Eiríkur Rögnvaldsson

Evrópskt máltæknisamstarf

Erindi flutt í Ríkisútvarpinu 25. október 2011

Um þessar mundir tekur Ísland þátt í víðtæku samstarf flestra Evrópuþjóða á sviði máltækni, META-NET. Innan META-NET eru svæðisbundin samstarfsnet, þar á meðal META-NORD sem öll Norðurlönd og Eystrasaltslönd eru aðilar að. Íslenski þátttakandinn í samstarfinu er Máltæknisetur, sem er samstarfsvettvangur Háskóla Íslands, Háskólans í Reykjavík og Stofnunar Árna Magnússonar í íslenskum fræðum um rannsóknir, kennslu og þróunarverkefni á sviði máltækni. META-NORD er tveggja ára verkefni sem hófst 1. febrúar sl. og er kostað af stefnumótunaráætlun Evrópusambandsins á sviði upplýsingatækni. Heildarstyrkur til verkefnisins er 2.250 þúsund evrur eða um 360 milljónir króna á núverandi gengi, en þar af er hlutur Máltækniseturs um 32 milljónir króna.

Máltækni felst í samþættingu tungumáls og tölvutækni og sú samþætting hefur ýmsar hliðar. Þannig má nota tölvutæknina til að hjálpa okkur á ýmsan hátt við meðferð og beitingu tungumálsins. Þar má nefna hugbúnað sem leiðréttir stafsetningu og málfar, svokallaða stafrýna og málrýna; vélrænar þýðingar, rafrænar orðabækur, kennsluhugbúnað og ýmislegt fleira. Á hinn bóginn má einnig snúa þessu við og nota tungumálið til að auðvelda hagnýtingu og beislun tækninnar. Þannig færist í vöxt að mannlegt mál sé notað til að stjórna ýmsum tölvustýrðum tækjum, allt frá heimilistækjum til bíla.

Í ýmiss konar þjónustuverum og upplýsingaveitum erlendis er nú algengt að tölva tengd talgreini túlki fyrirspurn notanda, leiti svars við henni í gagnabanka, semji svar og komi því til notandans með aðstoð talgervils. Möguleikar máltækni til hjálpar fólki með ýmiss konar fötlun eru einnig gífurlegir. Blindir, sjónskertir og lesblindir hafa t.d. ómetanlegt gagn af talgervlum, og talgreining er himnasending fyrir hreyfihamlaða sem eiga erfitt með að nota hendurnar til að stjórna tækjum eða slá inn texta.

Þróun máltæknibúnaðar er mjög dýr og kostnaðurinn er óháður fjölda málnotenda. Því er skiljanlegt að máltækni hafi átt erfitt uppdráttar á Íslandi – smæð markaðarins gerir það að verkum að fyrirtæki sjá sér ekki hag í því að leggja í mikinn þróunarkostnað sem lítil von er um að ná til baka. Það hefur hins vegar sýnt sig, bæði hér á landi og erlendis, að þekkingu og skilningi á máltækni og möguleikum hennar er verulega ábótavant, bæði hjá almenningi, fyrirtækjum og opinberum aðilum. Með vitundarvakningu á þessu sviði kæmu kröfur um íslenskan máltæknibúnað án efa frá fleiri hópum og yrðu bornar fram með meiri þunga. Þar með ykjust möguleikar á arðvænlegu þróunarstarfi á þessu sviði og jafnframt líkur á að íslensk fyrirtæki færu að sinna því.

Það er enginn vafi á því að þörf íslenskra málnotenda og íslensks málsamfélags fyrir máltækni er jafnmikil og gerist í öðrum tæknivæddum nútímasamfélögum. Verði henni ekki mætt með íslenskum máltæknibúnaði eru líkur á að enskan ryðji sér til rúms á sífellt fleiri sviðum daglegs lífs. Þess vegna er lögð áhersla á það í íslenskri málstefnu sem Alþingi samþykkti fyrir rúmum tveimur árum að mikilvægt sé að vinna að þróun íslensks máltæknibúnaðar.

En hvers vegna veitir Evrópusambandið stórfé til að styrkja máltækni? Fyrir því eru beinharðar efnahagslegar ástæður. Á síðustu 60 árum hefur Evrópa orðið að afmarkaðri pólitískri og efnahagslegri heild, en menningarlega og mállega er hún enn mjög fjölbreytt. Þetta þýðir að milli portúgölsku og pólsku, ítölsku og íslensku eru tungumálaþröskuldar sem torvelda dagleg samskipti milli Evrópubúa sem og samvinnu á sviði viðskipta og stjórnmála. Stofnanir Evrópusambandsins verja um milljarði evra á ári til að viðhalda stefnu sinni um fjöltyngi, þ.e. í að þýða texta og túlka tal. En þarf þetta að vera slík byrði? Nútíma máltækni og málfræðirannsóknir geta lagt mikið af mörkum til að lækka þessa tungumálaþröskulda. Með tengingu við vitræn tæki og búnað mun máltækni í framtíðinni geta gert Evrópubúum kleift að tala saman og eiga viðskipti jafnvel þótt þeir tali ekki sameiginlegt tungumál.

Markmið META-NET og META-NORD er að skapa tæknilegar forsendur fyrir margmála upplýsingasamfélagi í Evrópu þar sem allir geti notað móðurmál sitt við öflun og úrvinnslu hvers kyns upplýsinga. Þetta á að gera með því að efla máltækni fyrir allar þjóðtungur álfunnar og auðvelda tengsl milli þeirra með uppbyggingu margmála málfanga, s.s. textasafna, orðasafna og hugbúnaðar, sem nýst geti í margvíslegum máltækniverkefnum. Ekki er ætlunin að koma slíkum málföngum upp frá grunni, heldur ljúka við verk sem eru í vinnslu, staðla þau og gera aðgengileg á netinu. Með því að greiða leið milli tungumála og auðvelda mönnum að nota móðurmál sitt í fjölþjóðlegum samskiptum má koma í veg fyrir að enskan þrengi sér smátt og smátt inn á fleiri svið á kostnað þjóðtungna en varðveita þess í stað margmála evrópskt samfélag.

Hin hefðbundna leið til að komast yfir tungumálaþröskulda er að læra erlend mál. En án tæknilegs stuðnings er vald á um 30 opinberum málum Evrópuríkja og yfir 50 öðrum tungumálum álfunnar óyfirstíganleg hindrun fyrir Evrópubúa, sem og fyrir efnahag álