tölvutækan texta úr bókum, blöðum, tölvupósti, lagamáli o.s.frv. Um þetta er nánar fjallað í viðauka 3.

Þar þarf að safna miklum fjölda hljóðsýna sem síðan eru greind

Aðferðir við gerð tóla sem eru notuð í töluðu máli, hljóðtóla, eru öðru vísi. Þar þarf að safna miklum fjölda hljóðsýna sem síðan eru greind og ýmsir eiginleikar dregnir fram sem notaðir eru í hljóðtólin. Hér á landi er engin starfsemi á þessu sviði né á skyldum fræðasviðum og lítil eða engin þekking á aðferðunum. Einnig virðist sem þau fyrirtæki sem að þessu vinna hafi hvert um sig þróað ákveðnar aðferðir við greiningu; að vinnsla sé ekki stöðluð, né heldur sé grunnurinn sameiginlegur.

Þessar staðhæfingar þarf þó að taka með fyrirvara og athuga betur. Ekki er auðvelt að sjá hve langt er hægt að komast á þessu sviði tungutækni án samvinnu við erlenda aðila og þá líklega helst fyrirtæki.

Þess ber að geta að hljóðtólin nýta sér textatól tungunnar. Til dæmis mundi tölvan við greiningu tals hafa aðgang að forriti til leiðréttingar málfræði og nota það til þess að greina á milli tveggja möguleika sem hún teldi báða koma til greina. Orðmyndirnar *bíður* (af *bíða*) og *býður* (af *bjóða*) hljóma t.d. eins, en greining á setningafræðilegu og merkingarlegu umhverfi þeirra gerir mögulegt að ákvarða um hvora sögnina er að ræða. Þetta mundi að mestu vera sama forrit og notað væri til þess að greina málfræði í rituðu máli. Um þetta er nánar fjallað í viðauka 4.

Áður en langt er haldið í gerð hljóðtóla þarf að vera kominn góður grunnur textatóla

Af þessu leiðir að áður en langt er haldið í gerð hljóðtóla þarf að vera kominn góður grunnur textatóla. Eins og rætt er á öðrum stað í þessari skýrslu eru þau varla til fyrir íslensku enn sem komið er. Í ljósi þessa virðist farsælast að byrja á að byggja upp textatólin. Svolítið er til í landinu af málsöfnum og tólum sem má nýta og rétt er að reyna að fá aðgang að þeim og nýta það sem hægt er. Síðan þyrfti að auka við söfnin og samræma þau.

Erlendis er víða samvinna um slík söfn. Sem dæmi má nefna The Linguistic Data Consortium í Bandaríkjunum (sjá vísun í vefsíðu í viðauka 7). Að því koma opinberir aðilar og einkafyrirtæki sem hafa aðgang að textum á tölvutæku formi. Þar er um að ræða stjórnsýslu, blaða- og bókaútgefendur, rannsóknastofnanir og aðra slíka. Samtökin búa til, safna og dreifa málsöfnum, orðalistum og öðru hráefni fyrir tungutækni. Hér á landi eiga nokkrir aðilar slíkt efni og má þar t.d. nefna Orðabók Háskólans, Morgunblaðið, Alþingi og bókaútgefendur.

Hvaða vandamál tungutækni leysast sjálfkrafa og hver ekki?

Nú eru liðnir tveir áratugir síðan tölvuöld hófst á Íslandi. Frá upphafi hennar hefur þurft að laga tæki og hugbúnað að íslenskri tungu. Á fyrstu árum tölvualdar þurfti mikið fyrir þessu að hafa. Flestar tölvur notuðu sjö bita stafakóða og gátu ekki skráð íslenska stafi, skjám þurfti að breyta og prentara þurfti að laga til á ýmsan hátt. Sama átti við um hugbúnað; erlendur hugbúnaður síaði oft út íslenska stafi. Um stafatöflur og íslensku er nánar fjallað í viðauka 2.

Smátt og smátt réð upplýsingatæknin við íslenska tungu

Eftir allmiklar umræður og deilur varð hér á landi samkomulag um að tæki og forrit þyrftu að geta skráð íslenska stafi. Innflytjendur létu breyta tækjum og forritum sem þeir seldu og smátt og smátt réð upplýsingatæknin við íslenska tungu. Tölvupóstur réð þó lengi vel ekki við íslenska stafi og muna flestir það sem stundum var nefnt enskíslenska, þar að segja *th*, *ae* og annað slíkt í tölvuskeytum.

Mikið af því sem gert var á þessum árum var sértækt fyrir íslensku, íslenskum stöfum var skotið inn í göt í stafatöflum skjáa og prentara o.s.frv. Slík vinna nýtist aðeins fyrir ákveðna gerð búnaðar og úreldist þegar nýjar útgáfur tækja og forrita koma á markað. Þetta er því dýrt og vinnufrekt, en á fyrstu árunum voru vörumerki fá og innflytjendur fáir og nokkuð langt var á milli nýrra útgáfna af forritum og tækjum þannig að þetta gekk upp að mestu.

Framleiðendur fóru að gera ráð fyrir mismunandi tungumálum í tækjum og hugbúnaði

Smám saman leystust sum vandamál íslenskunnar af „sjálfu sér". Því réð einkum tvennt:

- Tölvur urðu öflugri og réðu við stærri stafasett og fleiri möguleika.
- Iðnaðurinn breyttist frá því að vera bandarískur í að vera alþjóðlegur og framleiðendur fóru að gera ráð fyrir mismunandi tungumálum í tækjum og hugbúnaði.

Þetta var reyndar eins gott, því vörumerkjum fjölgaði sífellt og líftími tækja og hugbúnaðar styttist, þannig að erfiðara og erfiðara varð að beita þeim íslensku sérlausnum sem notaðar voru í upphafi tölvualdar. Einstaka atriði urðu þó eftir. Til dæmis er ekki enn almennt viðurkennt að viðmót forrita þurfi að vera á íslensku og fæst forrit hafa íslenskt viðmót. Undantekningar eru nokkrar, t.d. hefur viðmót hugbúnaðar á Macintosh ávallt verið á íslensku.

Windows og önnur forrit frá Microsoft eru hins vegar öll á ensku og hafa ekki verið fáanleg á íslensku þrátt fyrir allmikinn þrýsting íslenskra stjórnvalda. Á því hefur þó nýlega verið ráðin bót. Á síðasta ári birtist fróðleg grein um baráttu Íslendinga fyrir því að fá Windows kerfið þýtt í dagblaðinu Los Angeles Times. Greinin er birt aftan við þessa skýrslu.

Af sögunni má draga þá ályktun að sum vanda-mál tungutækn-innar leysist sjálfkrafa

Af sögunni má draga þá ályktun að sum vandamál tungutækninnar leysist sjálfkrafa vegna öflugri tækni og breyttrar stefnu framleiðanda gagnvart erlendum mörkuðum, en önnur verði Íslendingar að leysa sjálfir. Hér skiptir höfuðmáli að reyna að tryggja að á öllum sviðum sé tekið tillit til íslenskrar tungu og sérkenna hennar strax við framleiðslu búnaðar. Einnig þarf að ganga hart fram í að koma íslensku inn í alþjóðlega staðla.

Almennt þarf að nota altækar lausnir í stað sértækra

Almennt þarf að nota altækar lausnir í stað sértækra. Þetta er eina stefnan sem getur tryggt að íslenska sé nothæf í upplýsingatækni í framtíðinni. Sérlausnir eru dýrar, þær hafa stuttan endingartíma og eru mjög erfiðar og mannfrekar í viðhaldi og þeim ætti ekki að beita nema í brýnustu neyð. Þetta er stefna íslenskra stjórnvalda og mikil vinna er lögð í að koma þessum skilaboðum á framfæri erlendis, t.d. í vinnu staðlaráða.

Hér á landi hafa á síðustu áratugum verið þróuð ýmis gagnleg tungutæknitól svo sem orðskiptiforrit, villupúkar og orðasöfn. Því miður er flestum þessum tólum sammerkt að þau urðu skammlíf. Það er blóðugt að þurfa á slíkum tólum að halda og vita að þau eru til, en eru ónothæf vegna þess að þau byggðust á sérlausnum. Slíku þarf að halda í lágmarki í framtíðinni.

Í ljósi þessarar reynslu þarf að líta á ný vandamál. Skilja þarf á milli þeirra vandamála sem ný tækni mun leysa sjálfkrafa eða sem verða leyst með öflugu starfi í staðlamálum, og þeirra vandamála sem eru sértæk fyrir íslenska tungu og Íslendingar þurfa að leysa sjálfir. Þetta er ekki alltaf auðvelt. Stundum þarf að sýna þolinmæði gagnvart vandamálum í fyrri flokknum. Það getur borgað sig að bíða eftir að tækni þróist, og vera á meðan án ákveðinna lausna.

Vandamál úr seinni flokknum er oftast best að ráðast á sem fyrst. Staðlamálum er alltaf best að taka á eins fljótt og kostur er. Það segir sig sjálft að ekki er víst að allir séu sammála um í hvaða flokk vandamál falla, og deilur geta risið um hvort þjóðin sé að missa af tækifærum vegna aðgerðaleysis eða hvort hún sýni góða búmennsku með því að bíða eftir almennum lausnum.

Fjölmörg einfaldari vandamál eru vel þekkt og bíða þess einungis að verða leyst

Nú er staðan sú í íslenskri tungutækni að fjölmörg einfaldari vandamál eru vel þekkt og bíða þess einungis að verða leyst. Önnur vandamál er erfitt að sjá hvort muni leysast almennt, að hluta til eða að öllu leyti, með nýrri og betri tækni og þekkingu, eða hvort Íslendingar verði að leysa þau sjálfir. Einnig getur verið erfitt að ákveða á hvaða tæknistigi á að fara af stað með að leysa málið. Sé farið of snemma af stað getur sú staða komið upp að lausn fáist sem fljótlega verður úrelt og síðan sé ekki til fé eða fólk til að leysa málið aftur. Þjóðin situr þá uppi með lélega lausn.

Sé farið of seint af stað getur það hins vegar orðið til þess að vandamálið verði aldrei leyst. Þekkingin sem þarf verður aldrei til í landinu. Einnig getur svo farið að fólk finni lakari lausn og sætti sig við hana. Dæmi um hið síðastnefnda er að nú er svo komið að mikill hluti þeirra Íslendinga sem nota tölvur kýs ef til vill frekar að nota enskt viðmót forrita fremur en íslenskt; menn kunna á það enska og vilja ekki skipta yfir í eitthvað nýtt og ókunnuglegt.

Í umræðum um hljóðtól tungutækni hafa komið fram þau rök að tími verksins bæði komi og fari, það verði einfaldlega of seint að vinna verkið sé það ekki unnið strax

Í umræðum um hljóðtól tungutækni, það er að segja þau tól sem breyta tali í texta og texta í tal, hafa komið fram þau rök að tími verksins bæði komi og fari, það verði einfaldlega of seint að vinna verkið sé það ekki unnið strax, þjóðin missi af vagninum, vagni sem kemur einu sinni. Þetta hefur nefndin íhugað og rætt. Tvennt getur valdið því að of seint verði að vinna verkið.

1. Fyrri möguleikinn er að menningarlegi vagninn fari hjá, ef svo má segja. Dæmi um það er íslenskun á viðmóti forrita sem nefnd var hér að ofan. Annað dæmi gæti verið að allir kynnu því svo vel að tala ensku við tölvuna sína að þeir vildu ekki tala við hana íslensku, þótt forrit væru til sem gerðu það.

2. Síðari möguleikinn er að tæknin og vinnuferlið við að leysa vandann komist á eitthvert það stig að ekki verði snúið til baka, það sé ekki hægt að komast aftur í startholurnar til þess að vinna frumvinnuna. Tæknin verði of langt komin til þess.

Slík hljóðtól eru ennþá aðeins til fyrir nokkur helstu tungumál Vesturlanda og mikill fjöldi tungumála á eftir að verða sér úti um slík tól

Nefndin hefur rætt þetta. Erfitt er að meta áhættuna við fyrri möguleikann. Síðari möguleikann hefur hún einkum rætt í sambandi við hljóðtól og á erfitt með að sjá að þetta sé raunverulegur vandi. Vinnuferlið (sjá viðauka 4) við gerð slíkra tóla er í grófum dráttum þannig að safnað er miklu af hljóðsýnum, þ.e. dæmum um tal fólks. Hljóðsýnin eru síðan greind og síuð með ákveðnum aðferðum. Niðurstaðan er loks notuð til þess að búa til hugbúnaðarkerfi sem getur til dæmis breytt tali í texta. Slík hljóðtól eru ennþá aðeins til fyrir nokkur helstu tungumál Vesturlanda og mikill fjöldi tungumála á eftir að verða sér úti um slík tól. Því er ólíklegt annað en að um nokkurt skeið verði gott aðgengi að tækni til að búa tólin til og því líklegt Íslendingar geti unnið verkið síðar.

Í öðru lagi er eðli tækni venjulega þannig að hún er dýrust fyrst, á meðan verið er að ná inn kostnaði af rannsóknum og tilraunum, en síðan lækkar verðið og tæknin verður aðgengilegri. Þannig finnast smátt og smátt ódýrari lausnir og það sem er erfitt í dag er venjulega auðveldara á morgun. Það er erfitt að finna dæmi um aðra hegðun tæknilegra framfara.

Það kunna að vera viðskiptaleg rök fyrir því að líkur séu á að þjóðin missi af vagninum. Dæmi um það er að á ákveðnum tíma er verið að vinna að ákveðnum verkum og á þeim tíma er hægt að vinna með öðrum að verkinu. Nýlega tilkynnti Landssíminn hf um stafrænt sjónvarp sem komið verður á hér á landi í samstarfi við norræn fyrirtæki. Þar kemur fram að þeir sem að þessu standa vilja gjarna líta á norrænu þjóðirnar sem einn markað og finna lausn sem nær því fram. Því verður kostnaður við að koma þjónustunni upp á Íslandi ekki aðgreindur frá öðrum kostnaði, valdar verða lausnir þar sem með litlum aukakostnaði er unnt að taka Ísland með o.s.frv. Við þetta verður sértækur kostnaður Landssímans ekki hár. Síðar gæti verið miklu dýrara að koma upp sérstakri íslenskri lausn.

Um þetta er mjög erfitt að dæma. Almennt virðist samt samkeppni í markaðsþjóðfélagi tryggja framboð. Í ljósi þess sem að framan var sagt um fjölda tungumála sem eiga eftir að byggja upp sín hljóðtól virðast nefndinni þessi rök ekki mjög sterk. Þau þarf hins vegar að íhuga vandlega.

Til dæmis virðist ljóst að þjóðin fer einhvers á mis með því að hafa ekki villupúka á tölvum sínum

Þau rök sem eru sterkust fyrir því að taka upp nýja tækni eru að þjóðin sé að fara einhvers á mis við að gera það ekki. Um getur verið að ræða efnahagslegan ávinning og hagræðingu, eða menningarlegan ávinning. Til dæmis virðist ljóst að þjóðin fer einhvers á mis með því að hafa ekki villupúka á tölvum sínum, vinna við tölvur verður erfiðari og tekur lengri tíma. Því er ekki vafi á að slíkt tól mundi skila einhverjum arði, þótt ekki sé ljóst hvort sá arður sé nægur til þess að greiða fyrir gerð villupúka við núverandi aðstæður.

Hljóðtól tungutækni þróast mjög ört þessa mánuði. Þróun þeirra verður þó að teljast skammt á veg komin og enn er markaður fyrir þau ekki stór. Þegar svo er sagt verður þó að horfa til framtíðar. Gerð íslenskra hljóðtóla mun taka einhver ár og ekki er ólíklegt að um það leyti sem hljóðtólin verða tilbúin hafi tækninni miðað áfram og af tólunum verði töluverð hagræðing.

Þessi rök eru sterk og hafa verið hugleidd. Álit nefndarinnar er að til þess að hljóðtól verði nýtanleg þurfi að vera til fjölmörg önnur tól sem notuð eru í ferlinu, svo sem leiðréttingarforrit fyrir réttritun og málfræði. Þau tól þurfi að búa til fyrst og án þeirra verði ekki haldið áfram á braut tungutækni hér á landi. Því sé mesta áhættan falin í því að gera ekkert. Svo fremi sem vinna sé hafin við málsöfn og tól tengd réttritun og málfræði sé ekki mikil hætta á að vagninn fari fram hjá þjóðinni fyrir fullt og allt.

Markaðsmál og fjármögnun

Markaður fyrir tungutækni á Íslandi

Nefndinni virðist ljóst að sem stendur er markaður fyrir tungutækni á Íslandi ekki nægilega stór til þess að sviðið geti borið sig fjárhagslega og þróast á þann hátt sem þarf til þess að tryggja stöðu íslenskrar tungu í upplýsinga-samfélaginu. Þessa niðurstöðu byggir nefndin á sögulegum staðreyndum sem raktar eru víða í þessari skýrslu. Eins og fram kemur í skýrslunni hafa orðið til allnokkur tungutæknitól fyrir íslensku á undanförnum árum, en það er þeim öllum sammerkt að þau hafa ekki náð nægilegri festu á markaði og fallið út aftur. Í sömu átt bendir að mörg undirstöðutól tungutækni hafa ekki verið þróuð á Íslandi.

Íslendingar hafa hingað til greitt fyrir sína íslensku

Ekki er víst að þannig þurfi þetta að vera til frambúðar. Íslendingar hafa hingað til greitt fyrir sína íslensku, ef svo má segja, útgáfa er mikil af bókum og blöðum og þjóðin greiðir fyrir það efni hærra verð vegna þess að efnið er á íslensku og markaðurinn er lítill. Líklega mun þjóðin smátt og smátt greiða þann kostnað sem hlýst af því að íslenska upplýsingatæknina, á svipaðan hátt og hún nú greiðir kostnað við íslenska útgáfu blaða og bóka. Nefndinni virðist samt sem áður að átak þurfi að gera til þess að koma tungutækninni á fæturna og að það verði ekki gert án stuðnings hins opinbera. Nefndin telur líklegt að slíkt átak muni borga sig, sé rétt staðið að verki.

Verðlagning

Í viðræðum um aðgengi almennings að orðabókum og fleiru hefur oft komið upp sú spurning hvort taka eigi gjald fyrir notkun á slíkum upplýsingum eða ekki. Tvö sjónarmið eru algengust. Sumum finnst sem allur aðgangur almennings að upplýsingum um íslenska tungu eigi að vera ókeypis. Hitt sjónarmiðið er að afla allra þeirra tekna sem hægt er fyrir aðgang að slíkum upplýsingum.

Rök þeirra sem vilja að aðgangur sé ókeypis eru einkum þau að áríðandi sé að aðgengi að upplýsingum um tunguna sé sem best. Það verði til þess að almenningur fletti upp í þessum upplýsingum og auki við þekkingu sína. Það hafi marga kosti. Þekking fólks á málinu aukist. Síður sé hætta á að málið einfaldist við það að fólk hikar við að nota orð og orðasambönd sem það er ekki visst um hvernig á að rita. Málvillur nái síður fótfestu þar sem auðvelt sé að fletta upp réttu máli. Þannig má lengi telja.

Rök þeirra sem vilja taka gjald fyrir aðgang að upplýsingum um tunguna eru hins vegar að allt kosti fé og fyrirhöfn og fyrir slíkt eigi að greiða. Tekjurnar auka möguleika á frekari starfsemi á sama sviði. Virðing fólks er meiri fyrir því sem það greiðir fyrir o.s.frv.

Báðar aðferðir er verið að reyna hér á landi. Íslensk málstöð selur aðgang að flestum íðorðasöfnum sínum. Orðabók Háskólans veitir ókeypis aðgang að sínu safni, Ritmálsskrá Orðabókar Háskólans. Lítil reynsla er komin á þetta enn þá og rétt er að fylgjast með hver hún verður og taka mið af því.

Ekki verði greitt fyrir einstaka fyrirspurnir og ókeypis aðgangur verði að söfnum vegna rannsókna.

Báðir aðilar hafa vafalaust eitthvað til síns máls. Nefndin telur æskilegt að aðgengi sé sem allra best að því sem til er af tungtæknitólum. Í sambandi við orðasöfn og annað slíkt gæti verið lausn að fyrir einstaka fyrirspurnir sé ekki greitt og ókeypis aðgangur sé að söfnum vegna rannsókna. Þegar gögnin eru notuð í viðskiptalegum tilgangi sé greitt fyrir gögnin.

Best er að greiðslur fyrir gögn falli saman við tekjur af vörunni sem gögnin eru notuð í. Því gæti verið heppilegra að semja um hlutdeild í væntanlegum tekjum af vörunni frekar en að gögn séu keypt í upphafi verks. Það fyrirkomulag mundi líka ýta undir að sem flestir spreyttu sig á framleiðslu á tólum sem nýta söfnin.

Verði til ýmis málsöfn sem fyrirtæki geta fengið aðgang að á hóflegum kjörum mun það ýta undir að til verði markaður hér á landi

Eins og þegar hefur verið nefnt hefur fram til þessa ekki verið markaður á Íslandi fyrir tól tungutækninnar. Kemur þar tvennt til; annars vegar eru notendur fáir og hins vegar þarf að vinna öll verk frá byrjun. Verði til ýmis málsöfn sem fyrirtæki geta fengið aðgang að á hóflegum kjörum mun það ýta undir að til verði markaður hér á landi með því að gera aðgengilegra, fljótlegra og ódýrara fyrir fyrirtæki að framleiða þessa vöru.

Fjármögnun

Nefndin hefur hugleitt hvaðan fé geti komið til tungutækni á Íslandi og satt að segja eru ekki margar matarholurnar. Fé á fjárlögum þyrfti að koma til, eins og þegar hefur verið nefnt, sérstaklega í upphafi. Rannsókna- og tækniáætlanir ESB styrkja verkefni tungutækni og Íslendingar eiga aðgang að þeim. Hingað til hefur lítið sem ekkert komið úr þessum sjóðum til íslenskra verkefna og er skýringin sú að á Íslandi hefur nánast engin starfsemi verið á þessu sviði.

ESB sjóðirnir styrkja verkefni allt að hálfu gegn mótframlagi umsækjenda

ESB sjóðirnir styrkja verkefni allt að hálfu gegn mótframlagi umsækjenda og styrkir þeirra eru háir miðað við það sem gerist hér á landi. Búast mætti við að verkefni á sviði tungutækni yrði styrkt með 75-250 MKR á þriggja ára tímabili. Þar sem þátttakendur eru að jafnaði frá a.m.k. þremur löndum gætu íslenskir þátttakendur vænst þess að fá úr slíku verkefni 10 til 30 MKR á ári í þrjú ár auk þess sem þeir hefðu að sjálfsögðu aðgang að og nytu þess sem aðrir þátttakendur gerðu í verkefninu.

Nú er að hefjast ný rannsókna- og tækniáætlun hjá ESB, fimmta rammaáætlunin, og verða fyrstu styrkirnir væntanlega veittir á vordögum 1999. Í þessari áætlun verður í tungutækni lögð áhersla á yfirfærslu þekkingar og færni (sjá tilvísanir í vefsíðu í viðauka 7) og því gætu Íslendingar haft af henni gott gagn. Ástæða er til þess að ítreka að slíkt verður ekki nema starfsemi sé á sviðinu á Íslandi og fé fáist hér á landi fyrir mótframlagi, jafnmiklu og því sem fæst úr sjóðum ESB.

Nefndinni þykir sýnt að styrkja þurfi fyrirtæki til starfsemi á sviði tungutækni

Eins og kemur fram annars staðar í þessari skýrslu eru fyrirtæki á þessu sviði hér á landi mjög vanburða og hafa fram til þessa ekki sett mikið fé í að búa til tól tungutækninnar. Dæmi eru um að fyrirtæki hafi farið flatt á slíkri starfsemi og eru útgefendur orðabóka vel þekkt dæmi um það. Nefndinni þykir sýnt að styrkja þurfi fyrirtæki til starfsemi á sviði tungutækni, bæði með því að veita þeim aðgengi að söfnum og gögnum sem þau þurfa í sínar vörur á hagstæðum kjörum, og eins með beinum styrkjum til þróunar vöru.

Þess er ekki að vænta að fyrirtæki leggi mikið í almennar rannsóknir á sviði tungutækni á næstu árum. Ef vel tekst til er það þó von nefndarinnar að þessi fyrirtæki styrkist og eflist og staðan geti batnað til muna á næsta áratug og þau geti þá staðið undir almennri rannsókna- og þróunarstarfsemi í landinu.

Átaksverkefni

Markmiðið með átaki á sviði íslenskrar tungutækni ætti að vera að styrkja sameiginlegan grunn tungutækninnar og söfnun hráefnis fyrir tungutæknitólin og að hvetja fyrirtæki til að þróa tólin, meðal annars með því að nýta hráefnissafnið. Á þennan hátt gæti skapast nýr iðnaður í tungutækni og sá sem þegar er fyrir hendi mundi styrkjast. Með iðnaði er átt við ýmsan iðnað tengdan útgáfu og meðferð tungumálsins, svo sem útgáfu á orðabókum og orðasöfnum, hugbúnað til leiðréttinga á stafsetningu og málfari, ýmis hjálparforrit við textasmíð, talgervla og hljóðtól. Sjá viðauka 1, 3, 4 og 5.

Vænta má, og ýta ætti undir, að slíkur iðnaður á Íslandi mundi nýta þekkingu sína og færni til þess að sækja inn á erlenda markaði

Vænta má, og ýta ætti undir, að slíkur iðnaður á Íslandi mundi nýta þekkingu sína og færni til þess að sækja inn á erlenda markaði, en þar munu vafalaust bjóðast ýmis tækifæri á næstu árum og áratugum. Í Finnlandi eru þegar tvö allsterk fyrirtæki í tungutækni sem nýta sér sérstöðu finnsku á þennan hátt. Eitt af fremstu fyrirtækjum tungutækni er belgískt (sjá tilvísanir í vefsíður í viðauka 7). Þannig gæti þjóðin nýtt sér sínar gömlu hefðir og menningu og virðingu fyrir tungumálinu til sóknar á nýja markaði á nýju tæknisviði. Þótt þarna séu vafalaust mörg tækifæri til sóknar má ekki gleyma því að það eru stór skörð í sóknarfylkinguna og þau þarf að fylla í.

Nefndinni virðist sem fjórþætt átak þurfi til þess að fylla í þessi skörð:

1. Stuðla ætti að því að í landinu byggist upp sameiginleg gagnasöfn sem geti nýst fyrirtækjum sem hráefni í afurðir.
2. Styrkja ætti fyrirtæki og rannsóknaraðila til hagnýtra rannsókna á sviði tungutækni.
3. Styrkja ætti fyrirtæki til þess að þróa afurðir tungutækni.
4. Auka þarf menntun á þessu sviði.

Í slíkt átak þarf fjármuni.

Nefndin leggur til að átakið felist í því að koma upp þróunarmiðstöð í tungutækni

Nefndin leggur til að átakið felist í fyrsta lagi í því að koma upp þróunarmiðstöð í tungutækni sem verði falið að vinna með útgefendum og öðrum við að koma upp þeim grunnsöfnum tungumálsins sem þarf. Þróunarmiðstöðin fái fjárveitingu á fjárlögum til nokkurra ára, til dæmis fimm til tíu ára.

Í öðru lagi leggur nefndin til að fé verði lagt í rannsóknasjóð

Í öðru lagi leggur nefndin til að fé verði lagt í rannsóknasjóð sem styrki rannsóknir og þróun á sviði tungutækni. Sjóðurinn verði tvískiptur eins og Rannsóknasjóður Rannsóknarráðs er nú, og veiti annars vegar fé til hagnýtra grunnrannsókna, sem gagnist iðnaðinum til lengri tíma litið, og hins vegar til þróunarverkefna fyrirtækja, einkum til þess að smíða tungutól.

Þá leggur nefndin til að menntun á þessu sviði verði efld

Þá leggur nefndin til að menntun á þessu sviði verði efld. Bæði vantar góða almenna menntun sem gagnast í vinnu við tungutækni og dýpri fræðilega menntun sem er undirstaðan. Auk þess að vera nauðsynlegt tungutækninni mundi hvorutveggja styrkja almenna stöðu íslensks máls. Lagt er til að sett verði upp stutt hagnýtt nám í máltækni og meistaranám í tungutækni.

Erfitt er að meta hve mikið fé þurfi í átakið, en ef einhver árangur á að nást þarf það að vera allmikið og starfsemin þarf að fara fljótt í gang. Stefna ber að því að verkefnið sé tímabundið og starfsemin verði sjálfbær á fimm til tíu árum. Í viðauka 1 er gerð grein fyrir þeim verkefnum sem hópurinn telur brýnust.

Þróunarmiðstöð

Nefndin leggur til að þróunarmiðstöð í tungutækni verði a.m.k. fyrst um sinn valinn staður í nánd við Orðabók Háskólans og Íslenska málstöð sem nú eru undir einu þaki á Neshaga 16. Um þetta hefur nefndin hugsað vandlega og vegið rök með og á móti. Það sem ræður úrslitum um þessa staðsetningu er að á engum öðrum stað hér á landi er til þekking og málsöfn sem mundu nýtast tungutækninni á fyrstu árum hennar. Sjá nánar í viðauka 6.

Nauðsynlegt er að tungutæknin hafi önnur viðhorf og annað vinnulag en hefðbundnar málvísindastofnanir

Á móti staðsetningunni er að nauðsynlegt er að tungutæknin hafi önnur viðhorf og annað vinnulag en hefðbundnar málvísindastofnanir. Hún þarf á mun meiri tækni að halda, meira af tæknilegu starfsfólki og það sem mest ríður á, hún þarf að hafa annað viðhorf til vinnunnar en rannsóknastofnanir sem starfa á akademískum forsendum. Þetta þarf að vera miðstöð með iðnaðarsjónarmið og iðnaðarhugsunarhátt, miðstöð sem framkvæmir, miðstöð sem vinnur verk sem gagnast iðnaðinum og þjóðfélaginu.

Miðstöðin þarf líka að reka aðra starfsmannastefnu en rekin er á akademískum stofnunum í hugvísindum. Launakjör þess fólks sem slík þróunarmiðstöð þarf á að halda eru önnur og svo framvegis. Allt eru þetta menningarleg atriði sem erfitt er að skilgreina, en nefndin vill ítreka að hún álítur að hvernig að þeim verði staðið muni ráða úrslitum um framgang málsins.

Við þróunarmiðstöðina þyrftu að starfa 5-10 starfsmenn og sé reiknað með 5 MKR á starfsmann á ári, mundi verkefnið kosta 25-50 MKR á ári. Í rekstrarkostnaði er gert ráð fyrir launum og aðstöðu fyrir starfsmenn. Reikna má með töluverðum stofnkostnaði í tækjum en hann fer eftir því hvar miðstöðin verður staðsett og því erfitt að áætla hann hér og er hann því ekki með í þessum tölum.

Rannsókna- og þróunarsjóður tungutækni

Hvort heldur er væri hægt að stofna nýjan sjóð eða styrkja sjóði Rannsóknarráðs Íslands með eyrnarmerktu fé til þessa iðnaðar. Síðari kosturinn yrði væntanlega ódýrari. Rannsóknasjóðurinn þyrfti að geta styrkt 5 hagnýt rannsóknaverkefni um a.m.k. 10 MKR hvert á ári og 10 þróunarverkefni fyrirtækja um a.m.k. 10 MKR hvert á ári. Samtals þyrfti því sjóðurinn um 150 MKR á ári til styrkveitinga. Í byrjun er vart við því að búast að umsóknir yrðu þetta margar, en þangað til ætti sjóðurinn að styrkja námsfólk til meistara- eða doktorsnáms á sviðum tungutækni og fjölga þannig hæfu starfsfólki.

Það sem hér er lagt til mundi ekki nægja til stærri verkefna eins og t.d. talgreiningar og stærri evrópskra samstarfsverkefna

Það sem hér er lagt til mundi ekki nægja til stærri verkefna eins og t.d. talgreiningar og stærri evrópskra samstarfsverkefna. Lagt er til að á slíkum verkefnum verði tekið sérstaklega hverju sinni. Þar gætu komið upp vandamál eins og til dæmis hvernig íslenskur hlutur í stærri iðnaðarverkefnum yrði fjármagnaður, þar sem íslenskir þátttakendur þyrftu opinberan stuðning í verkefni sem annars væri iðnaðarverkefni og erlendu þátttakendurnir kostuðu sinn hlut sjálfir.

Í slíkum verkefnum gætu komið upp vandamál í sambandi við eignarrétt á verkefnunum þar sem íslenski hlutinn væri að mestu kostaður af opinberu fé. Einnig gætu komið upp vandamál í sambandi við jafna samkeppnisstöðu fyrirtækja hvað varðar opinbert fé, vandamál sem erfitt er að gefa algildar reglur eða leiðbeiningar um.

Styrkir til slíkra verkefna yrðu að koma til viðbótar við fyrrnefndar 150 MKR þar sem þær þyrfti til smærri verkefna og mætti því ekki skerða

Gera þyrfti ráð fyrir að minnsta kosti einu stóru slíku verkefni þar sem styrkur væri 30 MKR á ári í þrjú til fimm ár. Styrkir til slíkra verkefna yrðu að koma til viðbótar við fyrrnefndar 150 MKR þar sem þær þyrfti til smærri verkefna og mætti því ekki skerða.

Mannafli og menntun

Inngangur

Eitt stærsta vandamál sem Íslendingar standa frammi fyrir ef þeir ætla að hefja öflugt og markvisst starf á sviði tungutækni er skortur á fólki með menntun, reynslu og þekkingu á því sviði. Mikilvægt er að hafa í huga að enda þótt skammt sé síðan tungutækni varð að iðnaði í grannlöndum okkar styðst það starf við margra ára rannsóknir og kennslu á háskólastigi. Í enskumælandi löndum er máltölvun (Computational Linguistics) víða sérstök grein í háskólum, en einnig sums staðar innan málvísindadeilda eða tölvunarfræðideilda. Svipuðu máli gegnir um Komputerlinguistik í þýskumælandi löndum, datalingvistik á Norðurlöndum o.s.frv.

Langflest störf sem málfræðingum standa til boða við háskóla í ensku- og þýskumælandi löndum um þessar mundir eru t.d. á þessu sviði.

Mikill vöxtur hefur verið í þessum greinum á undanförnum árum. Langflest störf sem málfræðingum standa til boða við háskóla í ensku- og þýskumælandi löndum um þessar mundir eru t.d. á þessu sviði. Vitaskuld er beint samband milli þenslunnar á þessi sviði í háskólum og hinnar öru þróunar sem hefur verið í tungutækni sem iðngrein sem veltir háum fjárhæðum. Þörf atvinnulífsins fyrir fólk með menntun á þessu sviði hefur stóraukist, og þar með vilji yfirvalda til að efla slíka kennslu á háskólastigi.

Hér á landi hefur lítið sem ekkert verið fengist við kennslu og rannsóknir á þessu sviði. Innan íslensku og almennra málvísinda við Háskóla Íslands hafa einstöku sinnum verið haldin námskeið um tölvur og tungumál, en því fer fjarri að þar hafi verið um skipulagðar námsleiðir að ræða. Ekki er heldur vitað til þess að íslenskir málfræðingar sem hafa farið til framhaldsnáms erlendis hafi lagt stund á nám af þessu tagi. Nokkuð er til af rannsóknum á íslenskri orðtíðni sem unnar eru með aðstoð tölvu, og unnið hefur verið að aðlögun og frumsmíð talgervla fyrir íslensku, en enginn íslenskur málfræðingur fæst nú svo að heitið geti við rannsóknir á sviði máltölvunar.

Stutt námsbraut í máltækni

Þar er átt við fólk sem kalla mætti máltækna

Auk fólks með verulega menntun á sviði máltölvunar er ljóst að einnig verður mikil þörf á fólki með annars konar menntun til að vinna í íslenskri tungutækni. Þar er átt við fólk sem kalla mætti máltækna, sem hefði hlotið staðgóða hagnýta menntun í íslenskri málfræði og málnotkun, hvers kyns meðferð og frágangi texta, þýðingum o.fl. Þetta nám myndi vitaskuld einnig nýtast á ýmsum öðrum sviðum, svo sem í fjölmiðlun, við útgáfustörf og fleira. Nefndin leggur til að þessu námi verði komið á fót hið fyrsta.

Þegar hafa verið samdar tillögur um tveggja ára nám af þessu tagi; sjá *Skýrslu nefndar um stuttar, hagnýtar námsbrautir við Háskóla Íslands* frá 1998. Gert er ráð fyrir að kennsla hefjist á slíkri námsbraut í íslensku næsta haust og gæti það orðið vísir að því námi sem hér er lagt til.

Meistaranám í máltölvun

Óráðlegt er að ætla að Íslendingar geti byggt upp öflugt starf á sviði tungutækni án þess að hyggja að fræðilegum undirstöðum slíks starfs. Nauðsynlegt er að fá sem fyrst til starfa vel menntað fólk á sviði íslensks máls og tölvunarfræði sem gerir sér grein fyrir sérkennum íslenskrar málfræði og íslensks málsamfélags. Ef ekki verður byggð upp innlend þekking á þessu sviði innan menntastofnana verðum við um ófyrirsjáanlega framtíð þiggjendur á þessu sviði og höfum miklu minni möguleika á að bregðast við breyttum aðstæðum og nýjungum, og þróa þau tól og tæki sem henta best íslenskum aðstæðum.

Nefndin leggur því til að komið verði á fót kennslu í máltölvun við Háskóla Íslands

Nefndin leggur því til að komið verði á fót kennslu í máltölvun við Háskóla Íslands. Formið á slíkri kennslu gæti verið með ýmsu móti, en einn möguleikinn er að skipuleggja þverfaglegt meistaranám í samvinnu fjögurra deilda; heimspekideildar (einkum íslensku og almennra málvísinda en einnig heimspeki og erlendra tungumála), félagsvísindadeildar (einkum sálfræði), raunvísindadeildar (einkum tölvunarfræði) og verkfræðideildar (rafmagns- og tölvuverkfræði).

Fordæmi fyrir slíku þverfaglegu námi er þegar fyrir hendi í sjávarútvegsfræðum og umhverfisfræðum. Með þessu móti ætti að vera unnt að stefna saman kennurum og nemendum af ólíkum sviðum og skapa, ef vel tekst til, frjóan rannsóknarvettvang og fjölbreytt nám með miklum sveigjanleika.

Námsbrautirnar þyrftu fjóra kennara og sé þar einnig reiknað með 5 MKR á starfsmann, yrði rekstrarkostnaður um 20 MKR á ári.

Heildarkostnaður

Heildarkostnaður á ári, við átakið sem lagt er til, yrði því:

Þróunarmiðstöð	25 til 50 MKR
Rannsókna- og þróunarsjóður	150 MKR
Sérstakur styrkur til stærri alþjóðlegra verkefna	30 MKR
Stutt hagnýtt nám í máltækni	10 MKR
Meistaranám í máltölvun	10 MKR
Alls	**225 til 250 MKR á ári**

Mat nefndarinnar er að áætlunin sé mjög hófleg og raunhæf

Þetta kann að þykja allmikið fé og er það vissulega, en mat nefndarinnar er að áætlunin sé mjög hófleg og raunhæf og sé mikið úr henni dregið muni hún ekki ná tilætluðum árangri. Það hefur sem sagt ekki verið gert ráð fyrir órökstuddum niðurskurði í þessari áætlun.

Viðaukar

1. Verkefni í íslenskri tungutækni

Forgangsverkefni

Meginmarkmið Íslendinga hlýtur að vera að unnt verði að nota íslenska tungu, ritaða með réttum táknum, sem víðast innan tölvu- og fjarskiptatæknninnar. 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. Nefndin leggur til að á næstu fimm árum verði lögð áhersla á eftirtalin verkefni:

1. Helstu tölvuforrit á almennum markaði verði á íslensku (Windows, Word, Excel, Netscape, Internet Explorer, Eudora, ...)
2. Unnt verði að nota íslenska bókstafi (áéíóúýðþæöÁÉÍÓÚÝÐÞÆÖ) við allar aðstæður; í tölvum, GSM-símum, textavarpi og öðrum tækjum sem almenningur notar.
3. Unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði. Til þess að það sé hægt þarf að:
   3.1. Koma upp stórri tölvutækri textaheild með íslenskum textum af sem fjölbreyttustum toga til að byggja áframhaldandi vinnu á.
   3.2. Koma upp fullgreindu orðasafni (með málfræðilegri og merkingarlegri greiningu) til nota í áframhaldandi vinnu.
4. Til verði góð hjálparforrit við ritun texta á íslensku, s.s. orðskiptiforrit, stafsetningarleiðréttingarforrit, málfarsleiðréttingarforrit o.fl.
5. Til verði góður íslenskur talgervill sem geti lesið upp íslenskan texta með skýrum og auðskiljanlegum framburði og eðlilegu tónfalli og sem sé skiljanlegur án þjálfunar.
6. Unnið verði að þróun talgreiningar fyrir íslensku, með það að markmiði að til verði forrit sem geti túlkað eðlilegt íslenskt tal.
7. Unnið verði að þróun forrita til vélrænna þýðinga milli íslensku og annarra tungumála, m.a. til að auðvelda leit í gagnabönkum.
8. Ákveðnum aðilum (stofnunum eða fyrirtækjum) verði falin ábyrgð á einstökum verkefnum.

Nánari skýringar

1. Á þessu sviði hefur orðið afturför á undanförnum árum; fyrir tíu til fimmtán árum voru helstu ritvinnsluforrit á íslensku. Nauðsynlegt er að ýta á eftir því að Windows-stýrikerfið og önnur helstu forrit frá Microsoft verði íslenskuð, en einnig þarf að þýða ýmis önnur forrit sem almenningur notar hversdagslega, s.s. vefskoðara, póstforrit o.fl. Með samningi við Microsoft er vonandi vörn snúið í sókn.
2. Á þessu sviði hefur orðið mikil framför á undanförnum árum, og þar hafa íslenskir staðlamenn og málverndarmenn unnið gott starf. Þar eru þó ýmsar blikur á lofti. Þannig eru íslenskir stafir t.d. ekki í stafatöflu GSM-síma, sem er alvarlegt, ekki síst í ljósi þess að tengsl tölvutækni og fjarskiptatækni eru sífellt að aukast.
3. Með málgreiningu (parsing) er átt við vélræna greiningu texta í orðflokka, setningarliði og setningar. Slík greining er mikilvæg fyrir gerð málfræðileiðréttingarforrita, þýðingarforrita o.fl. Mörg tungumál eiga forrit til nokkuð fullkominnar málgreiningar, en lítið er um slíkt hér á

landi. Vísi að vélrænni orðflokkagreiningu má þó finna í Púka Friðriks Skúlasonar og forritum sem Stefán Briem skrifaði á sínum tíma fyrir Orðabók Háskólans, en engin vélræn setningagreining er til fyrir íslensku.

3.1. Til að unnt sé að útbúa forrit sem vinna með tungumál þurfa að liggja fyrir miklar og nákvæmar upplýsingar um málið og notkun þess. Einn meginþátturinn í öflun slíkra upplýsinga felst í því að koma upp sem stærstri textaheild (corpus) sem hafi að geyma tölvutæka íslenska texta. Þar þurfa að vera textar af ýmsu tagi; blaðatextar, fræðitextar af ýmsum sviðum, bókmenntatextar, talmál o.fl. Einnig þarf að gæta þess að textaheildin geymi mál bæði karla og kvenna, fólks af ýmsum aldri, úr ýmsum þjóðfélagshópum, sem víðast af landinu, o.s.frv. Úr þessum textum þarf síðan að vinna margs konar upplýsingar sem nauðsynlegar eru til að hægt sé að skrifa forrit til hvers kyns vinnu með málið. Engin slík textaheild er nú til, þótt mikið hráefni í hana sé fyrir hendi, ekki síst hjá Orðabók Háskólans. Gerð textaheildar af þessu tagi, og úrvinnsla úr henni, er forsenda markvissrar vinnu í íslenskri tungutækni.

3.2. Hér gildir í meginatriðum hið sama og segir um næsta lið á undan. Orðasafn með grunnorðaforða íslenskunnar (nokkrum tugum þúsunda orða) er forsenda ýmiss konar vinnu í tungutækni. Í þessu orðasafni þurfa að vera sem nákvæmastar upplýsingar um hvert orð; framburð þess, orðflokk, beygingu, setningarstöðu, merkingu, stílgildi o.s.frv. Slíkar upplýsingar koma að gagni við gerð málfræðileiðréttingarforrita, vélrænar þýðingar, leit í gagnabönkum o.fl. Grunnur að slíku orðasafni er til hjá Orðabók Háskólans. Þar vantar þó inn miklar upplýsingar, t.d. alla merkingargreiningu.

4. Til eru allgóð íslensk orðskiptiforrit, og einnig forrit til stafsetningarleiðréttingar (Púki Friðriks Skúlasonar o.fl.). Sá galli er þó á þessum forritum að þau eru ekki innbyggð í helstu ritvinnslukerfi á markaðnum (s.s. Word) og vinna ekki nægilega vel með þeim. Það dregur mjög úr notagildi og notkun þeirra. Nauðsynlegt er að koma góðum íslenskum forritum inn í Word. Forrit til málfarsleiðréttinga (grammar checker, style checker) aðstoða notendur við að útrýma beygingarvillum, rangri orðaröð og klúðurslegri setningaskipan. Slík forrit eru aftur á móti engin til fyrir íslensku, en væru mjög þörf.

5. Undanfarin 6-7 ár hefur verið á markaðnum íslenskaður talgervill frá sænska fyrirtækinu Infovox, en hann var gerður í samvinnu Málvísindastofnunar Háskólans, verkfræðideildar Háskólans og Öryrkjabandalags Íslands. Þessi talgervill er byggður á tækni sem nú þykir úrelt, og skiptar skoðanir eru um gæði hans. Ljóst er að framburði hans er um margt ábótavant, en þó hefur hann gagnast sumum mjög vel. Annar talgervill hefur einnig verið gerður, byggður á annarri tækni, en sá hefur ekki verið settur á almennan markað. Nauðsynlegt er að vinna áfram að því að útbúa fullkominn íslenskan talgervil.

6. Með talgreiningu (speech recognition) er átt við það að tölvur skilji talað mál. Mjög miklar framfarir hafa orðið á þessu sviði upp á síðkastið. Líklegt er að talgreining muni skipta miklu máli á ýmsum sviðum í framtíðinni, t.d. við upplýsingaleit og stjórn ýmiss konar tækja. Því er mjög mikilvægt að hefja skipulega vinnu að þróun talgreiningar fyrir íslensku, en á því sviði hefur ekkert verið gert.
7. Vélrænar þýðingar eiga sér langa sögu, en hafa gengið misjafnlega. Á seinustu árum hafa þó komið fram þýðingarforrit sem virka allvel, a.m.k. á afmörkuðum sviðum. Líklegt er að mikilvægi vélrænna þýðinga muni aukast verulega á næstu árum, t.d. í sambandi við leitir í gagnabönkum o.fl. Stefán Briem hefur unnið talsvert að tilraunum með vélrænar þýðingar milli íslensku og esperanto, en þær eru mjög takmarkaðar.
8. Færa má rök að því að það hafi staðið allri þróun á þessu sviði mjög fyrir þrifum að enginn aðili hefur borið ótvíræða ábyrgð á því að Íslendingar fylgdust þar með. Meðal þeirra sem eðlilegt er að standi að þeim verkefnum sem hér er lýst má einkum nefna Málvísindastofnun Háskólans, Íslenska málstöð, Orðabók Háskólans, Staðlaráð Íslands og Póst- og fjarskiptastofnunina, en einnig er eðlilegt og nauðsynlegt að einkafyrirtæki taki þátt í verkefnunum. Ef til vill væri æskilegt að koma á fót einhvers konar formlegum samstarfsvettvangi þessara aðila. Menntamálaráðuneytið þarf síðan að fela einstökum aðilum eða samtökum þeirra ábyrgð á ákveðnum verksviðum.

2. Staða íslenskra bókstafa

Staðlar, stafatöflur og leturgerðir

Þó að tölvur hafi í upphafi verið smíðaðar til útreikninga hefur notkun þeirra í ýmiss konar meðferð texta orðið æ mikilvægari. Til að gera þetta kleift hafa framleiðendur tölva smíðað margvísleg kerfi sem byggjast á tungumálakunnáttu. Staðlar hafa orðið til um þetta og eru sumir þeirra opinberir staðlar, evrópskir eða alþjóðlegir, en aðrir eru í einkaeign einstakra framleiðenda.

Staðall er í eðli sínu samkomulag hagsmunaaðila um hvernig eitthvað skuli gert og þarf því ekki að semja staðal nema hagsmunaaðilarnir séu fleiri en einn. Þegar t.d. er skipst á gögnum um Netið milli ólíkra vélartegunda er hagkvæmt að fylgja staðli en innan stýrikerfis einnar vélar er minni ástæða til að hafa gögn á öðru formi en því sem hverjum framleiðanda hentar best.

Gerðir staðla sem snerta tungutækni

1. Stafatöflur. Í stafatöflum er hverjum staf gefið nafn og honum er úthlutað sæti eða númeri í töflunni. T.d. er stafnum SMALL LATIN LETTER THORN gefið númerið 239 (EF) í stafatöflunni Latin-1 (ISO/IEC JTC1 8859-1). Stafatöflur skilgreina ekki útlit stafa, heldur ræðst útlitið af leturgerðum.
2. Leturgerðir. Í hverri leturgerð er útlit stafs skilgreint. T.d. er stafurinn g birtur sem g í Helvetica. Flestar leturgerðir eru í einkaeigu en ISO staðall er til um leturgerðir fyrir tölvulestur (OCR-B).
3. Hnappaborð. Staðsetning stafa á hnappaborðum fer eftir þörfum og venjum hverrar þjóðar. Um hnappaborð eru til landsstaðlar, t.d. ÍST 125 um íslenskt hnappaborð.
4. Stafrófsröð, ritun upphæða, ritun dagsetninga, og margt fleira. Framleiðendur safna upplýsingum um þarfir hvers málsvæðis ásamt upplýsingum um hnappaborð og mynda það sem kallað er „locale". Helstu staðlar um þetta eru í eigu einstakra framleiðenda. Íslendingar hafa gert íslenskan forstaðal FS130 um þessi atriði.

Íslensk þátttaka í staðlavinnu

Íslendingar eru aðilar að alþjóðlegu staðlasamtökunum ISO og evrópsku staðlasamtökunum CEN. Staðlaráð Íslands sér um að innleiða íslenska staðla en aðild að Evrópsku staðlasamtökunum hefur þýtt að skyldugt er að gera evrópska staðla CEN að íslenskum stöðlum. Íslendingar hafa tekið að sér rekstur einnar tækninefndar, CEN/TC304, sem hefur það verkefni að útbúa evrópska staðla tengda „alþjóðavæðingu". Starfsemi þeirrar nefndar hefur að miklu leyti verið fjármögnuð með styrkjum frá ESB en þessu starfi hafa Íslendingar getað fylgst með og haft áhrif á gerð staðla sem snerta hagsmuni þeirra.

Á vegum TC304 hefur verið gerður skráningarstaðall um þjóðlegar þarfir í upplýsingatækni. Nú er þar m.a. unnið að gerð staðla um leit í gagnagrunnum. Búast má við að á næstu árum verði fleiri svið tungutækni að viðfangsefni staðlastarfs. Má þar nefna aðferðir við leit að texta á netinu og við talgreiningu.

Íslendingar hafa gert staðal um íslenskt lyklaborð, ÍST 125, sem að mestu hefur verið fylgt af framleiðendum og þeir hafa einnig virt staðalinn FS130 um ritun upphæða, stafrófsröð og fleira. FS130 þarfnast endurskoðunar vegna þess að í honum eru m.a. taldir upp fornstafir til nota við útgáfu fornra texta sem ekki eru til í neinum stafatöflum og enginn áhugi er á að koma inn í stóru stafatöfluna.

Stafatöflur og letur

Flestir hafa orðið varir við að því fylgja oft ýmis vandamál að flytja skjöl á milli tölva, einkum þegar þær nota mismunandi stýrikerfi. Oft á þetta rætur að rekja til þess að stýrikerfin nota ekki sömu stafatöflur. Þó að mikið hafi verið unnið að því að staðla stafatöflur er enn langt í land með að stafatöflur hætti að vera til trafala, einkum þegar unnið er með tungumál eins og íslensku, þar sem notuð eru tákn sem þekkjast í fáum eða engum öðrum málum.

Líklega er best að gera grein fyrir ástæðum þessara vandamála og spá um framtíð þeirra með því að rekja stuttlega sögu hinna ýmsu stafataflna og skýra frá því í hvaða samhengi þær voru eða eru notaðar.

Hvað er stafatafla?

Hreinlegasta skilgreining stafatöflu er að hún innihaldi færslur sem hver um sig samanstendur af tölu (sæti) og heiti táknsins. Stafatöflur skilgreina hins vegar ekki útlit táknanna og hafa leturgerðarmenn svigrúm til að kveða á um hvernig t.d. stafurinn “g” lítur út í hinum ýmsu leturtegundum. Nafnið segir þó oft til um gerð stafsins, t.d. heitir danskt “Ø” nafninu “LATIN CAPITAL LETTER O WITH STROKE” og stafurinn “Á” heitir “LATIN CAPITAL LETTER A WITH ACUTE”. Stafirnir “Đ” og “Þ” heita íslenskum nöfnum sínum “LATIN CAPITAL LETTER ETH” og “LATIN CAPITAL LETTER THORN”. Stafurinn sem Króatar kalla “Djet” hefur fengið nafnið “LATIN LETTER D WITH STROKE”.

Í stöðlun stafataflna fyrir tölvur hefur verið unnið að samræmingu nafna en ekki að samræmingu útlits stafa. Stafurinn “GREEK CAPITAL LETTER UPSILON” í 8 bita gríska stafatöflu heitir þannig sama nafni í 16 bita stafatöflunni 10646 og verður þannig ekki ruglað saman við stafinn “LATIN CAPITAL LETTER Y” sem Íslendingar kalla “Upsilon”.

7 bita töflur

Langt er síðan flestar tölvur fóru að vinna með 8 bita bæti og ástæðulaust er að ræða það sem þar fór á undan, enda eru nú líklega engar menjar um það sem enn valda vandræðum. Smæsta eining sem tölvur vinna með er biti, sem getur einungis haft tölugildið 0 eða 1. Ef átta bitar eru settir saman í eitt bæti gefur það 2^8=256 möguleika.

ASCII

Þegar fyrst var farið að senda gögn á milli tölva þótti ráðlegt að taka einn bita frá, sem tölvurnar gátu notað til að fylgjast með sambandinu sín á milli. Þá voru eftir 7 bitar, sem gáfu 2^7=128 möguleika. Þegar búið var að taka frá 33 stafasæti fyrir ýmis stýritákn var eftir pláss fyrir hástafi og lágstafi enska 26 stafa stafrófsins, tölustafina 10 og helstu greinarmerki. Þannig varð til stafataflan ASCII (American Standard Code for Information Interchange).

Þessi tafla náði geysimikilli útbreiðslu í krafti yfirburða bandarískrar tölvutækni og má í raun segja að smæð hennar og fátækt að táknum fyrir önnur tungumál sé rótin að flestum þeim vandamálum sem til hafa orðið síðan og lúta að stafatöflum. Einnig má segja að orsök langlífis hennar sé að einhverju leyti sú að ekki aðeins var bandarískur tölvuiðnaður leiðandi á heimsmarkaði, heldur var stærstur hluti sölumarkaðarins einnig í hinum enskumælandi heimi, og því lítill þrýstingur á iðnaðinn að gera rúm fyrir fleiri tákn.

ASCII taflan ber það með sér að vera gerð fyrir Bandaríkjamenn, t.d. er í henni dollaramerkið en ekki t.d. pundmerkið sem Englendingar nota. Í flestum Evrópulöndum voru búnar til sérstakar útgáfur ASCII töflunnar þar sem tekin voru út tákn eins og t.d. „dauður hattur“ og þjóðleg tákn sett í staðinn. Þannig var hægt að bæta við nokkrum stöfum og dugði það mörgum þjóðum, en ekki Íslendingum þar sem sérstafir Íslendinga voru of margir.

Enn notast margir póstþjónar víða um veröld við ASCII-töfluna þegar tölvupóstur er sendur á milli staða. Nú orðið ráða þó flest forrit við að þýða önnur tákn, sem send eru gegnum póstþjónana með stafarunum á borð við „=F1“. Þannig er hægt að vísa í öll táknin í 8 bita töflum (sjá hér að aftan). Þau forrit sem ekki geta þetta (t.d. Unix-forritið „elm“) eru nú mjög á undanhaldi og veldur þetta því litlum vandkvæðum nú orðið.

GSM

Helsta svið upplýsingatækni þar sem 7 bita töflur eru enn notaðar er GSM-tæknin. Framleiðendur GSM nota þessa töflu vegna þess að styttri tíma tekur að flytja færri bita. Stafatöflurnar í símunum eru notaðar til að birta leiðbeiningar og skipanir á skjá, skrifa færslur í símaskrá, birta SMS-skilaboð o.s.frv. Nú er unnið að gerð síma með meiri virkni sem tengist samskiptum við tölvur, og eykst mikilvægi stafataflna í símum við það. Til dæmis framleiðir Nokia nú þegar síma sem einnig eru handtölvur sem nota þessa 7 bita töflu.

Til að útskýra hvernig stendur á mismun stafataflna í tölvum og í GSM símum verður að hafa í huga að engum datt í hug fyrir áratug að símar myndu breytast í tölvur og hugtök eins og netsími voru þá óþekkt. Stöðlun á sviði fjarskiptatækni hefur því farið fram á öðrum vettvangi en stöðlun á sviði tölvumála, bæði í Evrópu og á alþjóðavísu. Evrópska staðlastofnunin fyrir tölvur heitir CEN en alþjóðastofnunin ISO/IEC JTC1. Í stöðlun fyrir tölvur eru bandarísku fyrirtækin IBM og Microsoft ráðandi og ISO/IEC JTC1 gefur út þá staðla sem skipta máli.

Fjarskiptastaðlar eru á hinn bóginn gerðir af evrópsku stofnuninni ETSI en á alþjóðavettvangi af ITU. Í fjarskiptastöðlun eru evrópsk fyrirtæki leiðandi eins og Ericsson, Nokia og Siemens og ETSI hefur gefið út helstu staðla um GSM. 7 bita tafla ETSI fyrir GSM var eins konar evrópsk ASCII tafla þar sem mörg tákn ASCII töflunnar voru fjarlægð en evrópskir stafir settir í staðinn, að auki voru nafnareglur tölvustaðla ekki virtar. Þannig eru sumir gríska stafir í töflunni en ekki allir og ætlast er til að Grikkir noti t.d. “LATIN CAPITAL LETTER Y” fyrir “GREEK CAPITAL LETTER UPSILON” en þeir stafir hafa sama útlit en ekki lágstafirnir sem svara til þessara hástafa. Þetta veldur vandræðum þegar gögn eru flutt úr GSM símum í tölvur.

Í grunntöflu GSM (default töflu) er gert ráð fyrir nokkrum helstu Evrópumálum auk ensku. Staðalinn gerir ráð fyrir að notuð sé önnur tafla ef grunntaflan dugir ekki. Slíkar töflur eru til fyrir ýmis stafróf, t.d. kýrillískt letur. Hægt er að skrá slíkar töflur í GSM staðlana en ennþá hefur það ekki verið gert fyrir Ísland enda hafa Íslendingar ekki útbúið handa sér neina sértöflu. Íslendingar láta sér duga að sleppa broddum og rita „th“ og „d“ í stað „þ“ og „ð“. Ef send eru SMS-skilaboð úr tölvupósti með íslenskum stöfum birtast hin ýmsu erlendu sértákn í þeirra stað.

Lítið hefur verið unnið í því að bæta úr þessu ástandi. Póstur og sími og síðar Póst- og fjarskiptastofnun hefur farið með aðild Íslendinga að ETSI. Landssíminn hefur ekki haft frumkvæði að því að útbúa sértöflu með íslenskum stöfum. Í ljósi þess hve notkun GSM-síma er útbreidd hér á landi er þetta ástand engan veginn viðunandi.

Staða mála í dag, eftir að Póstur og sími var gerður að hlutafélaginu Landssíminn, er að hann hefur sagt upp aðild sinni sem íslensk staðlastofnun að ETSI og ætlast er til að Staðlaráð Íslands taki við því hlutverki. Staðlaráð hefur hins vegar neitað að taka við þessari skyldu nema henni fylgi fjárveitingar sem dugi til annars en að fara með aðild að nafninu til.

ETSI er vettvangur símaframleiðenda og þar eru á borðum tillögur að fjarskiptastöðlum framtíðarinnar þar sem stuðst verður við ISO/IEC JTC 1 10646 (sama og Unicode, sjá hér á eftir). Það er þó ljóst að fyrir hvert land eða landssvæði þarf að útbúa hlutmengi af Unicode og fylgjast þarf með því að í nýjum stöðlum verði gert ráð fyrir íslensku. Eins og er taka Íslendingar engan þátt í stöðlun á sviði fjarskipta.

8 bita töflur

Þegar ljóst varð að 7 bita töflur dygðu ekki til að fullnægja þörfum ýmissa tungumála var gripið til þess að nota áttunda bitann til viðbótar við hina sjö til að greina að mismunandi tákn (enda var hans ekki lengur þörf til að viðhalda samskiptum milli tölva).

Ef þetta hefði verið gert á einum stað á einum tíma, t.d. undir handleiðslu staðlastofnunar, hefði e.t.v. mátt komast hjá ýmsum þeirra vandamála sem hrjá okkur enn í dag. Raunin varð hins vegar sú að hinir ýmsu tölvuframleiðendur bjuggu til sínar eigin lausnir, sem skapaði mikinn glundroða í stafatöflum. Hér á eftir verður sagt frá helstu 8 bita töflunum sem enn eru í notkun, en ónefndar eru ótal töflur sem fram komu á 9. áratugnum, t.d. fyrir hinar ýmsu svokölluðu „heimilistölvur“, sem þá voru vinsælar en hafa að mestu vikið fyrir IBM PC tölvum með stýrikerfum frá Microsoft.

EBCDIC

EBCDIC-taflan varð til hjá IBM í kringum 1965 þegar settar voru á markað System 360 tölvur frá IBM. Taflan varð til svo styðja mætti letur á gataspjöldum sem fylgdu System 360 en er enn notuð í hugbúnaði frá IBM. Líkt og gert var við ASCII töfluna voru gerðar útgáfur af þessari töflu fyrir hin ýmsu markaðssvæði. EBCDIC-taflan er enn notuð í stór- og miðtölvustýrikerfum IBM, t.d. OS/400-kerfinu sem notað er á AS/400-miðtölvunum sem hafa notið töluverðra vinsælda hjá stærri fyrirtækjum hér á landi.

DOS-töflur

IBM þróaði sérstakar útgáfur af 8 bita töflum sínum til að sinna ákveðnum markaðssvæðum, t.d. varð til taflan 861 sem var gerð að mestu fyrir tilstuðlan Íslendinga þar sem tekin var IBM tafla fyrir portúgölsku og gerðar á henni nokkrar breytingar. Síðar voru þessar töflur gerðar þannig að þær samsvöruðu stöðluðum ISO töflum sem IBM átti sjálft mikinn hlut í að urðu til. Þá var gerð 850 taflan sem svarar til Latin 1 en Íslendingar hafa verið lengi að færa sig úr 861 töflunni í 850 töfluna og margs konar vandræði hljótast enn af þeim sökum. Enn eru til fyrirtæki og stofnanir á Íslandi sem nota 861 töfluna en sérfræðingar Microsoft hafa sagt að við fyrirhugaða þýðingu Windows á íslensku verði ekki hægt að styðja hana.

Macintosh-töflur

Þegar Macintosh-tölvur frá Apple komu fyrst á markaðinn árið 1984 var strax lögð mikil áhersla á að þýða stýrikerfi þeirra á ýmis tungumál og gera mönnum víða um heim kleift að nota þær til að skrifa á eigin móðurmáli. Eins og títt var á þeim tíma (sem minnst var á hér að framan) hönnuðu Apple-menn eigin stafatöflur til nota fyrir ýmis mál. Til varð ein stafatafla fyrir Vesturlönd, önnur fyrir Austur-Evrópu, grísk stafatafla, kýrillísk stafatafla (fyrir rússnesku, búlgörsku og fleiri mál) og þar fram eftir götunum.

Þó urðu nokkur tungumál milli vita, þar á meðal íslenska. Þegar Macintosh-stýrikerfið var þýtt á íslensku var gripið til þess að búa til afbrigði af töflunni sem notuð var fyrir Vesturlönd. Þetta var gert með því að taka út 6 lítt notuð tákn og setja í stað þeirra há- og lágstafi “Þ”, “Ð” og “Ý” (aðrir broddstafir voru þegar til í töflunni). Sama leið var farin fyrir fleiri mál, þannig eru einnig notuð afbrigði af Vesturlandatöflunni fyrir tyrknesku, rúmensku og írsku, og sérstakt afbrigði Austur-Evrópu-töflunnar er til fyrir króatísku.

Það sem nú veldur vandræðum við að flytja íslensk skjöl milli Macintosh-stýrikerfisins og Windows er fyrst og fremst sú staðreynd að við notum þetta sérstaka afbrigði í Macintosh-kerfinu. Þeir sem nota staðalútgáfu Vesturlandatöflunnar þurfa litlar áhyggjur að hafa af þessum málum, þar sem mörg algeng forrit þýða sjálfkrafa milli þessara taflna. Í slíkum þýðingum er hins vegar sjaldnast tekið tillit til íslensku (eða annarra mála sem nota afbrigði af staðaltöflum).

Þetta hefur sérstaklega valdið vandræðum eftir að Microsoft tók að selja Office pakka sinn fyrir Macintosh tölvur sem þýðir að nú eru í umferð á Íslandi skjöl fyrir MS-Word í ósamhæfðum útgáfum vegna íslenskra stafa. Þeir sem útbúa slík skjöl á Apple tölvu geta ekki búist við að íslenskir stafir

séu lesanlegir á PC-tölvu og öfugt. Þrátt fyrir ítrekaðar fyrirspurnir Apple á Íslandi hefur ekki fengist vilyrði hjá Microsoft að bæta úr þessu en Microsoft framleiðir bæði Office fyrir Windows og Macintosh.

Um framtíð Macintosh-taflnanna og hvaða úrlausna er að vænta í þessum málum má lesa í kaflanum um Unicode hér að aftan.

Unix-/Windows-töflur

Windows 95, svo og flest Unix-kerfi, nota nú stafatöflur sem byggja á ISO-staðlaröðinni 8859. Þannig er sú tafla sem Íslendingar sjá mest tafla sem gengur undir nafninu Latin-1 sem er styttra nafn á töflunni á ISO/IEC JTC1 8859-1. Taflan sem Windows notar er þó ekki eins og staðallinn því í henni eru notuð nokkur sæti sem skilin eru eftir auð í staðlinum og IBM og fleiri nota undir sérstök stýritákn. Í ISO 8859-1 er að finna öll rittákn sem notuð eru í venjulegum íslenskum texta, þökk sé ötulli baráttu íslenskra staðlamanna við að fá íslenska stafi inn í staðalinn.

Eins og mörgum er kunnugt þurftu þeir að etja kappi við Tyrki, sem sættu sig ekki að vera með sína stafi í töflu fyrir Suður-Evrópu (Latin-3). Þeir fengu gerða töflu sem var að öllu leyti eins og Latin-1 nema í stað “ð”, “Ð”, “ý”, “Ý”, “þ” og “Þ” komu tyrkneskir stafir. Reynt var að láta þessa töflu leysa Latin-1 af hólmi en Tyrkir urðu að sætta sig við að hún fengi númerið 8859-9 og Latin-5 (latínutafla númer 5). Eistar hafa einnig leikið þann leik að setja eistneska stafi í stað “ð” og “þ” í Latin-1 til síns heimabrúks en hafa ekki reynt að gera þá töflu að alþjóðastaðli.

Nýlega kom fram afbrigði af 8859-1, sem kallast 8859-15, eða Latin-9. Með henni fá Eistar og Finnar táknin s og z með hatti, sem þeir þurfa til að rita fjölmörg tökuorð úr rússnesku, og einnig bætast við tákn úr frönsku oe (œ) og Y með tvípunkti (Ÿ). Á móti er fórnað táknum sem sjaldan eru notuð eins og t.d. sérstakt tákn fyrir 3/4 og dauðir broddar. Loks bætist við evrutáknið, sem var ástæða þess að þessi staðall var gerður.

Eins og áður sagði hefur Windows notað staðlaðar 8 bita ISO töflur en hefur bætt við stöfum í sætum þar sem IBM hefur ekki pláss vegna gamalla kerfa sem nota þau sæti undir stýritákn. Microsoft bætti evrutákninu í töfluna með Latin-1, auk eistnesku og finnsku stafanna og þar með eru sömu bókstafir í þeirri töflu og í Latin-9, en Microsoft þurfti ekki að fórna eldri táknum.

Þessar viðbætur eru í Windows 98 en þær er einnig hægt að fá í eldri kerfi frá MS af vefsíðum MS. Um evrutáknið í stöðlum hefur tækninefnd CEN/TC304 gert sérstaka skýrslu sem er á vefsíðu Staðlaráðs Íslands, www.stri.is en Evrópusambandið hefur falið TC304 að fylgjast með nauðsynlegri stöðlun vegna evrunnar.

Í upphaflegum tillögum að nýju afbrigði af Latin-1 töflunni var lagt til að íslenskir stafir yrðu teknir út og tyrkneskir stafir settir í staðinn. Fallið var frá þessu vegna ótta manna við að slíkt yrði umdeilt og myndi tefja framgang hinnar nýju töflu vegna sterkrar stöðu Íslendinga í staðlamálum.

Um Windows NT er fjallað í Unicode-kaflanum hér að aftan.

DEC

Í eldri útgáfum VAX-kerfa frá Digital er notuð stafatafla sem nefnist DEC Multinational. Í þeirri töflu var lengi tekist á um hvort setja ætti íslenska eða tyrkneska stafi en DEC mátti ekki vera að því að bíða og lét þau að lokum standa auð. Þessi auðu sæti hafa verið notuð líkt og jókerar ýmist undir íslenska stafi eða annað. Nýrri gerðir VAX-kerfa nota hins vegar Latin-1 töfluna, eins og Windows og Unix-kerfi.

Hewlett Packard

Í eldri HP-kerfum má nota íslensku með stafatöflu sem nefnist Roman 8, en nýrri kerfi nota töfluna Latin-1.

Unicode

Hvað er Unicode?

Unicode eða UCS (Universal Character Set) er 16 bita stafatafla. Með 16 bitum má greina að 2^{16}=65536 tákn. Það nægir til að gefa öllum bókstöfum í öllum helstu tungumálum sem notuð eru í heiminum í dag einkvæmt sæti, ásamt fjöldanum öllum af stærðfræðitáknum, stýritáknum, hljóðritunartáknum og táknum fyrir gjaldmiðla, svo fátt eitt sé nefnt. Þó hefur orðið að takmarka fjölda ritmála sem nota myndletur og hafa t.d. Japanir og Kínverjar orðið að sætta sig við að nota sama myndmálið, þótt sumir viljir gera greinarmun þar á.

Unicode-hópurinn hefur náið samstarf við ISO-nefndina sem semur staðalinn 10646-1 og er 16 bita staðall beggja eins. Gert er ráð fyrir að staðallinn geti notað 32 bita en með því fæst um það bil milljón sæta viðbót. Með 32 bitum er gert ráð fyrir að nóg pláss sé til að koma fyrir rittáknum ýmissa horfinna tungumála og öllu myndletri. Hinsvegar hefur enn sem komið er öll stöðlun verið í 16 bitunum og þar hefur verið komið fyrir ýmsu sem upphaflega stóð til að setja í 32 bita staðalinn síðar meir. T.d. hafa Norðurlandabúar komið rúnaletri sínu í 16 bitana þótt almenna stefnan sé að þar sé ekki annað en lifandi ritmál þjóða.

Finna má ýmsar upplýsingar um Unicode undir slóðinni
http://www.unicode.org.

Flestir virðast sammála um að Unicode sé það sem koma skal í stafatöflumálum. Síðustu árin hafa framleiðendur allra helstu stýrikerfa sett það á stefnuskrá sína að vinna að stuðningi við Unicode í kerfum sínum. Vinna við þetta er mislangt komin, en er víða farin að skila sér nú þegar. Hér á eftir verður fjallað um stöðu Unicode í nokkrum nútímakerfum.

Windows NT

Windows NT 4.0 frá Microsoft, sem nú er víða í notkun í netkerfum, styður Unicode að fullu. Þess ber þó að geta að við gerð forrita fyrir NT þarf að gera sérstakar ráðstafanir til að Unicode sé notað rétt, en þróunartól Microsoft hjálpa forriturum við að gera þetta án þess að upp komi vandamál þegar forritin eru keyrð í Windows 95/98.

Java

Java er forritunarmál sem þróað hefur verið á síðustu árum hjá bandaríska fyrirtækinu Sun Microsystems. Af ástæðum sem of langt mál væri að rekja hér er Java þó annað og meira en forritunarmál, og má segja að það sé stýrikerfi út af fyrir sig. Þannig má ekki einungis búast við því á næstu árum að fram komi tölvur og forrit sem byggist á Java, heldur jafnvel að Java leynist á bak við tölvustýringar ýmissa smátækja og jafnvel greiðslukorta. Í Java hefur frá upphafi verið fullur stuðningur við Unicode.

Fullur stuðningur við Unicode þýðir þó ekki endilega að fullur stuðningur sé við íslensku. Til þess að íslenska sé studd þarf að vera til íslenskt „locale“ fyrir hana sem hefur í sér t.d. upplýsingar um staðlað hnappaborð fyrir íslensku. Í fyrstu útgáfum af Java var ekkert íslenskt „locale“ sem hefur þýtt að vegna vandræða við innslátt íslenskra stafa, t.d. á vefsíðum, hefur Java ekki orðið jafn vinsælt á Íslandi og efni hafa staðið til. Eftir að Staðlaráð Íslands tók að skipta sér af þessu hafa menn hjá JavaSoft unnið að úrbótum á þessu.

BeOS

Stýrikerfið BeOS kom fram á árinu 1997, en hefur hlotið litla útbreiðslu, enda eiga ný stýrikerfi erfitt uppdráttar í samkeppni við risana, hversu mikið sem þau kunna að hafa að bjóða. BeOS notar Unicode í alla textavinnslu.

Mac OS

Apple tók fyrstu skrefin í notkun Unicode með útgáfu 8 af Mac OS, sem fram kom árið 1997. Með þessari útgáfu kerfisins var skipt um grundvallaraðferð við að meðhöndla ólík ritkerfi og tekin upp aðferð sem byggir á Unicode. Þar að auki fylgir kerfinu sérstök eining sem sér um þýðingar milli stafataflna, í gegnum Unicode. Þar er þegar gert ráð fyrir séríslensku töflunni. Í október 1998 kom út útgáfa 8.5 af stýrikerfinu, en þar er stuðningur við Unicode stóraukinn. Hjá Apple er stefnt að því að stýrikerfið Mac OS X, sem á að leiða inn nýja kynslóð Macintosh-stýrikerfa, hafi fullan stuðning við Unicode.

Windows 95/98

Í nýlegri uppfærslu Microsoft á Windows 95, sem kölluð er Windows 98, er enn ekki að finna stuðning við Unicode.

Vefurinn

Í HTML-skjölum sem unnið er með á vefnum er gert ráð fyrir því að í skjali megi skilgreina hvaða stafatafla er notuð í því. Nú þegar þekkja helstu vafrar (Netscape Navigator, Microsoft Internet Explorer) skjöl sem skrifuð eru með Unicode (eða UTF-8 eða UTF-7, sjá hér að aftan) og geta birt þau rétt að því marki sem stýrikerfið ræður við það.

UTF-8

Ef 16 bitar eru notaðir fyrir hvern staf í stað 8, eins og hingað til hefur verið venjan, gefur auga leið að venjulegur texti hlýtur að taka tvöfalt minni miðað við það sem áður var. Þegar eingöngu er verið að nota fáar síður hlýtur að felast í þessu mikil sóun á minni, diskrými og vinnslugetu.

Því hefur verið brugðið á það ráð að hanna sérstakt snið sem nefnist UTF-8. Þegar UTF-8 er notað þarf aðeins eitt bæti (8 bita) fyrir táknin 128 í gömlu 7 bita ASCII töflunni. Hins vegar þarf tvö bæti fyrir næstu 3968 tákn þar á eftir í Unicode-töflunni og fyrir þau 61440 sem eftir eru þarf þrjú bæti. Þannig fæst sparnaður ef 7 bita grunnsettið er mest notað, en ef t.d. er um að ræða kínverskan texta borgar sig ekki að nota UTF-8.

UTF-7

UTF-7 er enn annað snið, sem ætlað er til að tryggja öruggar sendingar gagna um miðla sem ekki er hægt að treysta til að varðveita áttunda bitann, t.d. í tölvupósti. Í þessu sniði eru eingöngu notaðir 7 bitar af hverju bæti og er textinn sendur á formi sem svipar til „Base 64“-formsins sem oftast er notað til að senda viðhengi með tölvupósti.

Hvers vegna Unicode leysir ekki öll stafavandamál

Um 1995 var því almennt trúað að stóra stafataflan yrði brátt tekin í notkun og stafavandamál Íslendinga og togstreita um sæti í stafatöflum t.d. milli Tyrkja og Íslendinga yrðu brátt úr sögunni. Þróunin hefur hinsvegar orðið hægari en vonast var til og enn eru 8 bita stafatöflur ráðandi og líklegar til að verða það næstu áratugi. Í sumum tilvikum er sennilegt að um alla framtíð verði aðeins hægt að nota 7 bita ASCII stafi, t.d. í netföngum og í vegabréfum og skráningakerfum flugfélaga. Í einu stöðluðu leturgerðinni fyrir ljóslestur t.d. á vegabréfum og peningaseðlum hefur ekki tekist að fá því framgengt að bætt sé við íslenskum sérstöfum.

Þótt grunnkerfi tölva styðji Unicode er ekki þar með sagt að allir stafir í Unicode verði birtingarhæfir á skjám eða á prenturum. Til þess að stafir séu birtingarhæfir þurfa þeir að vera skilgreindir í þeirri leturgerð sem verið er að nota. Leturgerðarmenn munu ekki eyða fjármunum í að gera afbrigði af öllum leturgerðum t.d. fyrir norrænar rúnir eða margs konar akademísk táknróf sem slæðst hafa inn í Unicode.

Þess vegna munu verða búin til hlutmengi af Unicode fyrir hin ýmsu markaðssvæði og munu leturgerðarmenn selja leturgerðir fyrir hvert þeirra t.d. fyrir Evrópu eða hluta hennar. Hér er hætta á ferðum fyrir íslenska stafi því að nú þegar eru mörg dæmi um að íslenskum stöfum sé sleppt úr hlutmengi leturgerða fyrir Vestur-Evrópu.

Það gefur auga leið að enginn notandi mun kæra sig um að þurfa að meðhöndla alla stafi í Unicode, t.d. munu verða gerð hlutmengi fyrir farsíma með þeim stöfum sem notendur á hverju markaðssvæði eru líklegir til að þurfa. Þar sem Íslendingar eru einir um að nota ð og þ gæti þetta þýtt að þeir þyrftu að vera einir um slíkt hlutmengi og gætu jafnvel orðið verr settir en áður.

Íslensku sérstafirnir, ð, þ og ý eru ekki í mörgum leturgerðum og þeim fyrirtækjum sem bjóða leturgerðir fyrir íslensku til sölu hefur farið fækkandi síðasta áratug, að sögn sérfræðinga. Einnig hefur útliti stafanna ð og þ oft verið ábótavant. Sjá *Vandamál íslensks leturs efa ég að leysist sjálfkrafa* aftan við þessa skýrslu.

Í mörgum löndum Evrópu er nánast bannað að nota kennitölur til að auðkenna fólk og tengist það minningum um seinni heimsstyrjöld. Í þeim löndum eru nöfn fólks notuð líkt og við notum kennitölur og það er vandamál að ýmiss konar ruglingur skapast af óþekkjanlegum stöfum. Þannig er ekki líklegt að gagnagrunnar á Evrópska vegabréfasvæðinu verði t.d. með grískum stöfum sem lögregluþjónar og landamæraverðir ráða almennt ekki við. Sömu viðhorf eru í allri skráningu farþega hjá flugfélögum og t.d. hafa Flugleiðir og íslenskar ferðaskrifstofur kerfi í notkun sem ekki þekkja íslenska stafi þótt tölvur þeirra ráði við þá og engar ráðagerðir eru um að breyta því.

Af tillitsemi við útlendinga nota tímarit eins og Iceland Review ekki stafina ð og þ í enskum ritum um íslensk málefni þegar vitnað er í íslensk nöfn eða heiti. Sama er um rit sem Flugleiðir dreifa í vélum sínum þótt broddstafir og stafirnir æ og ö séu almennt notaðir. Íslendingar hafa því sætt sig við að íslensku stafirnir ð og þ séu eingöngu til heimabrúks sem aftur merkir að engin ástæða verður fyrir neinn að setja þá í hlutmengi með stöfum annarra þjóða.

Vegna þess að Íslendingar eru ekki í ESB nýtur íslensk tunga ekki þeirrar virðingar að vera talin með opinberum tungumálum ESB sem minnkar markað fyrir íslenska stafi þar sem engin þörf er fyrir þá í öllum þeim gagnagrunnum sem kallaðir eru evrópskir og tölvuframleiðendur vilja gjarna styðja.

Verkefni í stafatöflumálum

Helstu hagsmunamál Íslendinga eru ekki lengur að koma íslensku stöfunum í helstu stafatöflur þar sem þeir eru þar nú þegar, heldur er áríðandi að fylgja eftir ýmsu því sem staðla þarf ofan á stafatöflurnar. Til dæmis þarf að fylgjast með skilgreiningum á hlutmengjum af Unicode og gerð leturgerða (fonta) og hugsanlegri stöðlun þeirra. Ástæða er til að hafa áhyggjur og jafnvel grípa til aðgerða vegna þess að leturgerðarmenn hafa, vegna þrýstings frá Tyrkjum, útbúið leturgerðir þar sem þorn og eð hafa útlit tyrkneskra stafa. Þessar leturgerðir eru gerðar fyrir 8 bita stafatöflur en vegna þess að nöfnin eru þau sömu í Unicode munu þær verða fluttar óbreyttar í 16 bita umhverfið.

Íslendingar hafa almennt ekki gert sér grein fyrir að mikill kostnaður er því fylgjandi að halda úti sérstökum táknum í stafatöflum og leturgerðum heimsins. Þennan kostnað verða Íslendingar sjálfir að bera. Fáar þjóðir í Vestur-Evrópu hafa jafnmarga sérstafi í stafrófi sínu og Íslendingar. Sérstafirnir “ð” og “þ” eru ekki afbrigði af öðrum þekkjanlegum stöfum hefðbundins latínustafrófs og eru einsdæmi ef frá er skilið hið þýska “bollu-s”.

Þetta veldur Íslendingum sjálfum mestum vanda t.d. við innslátt með hnappaborði. Það er sjaldgæft að nokkur þjóð hafi ekki alla sérstafi sína á hnappaborðum tölva. Þannig hafa Danir og Þjóðverjar sérstaka hnappa undir alla sína stafi. Það hafa Frakkar einnig en þeir hafa farið þá leið að setja broddstafi sína þar sem á venjulegu QWERTY hnappaborði eru tölustafirnir 1, 2, 3,... o.s.frv., en til að komast að tölustöfunum er ýtt á skiptihnappinn. Þetta hafa einnig Tékkar gert.

Það sem mestu skiptir á næstunni varðandi staðla og stafatöflur er eftirfarandi:

1. Fyrst og fremst þurfa Íslendingar að gera sér grein fyrir að það kostar fé að halda íslenska stafrófinu áfram inni í alþjóðlegum stöðlum og leturgerðum. Það þarf einnig að vera ljóst að þetta er eingöngu hagsmunamál Íslendinga og því munu engir aðrir en Íslendingar greiða þennan kostnað.
2. Sérstakra aðgerða er þörf við að skilgreina útlit íslenskra stafa í leturgerðum (þ, ð) og þar sem engir aðrir nota þessa stafi er það eingöngu mál Íslendinga.
3. Sérstakra aðgerða er þörf hvað varðar GSM-síma. Sú þróun að farsímar og tölvur rynnu saman í eitt tæki hefur orðið hröð. Íslenskir sérstafir eru ekki í algengustu stafatöflu GSM-síma. Staðlavinna á því sviði hefur ekki verið unnin af sömu aðilum og á tölvusviðinu og þar er starf óunnið.
4. Íslendingar hafa ekki verið fljótir til að koma inn í Unicode ýmsum handritastöfum og eru líklega orðnir of seinir til þess að koma þeim í 16 bita hluta Unicode. Áður en unnið verður að því, þurfa Íslendingar að skilgreina þarfir sínar við varðveislu upplýsinga og vinnslu þeirra og varðveislu í fornskjölum frá öllum tímum Íslandssögunnar.

Það var Íslendingum mikið happ að fá til sín rekstur stafanefndar Evrópusambandsins TC304, en rekstur hennar hefur verið fjármagnaður af ESB. Þetta hefur tryggt Íslendingum mikil áhrif án eigin fjárútláta. Til samanburðar má taka hversu illa hefur tekist til um stöðlun á sviði fjarskipta en þar hafa Íslendingar hvorki haft áhrif né þekkingu á þeim vandamálum sem glímt hefur verið við.

Í framtíðinni er ekki hægt að búast við að ESB greiði reikninga fyrir íslenska sérstafi í tölvum. Því þurfa Íslendingar að vera reiðubúnir til að greiða kostnaðinn sem þeir hafa af hinum miklu kröfum sem þeir gera til stafrófs í tölvum. Rugl í stafatöflum og leturgerðum leiðir af sér bæði beinan og óbeinan kostnað t.d. vegna leiðréttinga á innslætti, sem sparast við stöðlun.

3. Ritað mál

Málsöfn

Öll hagnýting tungutækni við meðferð ritaðs máls er mjög háð tilvist tvenns konar gagnasafna sem byggja þarf upp á skipulegan hátt fyrir hvert tungumál. Annars vegar þarf að koma upp **textaheild** (corpus) málsins, og hins vegar **orðasafni** (lexicon) þess.

Með *textaheild* er átt við mengi texta af ýmsu tagi sem sett er saman eftir ákveðnum reglum og í ákveðnu augnamiði. Reglurnar geta t.d. varðað lengd textanna, tegundir þeirra (bókmenntatextar, fræðitextar o.s.frv.), höfunda (aldur, kyn, uppruni) o.fl. Þessar reglur eru settar til að hægt sé að halda því fram að textaheildin gefi raunsanna mynd af því sem hún á að bera vitni um, en sé ekki tilviljanakennt samsafn texta sem engar öruggar ályktanir verði dregnar af.

Nú á dögum eru textaheildir nær undantekningarlaust settar saman úr tölvutækum textum, og á seinustu árum hafa verið byggðar upp viðamiklar textaheildir á ýmsum tungumálum. Þessar textaheildir hafa verið hagnýttar á ýmsum sviðum, ekki síst í orðabókagerð. Textaheildir eru einnig nauðsynlegar við gerð málgreiniforrita. Þær auðvelda mönnum að sjá í hvaða samhengi tiltekið orð getur komið fyrir, og einnig hvað er algengt og hvað sjaldgæft í málinu. Slíkar upplýsingar er t.d. hægt að nýta til að auka möguleikana á réttri greiningu tvíræðra orða eða orðasambanda.

Íslensk orðtíðnibók sem Orðabók Háskólans gaf út 1991 er byggð á textaheild sem sett var saman í þessum tilgangi. Í hana vantar þó ýmsar tegundir texta; t.d. er þar ekkert talmál. Þar að auki eru þeir textar sem þar eru notaðir orðnir áratugar gamlir eða meira, en slíka textaheild þarf stöðugt að uppfæra til að hún hafi að geyma þann orðaforða sem notaður er hverju sinni. Meginatriðið er þó að þessi textaheild er mjög lítil, 500 þúsund lesmálsorð, en þær textaheildir sem nú er verið að koma upp fyrir ýmis grannmál hafa að geyma tugi og jafnvel hundruð milljóna orða.

Tölvutæk orðasöfn eru einnig nauðsynleg til að unnt sé að útbúa flest tól tungutækninnar. Slík orðasöfn eru á margan hátt hliðstæð venjulegum orðabókum; þar er að finna upplýsingar um framburð flettiorðanna, orðflokk, beygingu, merkingu og setningafræðilega stöðu. Það sem greinir milli venjulegra orðabóka og orðasafna til nota í tungutækni er einkum að upplýsingar í hinum síðarnefndu þurfa annars vegar að vera miklu ítarlegri og fjölbreyttari, og hins vegar settar fram á miklu skipulegri hátt.

Tölvutækt íslenskt orðasafn af þessu tagi er ekki til. Svokallaður *norrænn orðabókarstofn* sem Orðabók Háskólans hefur unnið að fyrir norræna styrki gæti þó orðið vísir að slíku safni. Þar eru margvíslegar upplýsingar um grunnorðaforða málsins skráðar á skipulegan hátt. Mjög mikið vantar þó á að þetta safn hafi að geyma allar þær upplýsingar sem nauðsynlegar eru til að það geti nýst í tungutæknitólum.

Hér skal einnig bent á að tölvutæk orðasöfn af þessu tagi gagnast ekki eingöngu innan tungutækni. Tilvist þeirra gæti einnig skipt sköpum fyrir framtíð orðabókagerðar í landinu. Mikill skortur er á góðum orðabókum milli íslensku og annarra mála. Ýmis dæmi má nefna um útgáfufyrirtæki sem hafa

reist sér hurðarás um öxl við útgáfu orðabóka, enda hefur þar oftast orðið að byrja frá grunni. Ef til væri vandaður tölvutækur orðabókarstofn sem fyrirtæki gætu fengið aðgang að myndu forsendur fyrir íslenskri orðabókagerð gjörbreytast.

Málgreining

Málgreining (parsing) fer yfirleitt fram í tölvu, og felst í ýmiss konar málfræðilegri greiningu texta. Slík greining er undirstaða ýmiss konar tungutækni, s.s. vélrænna þýðinga, málfræðiforrita o.fl.

Greiningin getur verið misjafnlega nákvæm. Stundum er textinn aðeins greindur í orðflokka, en einnig getur verið um að ræða greiningu í setningarliði og greiningu á venslum þeirra, þ.e. formgerð setninga. Slík greining verður sjaldan fullkomlega rétt, en þó er hægt að ná mjög góðum árangri ef greiningin byggist á vönduðu gagnasafni. Þar er annars vegar um að ræða textaheild og orðasafn sem forritið leitar í til að finna upplýsingar um einstakar orðmyndir, og hins vegar reglusafn sem hefur að geyma upplýsingar um leyfilega gerð setninga og setningarliða í málinu.

Einn meginvandinn í málgreiningu felst í greiningu tvíræðra orðmynda. Orðmyndin *á* í íslensku getur t.d. táknað býsna margt. Hún getur verið forsetning, eins og í *Ég bý á Íslandi*; nútíð so. *eiga*, eins og í *Stelpan á þessa bók*; nefnifall, þolfall eða þágufall kvenkynsorðsins *á*, eins og í *Þessi á er straumhörð*; og þolfall eða þágufall kvenkynsorðsins *ær*, eins og í *Hún gaf mér flekkótta á*. Auk þess getur verið um að ræða heiti bókstafsins *á*, upphrópunina *á*! sem táknar sársauka og e.t.v. fleira.

Til að greina *á* rétt verður því að fara eftir samhengi. Ef á stendur næst á eftir sögn en á undan nafnorði, eins og í *Ég bý á Íslandi* eru allar líkur á að um forsetningu sé að ræða, því að sagnir og nafnorð standa varla í þeirri stöðu. Standi á hins vegar næst á eftir nafnorði en á undan ábendingarfornafni, eins og í *Stelpan á þessa bók*, er líklegast að um sé að ræða sögn, því að sjaldgæft er að finna forsetningar í þessari stöðu, hvað þá nafnorð.

Í þessum dæmum dugir setningafræðileg greining yfirleitt, vegna þess að um er að ræða mismunandi orðflokka. Til að greina milli orðmynda af sama orðflokki getur þurft að grípa til merkingarlegrar greiningar. Í dæmunum *Ég sá straumharða á* og *Ég sá flekkótta á* stendur *á* í sams konar setningafræðilegu umhverfi, en lýsingarorðin sýna að í fyrra dæminu er átt við no. *á* en í því seinna no. *ær*. Í setningunni *Ég sá þessa á í gær* er hins vegar útilokað að greina um hvort orðið er að ræða. Að vísu er hugsanlegt að næstu setningar á undan eða eftir leysi málið (t.d. ef næst kæmi *Hún var kolmórauð eftir leysingarnar* eða *Hún hafði týnt öðru lambinu*); en óvíst er hvort málgreiniforritið gæti nýtt sér slíkar upplýsingar.

Leiðréttingaforrit

Til eru ýmis forrit sem lesa tölvutæka texta og benda á villur eða hugsanlegar villur í þeim. Einföldustu forritin af því tagi leita að **stafsetningarvillum** (spell checkers). Slík forrit hafa þá innbyggt safn rétt ritaðra orðmynda. Þau lesa textann, orð fyrir orð, og bera saman við orðasafnið. Ef þau finna

orðmynd sem ekki er í orðasafninu stansa þau og vekja athygli notandans á þessari orðmynd. Stundum er um að ræða rétt ritað orð sem ekki er í safninu, og þá er gefinn kostur á að bæta því í safnið; en sé um rangt ritað orð að ræða er hægt að leiðrétta það. Sum slík forrit, t.d. Púki, búa líka yfir upplýsingum um hvernig orð geti verið til í málinu.

Leiðréttingaforrit af þessu tagi byggjast yfirleitt ekki á málgreiningu og finna því ekki villur þar sem leyfileg orðmynd er notuð á röngum stað. Í setningunni *Ég hitti Þórarinn* er *Þórarinn* í þolfalli og á aðeins að hafa eitt *n*; en vegna þess að *Þórarinn* með tveimur *n*-um er leyfileg mynd (rétt nefnifallsmynd) gerir forritið ekki athugasemd. Í setningunni *Vatnið síður* er seinna orðið rangt ritað; þar á að vera *sýður* (af so. *sjóða*). Vegna þess að *síður* með *í* er leyfilegt orð (atviksorð) gerir forritið heldur ekki athugasemd í þessu dæmi. Til að svo mætti vera þyrfti frekari greiningu textans.

Af svipuðum toga eru **orðskiptiforrit** (hyphenation programs) sem skipta orðum milli lína í samræmi við reglur viðkomandi tungumáls. Slík forrit byggjast yfirleitt á ákveðnum reglum um það hvar í stafasamböndum megi skipta orðum. Í íslensku verður t.d. að vera a.m.k. eitt sérhljóð í hvorum hluta; *skrjóðs* má hvorki skipta *skrj-óðs* né *skrjó-ðs*. Ekki má heldur flytja aðeins eitt sérhljóð milli lína; og seinni hlutinn á yfirleitt að hefjast á sérhljóði. Undantekningar frá síðastnefndu reglunni eru þó fjölmargar, einkum í samsettum orðum. Orðskiptiforrit geta því ekki unnið eingöngu eftir almennum reglum, heldur verða líka að byggjast á orðasafni þar sem talin eru upp helstu orð sem víkja frá reglunum.

Til eru leiðréttingarforrit fyrir íslenska stafsetningu, t.d. *Púki* Friðriks Skúlasonar, og einnig hafa verið gerð nokkur íslensk orðskiptiforrit, t.d. *Skipta* hjá Íslenskri málstöð. Þessi forrit hafa þó að því er virðist ekki náð verulegri útbreiðslu, a.m.k. ekki í seinni tíð, e.t.v. vegna þess að þau eru ekki samhæfð helstu ritvinnsluforritum á markaðnum.

Erlendis hafa einnig verið skrifuð ýmis forrit sem skoða **málfar** og **stíl** (grammar / style chekers). Þau geta t.d. gert athugasemdir við orðaröð, orðanotkun o.fl., allt eftir því hversu fullkomin þau eru. Margir munu hafa reynslu af því að slík forrit fyrir ensku eru orðin mjög þróuð. Málgreining er hins vegar forsenda forrita af þessu tagi, og engin slík eru til fyrir íslensku.

Orðabækur

Orðabækur af ýmsu tagi eru nauðsynleg hjálpargögn við ritun texta. Þar er einkum um að ræða **stafsetningarorðabækur**, eins og *Réttritunarorðabók* Íslenskrar málnefndar og Námsgagnastofnunar; **samheitabækur**, eins og *Íslenska samheitaorðabók*; og að sjálfsögðu hefðbundnar **orðabækur** með beygingarupplýsingum og merkingarskýringum, eins og *Íslenska orðabók handa skólum og almenningi* (Orðabók Menningarsjóðs / Máls og menningar). Auk þess má auðvitað nefna ýmiss konar tvímálaorðabækur, s.s. íslensk-enska og ensk-íslenska orðabók.

Á síðustu árum hefur verið ör þróun í þá átt víða erlendis að gera slíkar orðabækur tölvutækar og tengja þær beint við ritvinnsluforrit. Notandi getur þá hvenær sem er kallað á upplýsingar úr orðabókunum og nýtt þær við það sem hann er að gera. Þetta er að sjálfsögðu mun fljótlegra og þægilegra en

þurfa að vera með margar orðabækur sér við hlið og fletta upp í þeim á hefðbundinn hátt. Einnig má halda því fram að tilvist tölvutækra orðabóka auki líkur á að menn nýti sér þær upplýsingar sem þar er að finna og skrifi þannig vandaðri texta.

Engar tölvutækar íslenskar orðabækur eru til, að frátalinni ensk-íslenskri orðabók Aldamóta sem Mál og menning gefur nú út. Mál og menning stefnir reyndar að því að gefa *Íslenska orðabók* út tölvutæka síðar á þessu ári, en óljóst er hvaða form verður á þeirri útgáfu og hvernig hún mun vinna með ritvinnsluforritum.

Bæði *Réttritunarorðabók* og *Íslensk samheitaorðabók* eru til tölvutækar, og því mætti virðast einfalt að gefa þær út tölvutækar og nýta á þann hátt sem lýst er hér að framan. Hér verður þó að hafa í huga að þær eru samdar sem hefðbundnar bækur, og notkun þeirra sem tölvutækra safna lýtur allt öðrum lögmálum. Því þyrfti að leggja mikla vinnu í að breyta formi þeirra og uppsetningu ef nýta ætti þær tölvutækar, og hugsanlegt að borgi sig eins vel að byrja frá grunni.

4. Talað mál

Talgervlar

Talgervlar eru forrit sem bera fram texta sem skráður er í tölvu. Áður voru þetta gjarna sérstök tæki þar sem hljóðin voru mynduð í sérstökum rafeindarásum, en nú eru algengustu talgervlar forrit sem keyrð eru á venjulegri PC tölvu og nýta hljóðkerfi hennar til að bera fram hljóðið. Þá er gjarna sérstök skipun í ritvinnsluforritum sem kallað er á til þess að bera fram textann sem er verið að skoða eða skrifa í ritvinnsluforritinu. Notandi þarf því ekki annað en forritið og venjulega margmiðlunartölvu.

Hvernig vinna talgervlar?

Nokkrar gerðir talgervla eru á markaði sem vinna að hluta til misjafnt. Það er sameiginlegt talgervlum að þeir taka texta og hljóðrita hann yfir í hljóðstafróf (hljóðön) á sama hátt og gert er t.d. í orðabókum. Við þetta eru notuð sérstök forrit sem eru sértæk fyrir hvert tungumál. Þar er um að ræða reglur um hvernig hljóðrita skuli ákveðna orðhluta og orð. Gerð slíkra forrita krefst góðrar þekkingar á hljóðfræði viðkomandi tungumáls.

Slík forrit geta verið misítarleg, t.d. geta þau sleppt því að gera grein fyrir tónfalli, eða mun á áhersluatkvæðum og áherslulausum atkvæðum en við það minnka gæði talsins. Í sænsku er t.d. nauðsynlegt að gera greinarmun á orðunum „anden“ í merkingunni „andinn“ og „anden“ í merkingunni „öndin“, en þessi orð eru ekki borin eins fram í sænsku, þótt einföld hljóðritun segi að svo sé, heldur er merkingarmuni komið til skila með hljómfalli sérhljóða.

Eftir að texti hefur verið hljóðritaður skilur á milli aðferða:

Ein aðferð er að líkja eftir því sem gerist í öndunarfærum fólks þegar það talar. Þessa aðferð má greina í tvennt. Lengi vel voru notuð sérstök tæki til þess að bera fram hljóðstafina. Í þeim voru rafrásir, hljóðsíur, sem líktu eftir því sem gerist í munni fólks þegar það talar. Þegar tölvur urðu öflugri var farið að gera stafræn líkön af öndunarfærunum og þau var hægt að keyra á venjulegum tölvum og þurfti þá ekki lengur þessar sérstöku rafrásir heldur nýtir líkanið sér hljóðkort tölvunnar til þess að bera fram orðin.

Kosturinn við talgervla sem líkja eftir öndunarfærum á einn eða annan hátt er sá að þeir eru alþjóðlegir á sama hátt og hljóðtáknin. Þetta byggist á því að öndunarfæri fólks eru eins hvaða tungumál sem það talar. Með því að nota mikinn fjölda hljóðtákna og stilla nákvæm hljóðgildi fyrir hvert tungumál er auðvelt að laga slíka talgervla að nýju tungumáli. Galli við þessa aðferð er að ekki hefur tekist að ná fram miklum hljómgæðum með einföldum líkönum af talfærunum.

Hér þarf að taka fram að mörg nöfn eru á þessum líkönum en þau eru að mestu jafngild. Á ensku er talað um vocal-tract model, formant synthesis, Liniear Predictive Coding og margt fleira. Til þess að auka gæðin hefur reynst nauðsynlegt að gera líkönin svo flókin að kostir einfaldleikans hverfa og þá hefur reynst auðveldara að ná fram miklum gæðum með annarri aðferð.

Hin aðferðin er að nota búta úr tali til þess að bera fram orðin. Þegar aðeins er fengist við takmarkaðan orðaforða er auðvelt að gera fullkomna talgervla af þessari gerð. Einföld notkun slíkra talgervla sem flestir þekkja eru símsvarar, ungfrú klukka og bankavélar. Þar eru fá orð notuð og verkefnið því miklu léttara en í almennu tali. Þessi aðferð hefur þó orðið ofan á í öllum nýrri gerðum talgervla sem sagt geta hvaða orð sem er í tilteknu tungumáli. Þá er tekin upp rödd manns (eða konu) og hún bútuð niður í einingar sem síðan eru settar saman aftur eftir því sem hljóðritunin segir til um.

Það fer eftir krafti tölvunnar hvernig þetta er gert. Minnstu bútarnir eru ekki hefðbundin hljóðtákn, heldur er tekin nokkur hljóðtákn saman (á ensku diphones). Þótt þannig verði að geyma upptökur af nokkrum tugum þúsunda hljóðbúta er slíkt vel viðráðanlegt fyrir nútíma heimilistölvu sem með þessari aðferð getur borið fram hvaða texta sem er. Gallinn er að til þess að bæta gæði talsins þarf að fjölga hljóðbútum og stækka þá og fjöldinn getur orðið of mikill.

Í frumstæðum talgervlum af þessari gerð er hljóðbútum oft illa raðað saman og hlustandinn heyrir óþægilegt brak. Í nýrri gerðum eru notaðar ýmsar síur og forrit sem gera samsetninguna svo til fullkomna og ennfremur geta þau mótað talið sem framleitt er þannig að hægt er að stýra talhraða, hljómfalli og áherslum.

Allmörg vandamál eru enn óleyst í sambandi við talgervla þannig að þeir verði áheyrilegir. Þau eru bæði tæknileg og hljóðfræðileg. Hljómfall er ekki eðlilegt og áherslur rangar, en erlendis er mikil vinna lögð í að bæta þá.

Talgervlar fyrir ensku eru algengir. Þeir eru vel skiljanlegir og nýjustu og bestu gerðirnar eru vel áheyrilegar. Þá er auðvelt að fá fyrir venjulegar PC tölvur. Einn íslenskur talgervill er í notkun. Hann er byggður á talgervli frá sænska fyrirtækinu Infovox og var aðlagaður að íslensku af Málvísindastofnun Háskólans, verkfræðideild Háskólans og Öryrkjabandalaginu á árunum 1989 til 1993. Verkið var að mestu unnið af Pétri Helgasyni málfræðingi og styrkt af Nordiska nämden for handikappfrågor.

Talgervillinn hefur verið notaður af blindum og gagnast vel. Gæðin eru samt ekki nægilega mikil til þess að hann sé notaður af almenningi. Nýlega var talgervillinn lagfærður og ný útgáfa er væntanlega á næstu vikum. Talgervillinn er af þeirri gerð sem líkir eftir öndunarfærum fólks.

Á Íslandi væri hráefni fyrir talgervla hljóðritað orðasafn og upptökusafn af texta sem hægt væri að nota til að afla tölfræðilegra upplýsinga hluti eins og lengd, áherslu og hljómfall. Einnig væri gagn af sjálfvirkri setningagreiningu sem myndi auðvelda innsetningu á réttu hljómfalli.

Sýnishorn af tali talgervils frá Infovox sem er nú í eigu Telia er að finna á vefsíðunni: www.promotor.telia.se/infovox/ Á þessari vefsíðu er bæði hægt að hlusta á tal með eldri og nýrri gerð talgervla (diphone-gerð). Ekki stendur til hjá Telia að þróa nýrri gerð talgervla fyrir íslensku; til þess er kostnaður talinn of mikill og íslenski markaðurinn of lítill.

Talgreining

Talgreining fer þannig fram að hljóðnemi er tengdur við hljóðkort tölvunnar. Hljóðneminn þarf að vera góður og til þess gerður að lágmarka hljóð í bakgrunni, en slík hljóð valda vandræðum því þau leggjast við það sem notandi segir og aflaga hljóðið sem tölvan heyrir. Forritin búta talið í sundur og tíðnigreina hvern bút og bera saman við mikið safn hljóðeininga sem geymdar eru í minni tölvunar. Forritin vinna ekki með einstök hljóðtákn heldur þekkja þau stærri máleiningar. Notaðar eru tölfræðilegar aðferðir og þau orð valin sem mestar líkur virðast vera á að séu rétt.

Til þess að geta gert þetta þarf forritið helst að vita hvernig notandinn talar, og því þarf notandinn áður en hann notar forritið í fyrsta skipti að bera fram nokkurn fjölda þekktra setninga. Í forritið er innibyggð þekking á tungunni svo að þekking á stöðu orða í setningu getur að einhverju leyti auðveldað greininguna. Staða orðs í setningu er einnig notuð til þess að greina á milli orða sem hljóma eins eða svipað. Dæmi um það eru ensku orðin „to“, „too“, og „two“ sem hljóma eins. Tölvan giskar á hvert þeirra er um að ræða út frá þeim orðum sem næst orðinu standa eins og til dæmis: „walk to London“, „walk two miles“, „walk too far“.

Þessi vinnsla krefst mikils tölvuafls. Þær PC tölvur sem nú eru mest notaðar og hafa 200 - 300 MHz klukku gera ekki mikið meira en að ráða við þetta og geta t.d. ekki kannað nema þrjú nálæg orð. Miklar framfarir hafa orðið í talgreiningu á undanförnum árum og þótt aðferðirnar sem notaðar eru byggi fyrst og fremst á stórum gagnagrunnum og mikilli reiknigetu tölva, hefur í raun reynst auðveldara að ná fram fullkominni talgreiningu en fullkomnum talgervlum.

Til þess að búa til forrit til talgreiningar þarf stór hljóðsöfn töluð af mörgum einstaklingum þar sem eru allir orðhlutar tungunnar. Úr þessum hljóðsöfnum þarf síðan að draga upplýsingar um málhljóðin, tölfræðilegar upplýsingar um stöðu orða í setningum og annað slíkt. Frekari fræðilegan bakgrunn er að finna á vefsíðum um tungutækni sem vísað er til í viðauka 7. Sjá einnig viðauka 3. Í viðauka 7 er vísað til heimasíðna helstu framleiðenda og dóma um þau forrit til talgreiningar sem nú eru á markaði.

5. Vélrænar þýðingar og leitir á vefnum

Vélrænar þýðingar

Stefán Briem hefur ritað eftirfarandi greinargerð um vélrænar þýðingar fyrir starfshópinn:

Hvað eru vélrænar þýðingar?

Vélrænar þýðingar á rituðu efni milli tungumála felast í notkun þar til gerðra tölvuforrita sem taka við texta á tilteknu *tungumáli*, *frummálinu*, og skila efni hans í samsvarandi texta á öðru tungumáli, *viðtökumálinu* eða *markmálinu*.

Ef forritið vinnur sjálfstætt án afskipta manna frá því að forritið tekur til starfa og þangað til það skilar niðurstöðu er talað um *sjálfvirkar þýðingar*. En meiri gæði nást að jafnaði í þýðingum með samspili tölvu og mannshugar. Það getur t.d. verið með þeim hætti að tölvan varpar spurningu til notanda tölvunnar þegar álitamál kemur upp og bíður svars en heldur áfram vinnslu að því fengnu. Má þá tala um *hálfsjálfvirkar þýðingar*. Annar háttur er sá að tölva *hráþýðir* eða *grófþýðir* vélrænt en mannlegur þýðandi tekur síðan við niðurstöðunni, leiðréttir villur og lagar málfar. Og enn einn háttur er sá að forvinna textann áður en hann er vélþýddur þannig að þýðingarforritið ráði betur við hann en ella.

Vélrænar þýðingar eru eitt helsta viðfangsefni tungutækni og tengjast flestum öðrum þáttum hennar. Menn hófu rannsóknir og þróun á sviði vélrænna þýðinga fyrir fimmtíu árum en framfarir hafa orðið mun hægari en menn gerðu sér vonir um í upphafi. Erfiðasti þátturinn í vélrænum þýðingum er merkingargreiningin, þ.e. að láta forrit greina merkingu textans sem þýða á, en gæði vélrænna þýðinga eru mjög undir því komin hvernig til tekst um þennan þátt. Tölvur munu aldrei leysa alveg af hólmi mannlega þýðendur en þær geta nú þegar flýtt fyrir og létt störf þýðenda verulega, a.m.k. á afmörkuðum sviðum, ef skynsamlega er staðið að málum. Á Íslandi hefur vélrænum þýðingum lítið verið sinnt hingað til nema í tómstundum áhugamanna.

Talað mál og táknmál

Fyrir vélrænar þýðingar á *töluðu máli* nýtast sömu aðferðir og í vélrænum þýðingum á rituðu máli en auk þeirra þarf *talgreini*, búnað sem umbreytir tali í texta, og *talgervil*, búnað sem umbreytir texta í tal. Íslenskir talgervlar eru til en þarfnast endurbóta. Hins vegar er íslenskur talgreinir ekki til enda er mun erfiðara viðfangs að láta tölvu greina tal en að búa til tal úr texta. Erlendis hafa allra síðustu árin orðið talsverðar framfarir í þróun talgreina. T.d. kom síðla árs 1998 á almennan markað enskur talgreinir fyrir einmenningstölvur. Þess er vænst að vélrænar þýðingar á töluðu máli muni í framtíðinni gera mönnum af ólíku þjóðerni kleift að talast við, t.d. í síma, þannig að hvor viðmælandi noti eingöngu sitt eigið móðurmál. Sama tækni mun þá nýtast til að láta tölvu taka við hlutverki túlks að þeim takmörkum sem gæðakröfur setja.

Tungutækni getur líka verið til gagns fyrir fólk sem er ekki fært um að tjá sig á venjulegu tungumáli en notar þess í stað *táknmál* sem „talað“ er með hreyfingum handa og höfuðs. Með þar til gerðum búnaði er möguleiki á að umbreyta þessum hreyfingum í tölvuskráð merki sem síðan má beita aðferðum

tungutækni og umbreyta að einhverju leyti í venjulegt tungumál, og öfugt. Rannsóknir og þróun á þessu sviði eru skammt á veg komnar erlendis og enn skemur hérlendis.

Markmið

Markmið með vélrænum þýðingum á rituðu máli geta verið ýmiss konar eftir því að hvers konar notkun og að hve miklum gæðum er stefnt.

Við leit á *veraldarvefnum* að erlendu efni getur íslenskum notanda t.d. oft nægt að fá óvandaða þýðingu á íslensku sem sýnir um hvað er fjallað án þess að allt sé rétt þýtt. Þeir sem hafa notað leitarkerfið *AltaVista* á veraldarvefnum ættu að kannast við þá þýðingarþjónustu sem þar er nú veitt til að þýða texta og *vefsíður* vélrænt og sjálfvirkt milli ensku annars vegar og frönsku, ítölsku, portúgölsku, spænsku og þýsku hins vegar. Þær þýðingar sem AltaVista skilar eru ekki áreiðanlegar þó að þær gefi að jafnaði grófa hugmynd á viðtökumálinu um efni textans. En þarna sakna Íslendingar, og reyndar flestar þjóðir heims, sárlega móðurmáls síns.

Í öðrum tilvikum getur verið fullnægjandi að upplýsingar frumtextans komist rétt til skila á viðtökumálinu þó að málfari sé ábótavant. Þetta getur t.d. átt við vörulýsingar á umbúðum, orðsendingar í tölvupósti, veðurskeyti og önnur fréttaskeyti þar sem hraði í þýðingum skiptir jafnan meira máli en fagurt málfar.

Oft er þó þörf á þýðingum sem koma efni frumtextans til skila bæði óbrengluðu og á vönduðu máli. Forsenda þess að samin verði þýðingarforrit sem rísa undir slíkum kröfum eru stóreflda rannsóknir í mörg ár eða áratugi á sviði vélrænna þýðinga.

Hvað þarf til að geta þýtt vélrænt af og á íslensku?

Í fyrsta lagi þarf forrit til að þýða af a.m.k. einu erlendu tungumáli á íslensku og forrit til að þýða af íslensku á a.m.k. eitt erlent tungumál. Með tengingu við önnur *þýðingarforrit* má svo ná til annarra tungumála. Meginhugmynd um slíka tengingu er að nota eitt sameiginlegt millimál í þýðingum milli allra þjóðtungna. Alþjóðamálið esperanto hefur verið nefnt til leiks sem slíkt *millimál* vegna einfaldleika þess, röklegrar uppbyggingar og miklu minni *margræðni* en þekkist í þjóðtungum. En einnig kemur til greina að millimálið sé ekki mannamál heldur einhvers konar tölvuvinnslumál sem geymir tungumálaupplýsingar í öðru formi og ekki er hægt að lesa eða tala eins og venjulegt tungumál.

Í öðru lagi þurfa þýðingarforritin að hafa aðgang að umfangsmiklum tölvuskráðum upplýsingum um þau tungumál sem þýðingarnar taka til. Fyrir íslensku þarf tölvuskráð *orðasafn*, sem tengir hana við a.m.k. eitt annað mál, tölvuskráð *beygingarkerfi*, sem tekur til allra beygjanlegra íslenskra orða, og fjölbreyttar upplýsingar um notkun einstakra orða og orðasambanda. Undir síðastnefnda þáttinn má einnig fella ítarlegar tölvuskráðar upplýsingar um *umheiminn* sem þarf til að skera úr þegar orð eða setningahlutar geta haft margar merkingar. Rétt merking ræðst einkum af samhengi og þekkingu á umheiminum. Maðurinn ratar oftast á rétta merkingu ómeðvitað eftir reynslu sinni og þekkingu en þýðingarforrit þarf að hafa ílæga aðferð til að velja

líklegustu merkinguna, aðferð sem er hugsuð út við samningu forritsins og látin byggjast á þekkingu tölvunnar og ef til vill einnig reynslu hennar.

Á hverju á að byrja?

Brýnustu verkefni Íslendinga á sviði vélrænna þýðinga, sé litið til skamms tíma, eru að mati þess sem þetta ritar að setja saman grunnhugbúnað fyrir íslenska tungu, sem nýtist einnig á öðrum sviðum tungutækni og jafnvel víðar, og gera hann aðgengilegan. Hér er einkum átt við tvennt:

1. Að koma á fót og halda við stóru tölvuskráðu tvíátta orðasafni íslensku og einhvers annars tungumáls og veita af því frjáls afnot og aðgengi m.a. á *Lýðnetinu* (Internetinu). Orðasafnið ætti að vera þannig fram sett að það geti nýst til margvíslegra verkefna, ekki aðeins fyrir vélrænar þýðingar heldur einnig fyrir önnur verkefni sem stofnanir, fyrirtæki og einstaklingar kunna að stofna til á sviði tungutækni, t.d. *orðskiptiforrit*, *leiðréttingarforrit* og hjálparbúnað fyrir þýðendur, sem gæti verið ílægur í notkunarforritum á borð við ritvinnsluforritið *Word*.
2. Að koma á fót og halda við tölvuskráðu beygingarkerfi fyrir íslenska tungu, með sams konar aðgengi og ofannefnt orðasafn, þar sem skráðar yrðu tæmandi upplýsingar um beygingar íslenskra orða. Tölvuskráð beygingarkerfi af þessu tagi myndi ekki aðeins nýtast á mörgum sviðum tungutækni heldur einnig í skólum og fyrir almenning til uppflettingar.

Þessi tvö verkefni eru þess eðlis að mjög hlýtur að koma til álita að þau verði kostuð að mestu eða öllu leyti úr sameiginlegum sjóðum landsmanna. Annars er hætta á að þeim verði ekki sinnt með verðugum hætti og að aðgengi að þeim verði ekki eins gott og vera þarf til að þau nýtist vel. Þessi verkefni eru um leið forsenda fyrir góðum árangri í næstu verkefnum.

Sé litið til lengri tíma er áríðandi að bíða ekki með næstu skref heldur að hefjast handa sem fyrst og ef fjármunir þykja takmarkaðir að beina þá kröftunum fyrst að verkefnum sem munu skila sér fjárhagslega fyrir íslenskt samfélag í fyrirsjáanlegri framtíð. Meðal vænlegra verkefna eru eftirtalin:

1. Hönnun takmarkaðra kerfa fyrir vélrænar þýðingar, annars vegar þar sem þýðingar þurfa ekki að vera villulausar né gæði mikil og hins vegar þýðingar á takmörkuðum efnissviðum þar sem texti er annaðhvort mjög einfaldur eða auðugur af endurtekningum. Eitt gagnlegasta verkefni af þessu tagi eru vélrænar þýðingar á borð við þær sem AltaVista býður upp á og nefndar hafa verið.
2. Þróun talgreinis fyrir íslensku og endurbætur á þeim íslenska talgervli sem bestur þykir. Íslenskur talgreinir mun skila sér fjárhagslega í stórfelldum vinnusparnaði við tölvuskráningu miðað við innslátt á hnappaborð.

Ýmis þeirra verkefna sem hér hafa verið nefnd kalla á samstarf við aðrar þjóðir, bæði við rannsóknir og hönnun búnaðar og við að gera árangur úr verkefnunum aðgengilegan í algengum notkunarhugbúnaði á heimsmarkaði. En þó er ljóst að það sem snýr beint að íslensku hlýtur að hvíla að langmestu leyti á Íslendingum sjálfum.

Íslenska á vefnum — leitarvélar

Mikilvægt er að á netinu séu aðgengilegar leitarvélar sem ráða við að leita að skjölum á íslensku.

Íslenskir stafir

Í þessu tilliti er auðvitað brýnast að vélarnar taki við íslenskum stöfum og leiti rétt að þeim. Þær þurfa að geta áttað sig á samhengi hástafa og lágstafa, t.d. að ‘Þ’ sé sambærilegt við ‘þ’. Sumar algengar leitarvélar virðast bjóða upp á þetta nú orðið, og má þar nefna eina vinsælustu leitarvélina, AltaVista www.altavista.com. Helsta forsenda þess að þessir hutir séu í lagi er að íslenskir stafir séu í algengustu stafatöflum, eins og nú er raunin með ISO-8859-1. Rétt er að fylgjast með helstu leitarvélum og hvetja aðstandendur þeirra til að sjá til þess að í þeim megi nota íslenska stafi sem aðra.

Beygingar

Sumar leitarvélar ráða við að leita að mismunandi beygingarmyndum sama orðs, sem gerir allar leitir mun skilvirkari. Þetta er misjafnlega flókið í framkvæmd, allt eftir beygingarreglum málsins. Fyrir sum tungumál er nóg að greina stofninn og leita svo að öllum orðmyndum sem byrja á honum. Hins vegar þarf flóknari aðferðir til að greina orðmyndir í tungumálum á borð við íslensku, sem hafa innri beygingu, þ.e.a.s. mál þar sem beygingar koma ekki eingöngu fram í endingum, heldur einnig breytingum á stofni orðanna (saga ~ sögu, fara ~ fór).

E.t.v. er þó ekki ástæða til að Íslendingar ráðist í sérstakar framkvæmdir til að sinna þessum málum. Huga þarf að svipuðum atriðum í tengslum við mörg útbreiddari mál (t.d. þýsku, dönsku, arabísku), og líklegt er að fram komi lausnir fyrir þessi tungumál sem byggja megi á við gerð öflugrar leitarvélar fyrir íslensku. Þó má athuga hvort hægt væri að nýta til þessa verks greiningartæki á borð við villupúka Friðriks Skúlasonar, sem greinir beygingarmyndir.

Rökvirkjar

Í mörgum leitarvélum má nota rökvirkja til að skilgreina nákvæmari leit. Þannig má nota „and“ til að takmarka leit við skjöl sem innihalda tvö orð sem tilgreind eru, „or“ til að biðja um öll skjöl sem innihalda annaðhvort orðið og „not“ til að losna við skjöl sem innihalda tiltekið orð. Þannig gæfi leit að „(tungutækni or málgreining) and íslenska and not enska“ skjöl þar sem fjallað er um tungutækni eða málgreiningu í samhengi við íslensku, án þess að minnst sé á ensku.

Í flestum eða öllum leitarvélum eru orðin sem standa fyrir þessa rökvirkja eingöngu á ensku. Þótt vel mætti hugsa sér leitarvél sem skildi orðin „og“, „eða“ og „ekki“ og meðhöndlaði þau sem rökvirkja, getur það tæplega talist forgangsverkefni. Ensku orðin eru víða notuð í þessum tilgangi, t.d. í mörgum algengum forritunarmálum, og eru margir vanir þeim. Í hugum flestra

eru þau í þessu samhengi frekar einhvers konar tákn en eiginleg orð. Loks má benda á að sé mönnum illa við að nota þessi orð má í mörgum leitarvélum nota sérstök tákn í þeirra stað, t.d. ‘&’ fyrir ‘and’, ‘|’ fyrir ‘or’ og ‘!’ fyrir ‘not’.

Gervigreind

Leitarvélin AltaVista býður nú upp á að í stað leitarorða séu slegnar inn heilar spurningar (dæmi úr auglýsingu AltaVista: „What is the latest news coverage on the Monica Lewinsky scandal?“). Leitarvélin beitir gervigreind til að átta sig á því eftir hvers konar skjölum er óskað og sækir þau. Vissulega væri gott ef unnt væri að spyrja slíkra spurninga á íslensku. Hins vegar er þessi tækni að taka sín fyrstu skref, og enn sem komið er hvergi boðið upp á hana á öðrum málum en ensku. Líklega er ekki tímabært að leggja áherslu á að ganga megi að slíkri þjónustu á íslensku, en rétt er að fylgjast með þróuninni á þessu sviði og reyna að fylgja straumnum þegar fleiri tungumál bætast í hópinn.

6. Stofnanir á sviði tungutækni

Nefndin telur ljóst að grunnvinna á sviði íslenskrar tungutækni verði ekki unnin án atbeina ríkisins. Þar er átt við uppbyggingu viðamikilla gagnasafna á sviði ritaðs og talaðs máls, en slík gagnasöfn eru að allra mati forsenda allrar frekari vinnu, þ.e. smíði einstakra forrita og markaðsvara. Víðast hvar í nágrannalöndum okkar hefur slíkum gagnasöfnum verið komið upp fyrir opinbert fé, bæði frá einstökum ríkisstjórnum og Evrópusambandinu. Vegna smæðar íslensks málsamfélags og þar með markaðar er borin von að ætla að einkafyrirtæki geti risið undir þeim kostnaði sem í þessu felst.

Háskólastofnanir

Nefndin telur sjálfsagt að nýta eftir því sem kostur er þær stofnanir ríkisins sem þegar eru fyrir hendi og tengjast þessu sviði. Þar er einkum um að ræða þrjár stofnanir í tengslum við Háskóla Íslands; Orðabók Háskólans, Íslenska málstöð og Málvísindastofnun Háskólans. Þessar stofnanir eiga þrennt sameiginlegt sem hér skiptir máli. Í fyrsta lagi er hlutverk þeirra skilgreint þannig í lögum og reglugerðum að eðlilegt er að þær eigi hér hlut að máli. Í öðru lagi hafa þær allar yfir að ráða mikilvægum gagnasöfnum sem nýtast í uppbyggingu íslenskrar tungutækni. Í þriðja lagi búa starfsmenn þeirra allra yfir sérþekkingu sem skiptir máli í þessu starfi.

Samkvæmt reglugerð Orðabókarinnar er hún „vísindaleg orðfræðistofnun“ sem hefur m.a. það hlutverk að „sinna hvers kyns rannsóknum á orðaforðanum og þróun hans. Stofnunin gegnir jafnframt þjónustuhlutverki við fræðimenn og almenning“. Orðabókin hefur m.a. yfir að ráða mjög miklu safni tölvutækra íslenskra texta frá síðustu 15 árum. Auk þess hefur hún unnið að samningu íslensks orðabókarstofns sem hefur að geyma grundvallarorðaforða málsins. Þetta tvennt er mjög mikilvægt í uppbyggingu þeirra gagnasafna sem nauðsynlegt er að koma upp.

Íslensk málstöð er rekin af Íslenskri málnefnd í samvinnu við Háskóla Íslands. Samkvæmt lögum er meginhlutverk málnefndarinnar „að vinna að eflingu íslenskrar tungu og varðveislu hennar í ræðu og riti“, og skal hún vera „stjórnvöldum til ráðuneytis um íslenskt mál“. Sé litið á þátttöku stjórnvalda í uppbyggingu íslenskrar tungutækni sem mikilvægan þátt í verndun íslensks máls er því eðlilegt að málnefndin og málstöðin eigi þar hlut að máli. Að auki hefur málstöðin yfir að ráða ýmsum gögnum sem myndu nýtast við uppbyggingu gagnasafna, s.s. orðabanka, safni fyrirspurna um málfarsleg atriði o.fl.

Málvísindastofnun er ein af stofnunum heimspekideildar og er „vísindaleg rannsókna- og fræðslustofnun“ skv. reglugerð, sem „skal annast grundvallarrannsóknir í íslenskum og almennum málvísindum“. Einnig skal hún „safna gögnum (segulbandsupptökum o.fl.) um íslenskt nútímatalmál og varðveita þau“, og auk þess „sinna verkefnum í hagnýtum málvísindum“. Þetta fellur vel að því sem hér er til umræðu, og með þátttöku Málvísindastofnunar fást nauðsynleg tengsl við kennslu- og rannsóknarumhverfi í íslenskri málfræði innan Háskólans, auk þess sem stofnunin ræður yfir miklum gögnum um íslenskt talmál.

Nú er Íslensk málstöð nýflutt að Neshaga 16 þar sem Orðabókin var fyrir. Við það skapast góð skilyrði fyrir auknu samstarfi þessara stofnana. Æskilegt er að þeirri starfsemi sem ríkið stæði fyrir á sviði tungutækni yrði komið fyrir á sama stað og nyti þannig nábýlis við þessar stofnanir, þau gögn sem þar eru vistuð og þá sérþekkingu sem starfsmenn þeirra búa yfir. Á þessu stigi eru ekki gerðar ákveðnar tillögur um formlega stöðu starfsemi í tungutækni eða skipulagsleg tengsl þeirrar starfsemi við umræddar stofnanir. Slíkt verður að skoða nánar og í samvinnu við forráðamenn stofnananna.

Þýðingamiðstöð utanríkisráðuneytisins

Þýðingamiðstöð utanríkisráðuneytisins var stofnuð árið 1990. Tilgangurinn með miðstöðinni var í upphafi að þýða EES-samninginn og gerðir sem heyra undir hann (tilskipanir, reglugerðir og fleira). Frá upphafi var lagt upp með þá hugsun að stefna ætti að því að nýta tölvur eins og kostur væri við að samræma og auðvelda þýðingastarfið. Til að mynda var útbúið forrit sem reiknaði líkindastuðul með skjölum og gerði þannig auðvelt að finna skjöl sem nýta mætti sem fyrirmyndir. Þessu forriti hefur þó ekki verið haldið við, og vinnur það ekki rétt með breytt form sem nú er á frumtextunum. Þá er notað villuleitarforrit og forrit sem leitar að tvíteknum orðum.

Þýðingamiðstöðin býr yfir tölvutækum orðalista, sem er lykiltæki við samræmingu orðaforðans. Í orðalistanum er mikill fjöldi íðorða af ýmsum sviðum, alls hátt í 20.000 færslur. Í hverri færslu er tilgreint íslenskt orð eða frasi og erlend (oftast ensk) samsvörun. Þá er gjarnan gefið dæmi um notkun. Hins vegar eru oftast nær ekki skráðar málfræðilegar upplýsingar.

Framan af var talsvert unnið að því að gera sem mest af þýðingunum sjálfvirkt, einkum þýðingar á ýmsum heitum, föstum frösum og stöðluðum hlutum gerðanna. Þessi viðleitni náði þó aldrei til flóknari þýðinga, og hluti sjálfvirkninnar er ekki lengur notaður. Þótt á síðustu árum hafi ekki verið starfað frekar á þýðingamiðstöðinni að sjálfvirkum þýðingum og gerð annarra hjálpartóla fyrir þýðendur er áhugi þar á slíkum málum mikill. Starfsmönnum miðstöðvarinnar er ljóst að útilokað er að ná fram þeirri samræmingu og stöðlun sem krafist er við þessar þýðingar nema með aðstoð tölvutækninnar.

Staðlaráð Íslands

Frá því um áramót 1992-1993 hefur Staðlaráð Íslands (STRÍ) starfað samkvæmt þeim lögum sem gilda um staðla hér á landi. Hlutverk þess er að vera samstarfsráð þeirra sem áhuga hafa á og hagsmuna hafa að gæta af gerð og notkun staðla hérlendis. Í starfi sínu hefur STRÍ það að leiðarljósi að auka vöxt og nýsköpun íslensks atvinnulífs og bæta starfsskilyrði þess, og bæta auk þess vernd og öryggi neytenda. Viðamesta verkefni Staðlaráðs tengist aðild ráðsins að evrópsku staðlasamtökunum CEN og CENELEC því aðildinni fylgir sú skuldbinding að gera alla staðla sem frá samtökunum koma að íslenskum stöðlum.

STRÍ gefur út séríslenska staðla og hefur umsjón með gerð þeirra. Séríslenskir staðlar eru samdir þegar hagsmunaaðilar telja þörf á því vegna sérstakra aðstæðna hér á landi eða vegna þess að ekki eru til evrópskir eða alþjóðlegir staðlar um tiltekið efni. Í einstaka tilfellum eru staðlar þýddir á íslensku og hefur STRÍ þá umsjón með þeirri vinnu og sér um útgáfuna.

Staðlaráð veitir upplýsingar um hvaðeina er lýtur að stöðlum og stöðlun og sér um sölu staðla frá fjölmörgum staðlastofnunum. Auk þessa er eitt af hlutverkum STRÍ að stuðla að framgangi stöðlunar á Íslandi með fræðslu og kynningum.

7. Áhugaverðar vefsíður

Stefna íslenskra stjórnvalda

- Framtíðarstefna ríkisstjórnarinnar á íslensku.
  URL: www.stjr.is/framt/syn00.htm

- Framtíðarstefna ríkisstjórnarinnar á ensku.
  URL: www.stjr.is/framt/vision00.htm

- Vefsíður RUT nefndar. Þessi vefsíða segir frá stefnumótun ríkisstofnana í upplýsingatækni. Þarna er ekkert um íslensku.
  URL: www.stjr.is/fr/rut/hugb97/gsigurd/index.htm

- Vefsíða úr vinnunni við stefnumótun ríkisstjórnarinnar. Hér er kafli um tunguna og einnig vísað í forstaðalinn FS130.
  URL: eldur.stjr.is/fr/rut/uppsam95.htm#n1u

- Vefsíða verkefnisstjórnar um upplýsingasamfélagið. Hér eru tilvísanir í ýmsar skýrslur.
  URL: brunnur.stjr.is/interpro/for/for.nsf/pages/verk

Tungutækni, almennt

- Survey of the State of the Art in Human Language Technology. Greinasafn um tungutækni, tæknilegt og mjög ítarlegt.
  URL: www.cse.ogi.edu/CSLU/HLTsurvey/

- Sænsk námskeið í tungutækni (språkteknologi).
  URL: www.nada.kth.se/~viggo/sprakteknologi/kursplan.html

- Námskeið Eiríks Rögnvaldssonar prófessors um tölvur og tungumál.
  URL: www.hi.is/~eirikur/ttt.html

Tungutækni og málvísindi

- Speech on the Web, list of pages related to phonetics and speech sciences. Vefföng tengd hljóðfræði og tali.
  URL: fonsg3.let.uva.nl/Other_pages.html

- WWW Indices related to Computational Linguistics.
  URL: www.ims.uni-stuttgart.de/info/Indices.html

- Conference Schedule for Linguists, Translators, Interpreters and Teachers of Languages. Upplýsingar um ráðstefnur fyrir málvísindafólk, þýðendur, kennara og aðra slíka.
  URL: www.clark.net/pub/royfc/confer.html

Opinberar stofnanir og nefndir sem fjalla um mál og málsöfn

- Orðabók Háskólans. URL: www.lexis.hi.is/

- Íslensk málstöð. URL: www.ismal.hi.is/

- Lög um Íslenska málnefnd.
  URL: www.althingi.is/dba-bin/unds.pl?txti=/
  wwwtext/htdocs/lagas/122b/1990002.html&leito=
  Fyrirspurnir%5C0Sv%F6r#word1

Erlendar rannsóknastofnanir

Vefföng nokkurra máltölvunardeilda á Norðurlöndum:

- Universitetet i Stockholm. URL: www.ling.su.se/dali/dali.htm

- Universitetet i Göteborg. URL: svenska.gu.se/sprakdata og
  URL: www.cling.gu.se/

- Uppsala Universitet. URL: stp.ling.uu.se/educa/fudl.html

- Universitetet i Oslo. URL: www.hf.uio.no/ilf/fou/fag/sli.html

- Universitetet i Bergen. URL: www.hf.uib.no/i/LiLi/SLF/undervisning/aktuell-undervisning.html

- University of Helsinki.
  URL: www.ling.helsinki.fi/research/rumlat.html

- Handelshøjskolen i København.
  URL: www.cbs.dk/departments/dl/index.html

- Máltæknalína í Stokkhólmsháskóla, Utbildningsplan för Språkkonsultlinje. URL: www.nordiska.su.se/konsult.htm

Háskólarannsóknir í ýmsum öðrum háskólum:

- Speech at Carnegie Mellon University.
  URL: drum.speech.cs.cmu.edu/speech/

- Massachussets Institute of Technology, MIT.
  URL: www.sls.lcs.mit.edu/sls/

- University of California, Berkeley.
  URL: www.icsi.berkeley.edu/real/speech.html

- Missisippy State University. URL: www.isip.msstate.edu/

- The Circuit Theory and Signal Processing Lab of the Faculté Polytechnique de Mons í Belgíu. URL: tcts.fpms.ac.be/

Samtök og félög, tungutækni og málvísindi

- Safn af vefsíðum um tungutækni (speech technology), framleiðendur, félög og blöð. URL: www.speechtechmag.com/hotlinks.htm

- ELRA European Language Research Association.
  URL: www.icp.grenet.fr/ELRA/home.html

Evrópusambandið (ESB), tungutækni og tungumál

Hér eru nokkur veföng um sem vísa í tækniáætlanir Evrópusambandsins. Evrópusambandið styrkir tungutækni, vélþýðingar og ýmislegt annað á sviði tungutækni og tungumála. Vorið 1999 hefst ný tækniáætlun ESB, 5. rammaáætlunin.

- Telematics Application Programme, official homepage. Sú áætlun í 4. rammaáætlun ESB sem fjallar um tungutækni (Language Engineering, LE). URL: www2.echo.lu/telematics/

- Concord, directory to the Telematics Applications Programme. Listi yfir verkefni áætlunarinnar. URL: www.concord.dcbru.be/

- The Multilingual Information Society (MLIS) Programme. Eitt af verkefnum ESB sem á að stuðla að fjöltyngi.
  URL: www2.echo.lu/mlis/mlishome.html

Fyrirtæki á sviði tungutækni

- Lingsoft ab í Finnlandi þróar og framleiðir ýmsan tungutækni hugbúnað fyrir mörg tungumál, t.d. finnsku, þýsku, rússnesku, norsku. URL: www.lingsoft.fi/sv/

Staðlar

- Staðlaráð Íslands. URL: www.stri.is/

- Innkaupahandbók um upplýsingatækni. Fjármálaráðuneytið. Rit 1998-3. Ritstjóri Jóhann Gunnarsson. URL: www.stjr.is/rut/

- Um sögu íslenskra bókstafa og stafataflna er hér vísað til vefsíðna á ensku www.stri.is/STRI/enska/Iceletters.shtml

- NIST: National Institute of Standards and Technology (USA) interacts with organizations in implementing Speech Processing Evaluations and Benchmark Tests. Bandaríska staðlastofnunin, tungutæknideild. URL: www.nist.gov/speech/online.htm

Stafir, letur, stafasett

- Unicode, stóra stafataflan. URL: www.unicode.org/

- Gunnlaugur SE Briem, leturgerðarmaður, heimasíða.
  URL: rvik.ismennt.is/~briem

Ýmsar greinar og ritsmíðar um íslensku og tungtækni

- Íslensk stafasaga. Þorgeir Sigurðsson. Staðlaráð Íslands.
  URL: www.stri.is/STRI/enska/Iceletters.shtml

- Rúnastaðall. Staðall um að skrifa íslensku með16 rúnastöfum. Þorgeir Sigurðsson. URL: www.stri.is/traoh2.html

- Heimasíða Stefáns Briem. URL: www.ismal.hi.is/stefan/

Tungumál

- Ethnologue, Languages of the World. Vefsíða sem gefur upplýsingar um 6 700 tungumál sem töluð eru í 228 löndum, hve margir tala þau og ýmislegt fleira. URL: www.sil.org/ethnologue/

Vélrænar þýðingar

- Systran hugbúnaður til þýðinga.
  URL: www.systransoft.com/personal.html

Talgervlar

- Á heimasíðu Telia er að finna upplýsingar um talgervla og dæmi sem hægt er að hlusta á. Þar er m.a. íslenskur talgervill sem Blindrafélagið hefur staðið fyrir að gera. Þarna má heyra muninn sem er á þeim tveimur gerðum talgervla sem notaðir eru.
  URL: www.promotor.telia.se/infovox/index.htm

- Á ýmsum stöðum á netinu má finna síður þar sem hægt er að prófa talgervla á ýmsum tungumálum. Dæmi um þetta er hér.
  URL: www.elan.fr/speech/

Talkerfi

Hér eru vefslóðir í talkerfin sem eru til sölu í dag fyrir PC tölvur. Annars vegar eru slóðir framleiðenda en hins vegar í ýmsar umsagnir um kerfin.

Framleiðendur

- Lernout and Hauspie. Framleiða m.a. Voice Xpress.
  URL: www.lhs.com/

- IBM Voice Products. Framleiða m.a. ViaVoice.
  URL: www.software.ibm.com/speech/

- Dragon. Framleiða m.a. NaturallySpeaking.
  URL: www.dragonsys.com/

- Philips. Framleiða m.a. FreeSpeech98.
  URL: www.speech.be.philips.com/

Umsagnir

- What to Look For in a Speech Recognition Program.
  PC Magazine Online.
  URL: www.zdnet.com/pcmag/features/speech98/lookfor.html

- LET'S TALK! Speech technology is the next big thing in computing. Will it put a PC in every home? Business Week , 23.2.1998
  URL: www.sls.lcs.mit.edu/sls/news/article01.html

- Tilvísanir í umsagnir um tungutól í tölvublöðum. Hér má finna umsagnir um helstu forrit sem seld eru nú.
  URL: www.voicerecognition.com/1998/information/in_the_news.html

Alþjóðlegur og fjöltyngdur hugbúnaður

Þegar framleiða skal fjöltyngdan hugbúnað fyrir alþjóðlegan markað þarf að standa rétt að verki frá upphafi. Hér er efni um það.

- Globalisation of Software and User Interfaces. Richard ISHIDA.
  URL: ourworld.compuserve.com/homepages/Richard_Ishida/

- Multilingual Software Digest.
  URL: www.gy.com/home.html#MultilingualSoftware

Málsöfn

- Orðabanki Íslenskrar málstöðvar. URL: www.ismal.hi.is/ob/

- Links to 400 dictionaries. Vefsíður orðabóka af ýmsu tagi.
  URL: www.bucknell.edu/~rbeard/diction.html

- Dæmi um einkaframtak í því aðsetja tækniorð á Netið. Sá sem gerði þetta hefur verið að gefa út kennslubækur um Word og fleira.
  URL: www.prim.is/Ordasafn/

- The Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The University of Pennsylvania (www.upenn.edu/) is the LDC's host institution. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems (www.cise.nsf.gov/iis/index.html) division of the National Science Foundation (www.nsf.gov/).

- European Language Resources Association, ELRA (European Language Resources Association) was established in Luxembourg in February, 1995, with the goal of founding an organization to promote the creation, verification, and distribution of language resources in Europe. A non-profit organization, ELRA aims to serve as a focal point for information related to language resources in Europe. It will collect, market, distribute, and license European language resources. ELRA will help users and developers of language resources, government agencies, and other interested parties exploit language resources for a wide variety of uses. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world.
  URL: www.icp.grenet.fr/ELRA/home.html

- LDC related sites Mikið af vefsíðum tengdum mál- og textasöfnum.
  URL: www.ldc.upenn.edu/ldc/sites/index.html

- Following is the list of all the hyperlinks from the comp.speech FAQ. This is probably the biggest list of speech technology links available. The links are provided to WWW references, ftp sites, and newsgroups. Cross-references to the comp.speech WWW pages are also provided. Mikið safn af vefsíðum fyrirtækja sem framleiða hljóðtól, rannsóknastofnana o.s.frv.
  URL: www.speech.cs.cmu.edu/comp.speech/index.html

- The European Association for Machine Translation (EAMT) is an organization that serves the growing community of people interested in MT and translation tools, including users, developers, and researchers of this increasingly viable technology. The EAMT is one of three regional associations of the International Association for Machine Translation (IAMT), which counts a increasing number of members worldwide. URL: www.lim.nl/eamt/

- Electronic Publishing, R&D news and resources. ESB vefblað um slíka tækni. URL: inf2.pira.co.uk/

- The International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques for Speech Input/Output, COCOSDA, has been established to encourage and promote international interaction and cooperation in the foundation areas of Spoken Language Processing.
  URL: www.itl.atr.co.jp/cocosda/

Vandamál íslensks leturs efa ég að leysist sjálfkrafa.

Fyrir rúmum áratug fengu stafirnir Þþ Ðð og Ýý oft að fljóta með. Mér var sagt að það væri þrýstingi frá NATO að þakka; tókst aldrei að fá það staðfest. Sá þrýstingur dugir að minnsta kosti ekki lengur. Þeim leturframleiðendum sem selja íslenska stafi fer fækkandi. Samkeppnin er hörð og íslenski markaðurinn ekki ómaksins verður.

Margir leturhönnuðir misskilja stafina þ og ð. Þeir verða oft ónothæfir. (Ég setti leiðbeiningar á vefinn en ekki er fólk skyldugt til að fara eftir þeim.)

Mér kæmi ekki á óvart að Tyrkir sigruðu í togstreitunni um sess í stöðlum. Þeir eru núna að gera það sem Íslendingar ættu að gera: vinna framleiðendur á sitt band.

Hvað er til ráða?

Íslendingar geta komið sér vel við leturframleiðendur. Þeir fást trúlega til að hafa íslenska stafi á boðstólum ef þeir fá þá tilbúna og ókeypis. (Ég tek fram að ekki er ég í verkefnaleit.) Til lengdar dugir ekki að berja í borðið. Gagnlegra væri að afla sér bandamanna.

Íslendingar ættu að vingast við Tyrki og styðja þá í kröfum um staðalnúmer sem okkur er ekki sárt um.

Þetta er hvort tveggja augljóst. Hitt er líka sýnt að Íslendingum stendur flestum á sama. Ég reyndi lengi að gæta hagsmuna okkar, sérstaklega meðal kunningja minna í Association Typographique Internationale. Áhuginn hérlendis var minni en hjá útlendingum.

Ég vona að þetta svari því helsta sem um var spurt.

Hjartanlega. Gunnlaugur

Gunnlaugur SE Briem

briem@ismennt.is
http://rvik.ismennt.is/~briem

Phone 44 (UK) 181 749 2919
Phone 1 (US) 415 771 3212
Phone 354 (Iceland) 565 7198

Microsoft in War of Words

By MARY WILLIAMS WALSH,

Times Staff Writer

REYKJAVIK, Iceland—You think the Justice Department has it in for Bill Gates and the marketers of Microsoft Corp.? Try an earful from the Icelandic Language Institute. „They are nothing less than destroying what has been built up here for ages," says the institute's director, Ari Pall Kristinsson. Iceland, you must first understand, is a tough, proud island nation with language-preservation instincts that put the Academie Francaise to shame. Icelandic may be spoken by fewer than half a million people worldwide, but you should never mistake it for a beleaguered minority tongue. On the contrary: Up until now, Iceland could boast a minority-language success story extremely unusual in the world. The French may be fighting a losing battle with such creeping barbarisms as le hot dog; Germany may have succumbed to das midlifecrisis. But centuries of Icelandic isolation and vigilance have preserved a national grammar, vocabulary and spelling that are virtually identical to what the Vikings spoke when they settled this land in the 9th century. Startling though it may sound to an American who has struggled with the Middle English of Chaucer, any Olof Sixpack here can curl up with a saga—written a good century before „The Canterbury Tales"—and understand every word.

* * *

Today, however, Iceland's linguistic patriots say Gates stands poised to lay waste to all they hold dear. The reason, they say, has everything to do with the shamanistic powers computers seem to exercise over the minds of the young and with the marketing strategies of far-away Microsoft. Microsoft's sin: It refuses to translate Windows into Icelandic. Spokeswoman Erin Brewer notes that while the company has translated the popular program into „at least 30 languages," including such rarities as Slovenian and Catalan, it won't be doing Icelandic. „We are not localizing Windows 98 into Icelandic due to the size of the market," she said. Thus, Iceland's unique linguistic success may now prove its undoing. For even as its language specialists were defending the purity of their ancestral tongue, they were also making sure every schoolchild here learned English. With the entire population now proficient in English as a second language, Microsoft sees no point in translating Windows into their proud mother tongue; it can just sell them the English version. „As it looks now, Microsoft is the most powerful company in the world, and it can decide which way the computer world is heading," Kristinsson says. „This is a disaster. You cannot implement a language policy if the computer talks to you in some other language." To appreciate this sad, new Icelandic saga, you have to understand Iceland's linguistic achievements to date. In the early 1960s, the age of Sputnik and the transistor radio, all of the Nordic countries—Iceland, Norway, Sweden, Finland and Denmark—began setting up national language councils. Torrents of new products and ideas were washing out over the globe, and these countries wanted to set national policies for name creation. Iceland established its language council in 1964.

In many places, the English names for new inventions and processes aresimply incorporated into the language, as when the P.A. system in a German airport tells you to bring das ticket to dem check-in. But not in

Iceland. „It seems to us to be a very practical thing, not to absorb foreign words for new objects, but to make new words for things as they come up," Kristinsson says. So: Let the research labs of the world come forward with their hyperlinks and motherboards, their fuzzy-set logic and their geosynchronous satellite positioning systems. Up until now, Iceland's linguists have kept pace with them, creating perfectly pedigreed Icelandic words for anything new. An example: No self-respecting Icelander would think of arguing that this name game isn't worth the effort—that, say, AIDS should just be called AIDS, rather than alnaemi, an ancient Icelandic word for „totally vulnerable." And thus, a video monitor here is a skjar, which literally means „the amniotic sac of a calf." Generations ago, when Icelanders lived in sod houses, these membranes were dried and stretched across holes in the earthen walls for windows. Even today, when windows are made of glass, skjar still evokes the idea of a window. And since the centuries-old term had fallen into disuse, it was free for the taking and recycling by computer wonks. The etymology of the Icelandic word for computer, toelva, is similarly pure: It is a compound word, put together from the Icelandic words meaning „digit" and „prophetess," alluding to a computer's great store of knowledge. „You can say everything in Icelandic," says Kristjan Arnason, professor of Icelandic at the University of Iceland. „You don't need English to express yourself." Kristinsson adds: „They start teaching computers in kindergarten, and there's no way they would call these things anything but skjar and toelva. They're just words, like 'car,' or 'cup.' You don't have to be filled with national pride, or anything like that, to use them." Not that Icelanders are short of national pride, of course, but it's really logic and efficiency that fuel their crusade, say the specialists: By constantly making up new indigenous words for global concepts, Iceland has neatly avoided the costly language battles now plaguing other countries in the instant-communication age. All across industrialized Europe, schools, ministries of culture, writers and other leading linguistic lights are mired in debate over new foreign words and what to do about them. Consider Norway: Its language council made the mistake of letting foreign words worm their way in, and now its language is hopelessly cluttered with interlopers such as „entertainer," „fight," „insider" and „champagne." No one can agree on how to spell them in Norwegian, or what genders they should be assigned—an important point in the Germanic languages. Norwegian schoolteachers throw up their hands at the thought of teaching spelling anymore.

* * *

To the south, Denmark, a country that normally prides itself on unfettered free expression, had to pass a law last year forcing foot-dragging schools to comply with its new official spellings. How nice to be here in Iceland, above the fray. „As you know, Norwegian used to be the same language as ours," Kristinsson says with a smile. „Theirs has changed. Not ours." Thus, until recently, Kristinsson was the toast of the international linguistics crowd, holding his head high at professional meetings haunted by doleful Quebeckers, militant Basques and worried Russian delegates seeking ways of defusing the next language-based conflagration on their territory. But now comes Windows. Iceland can't avoid computers; on the contrary, because it is in the middle of the North Atlantic, far from any continent, it needs e-mail and the Internet just to function in modern times. Iceland has worked hard to promote a computer-literate society. „It's a very big danger," says Arnason, „because schoolchildren need computers, and the language of computers soon becomes the language of the kitchen." A few years back, Apple Computer Inc. spotted a business opportunity in Iceland's fear of electronic

English infiltration. It translated its software into Icelandic and mounted a marketing campaign on the theme of minority-language protection. Kristinsson has one of Apple's promotional posters on his office wall. It features a list of bastard words that an Icelandic kid might pick up via English-language Windows, with the legend, „What kind of child do you want to have?" Kristinsson heartily approves—but he also knows the placation of cultural preservationists isn't what sells computers these days. „I'm afraid the Microsoft systems are overwhelmingly used here," he says. Unable to stop the influx of Windows, Iceland's cultural authorities began petitioning software importers, asking for the right to translate Windows into Icelandic. That proposal went nowhere, Arnason says, because the programs can't be translated without the translator's going into the main operating system, something Microsoft won't allow. Iceland then offered to pay Microsoft to do the translation itself, but Microsoft refused to quote Iceland a price. „The Microsoft people say we have to do it, but we're not allowed to do it," Arnason says. „It's a—a what do you call it?—a Catch-22." Last fall, Iceland's minister of culture bypassed the importers and wrote directly to the Redmond, Wash., headquarters of the software giant, warning Microsoft that if a translation wasn't forthcoming, this country would find other ways to computerize its schools. That at least elicited a letter, saying that Microsoft wouldn't translate Windows 95, but it might translate Windows 98. Since then, nothing. So now, Iceland is bringing in the heavy artillery: President Olafur Grimsson himself is about to join the campaign. Details of his spin offensive are still under wraps, but the push is expected to coincide with other U.S.-Nordic friendship gestures scheduled for the turn of the millennium. Will it be enough to change policy at a company that has already shown itself willing to take on the Justice Department and 20 American states? „I am not pessimistic in any other area, but we have no control over this," Kristinsson says. „Bill Gates doesn't even listen to Bill Clinton, so I don't think he will listen to us."

Menntamálaráðuneytið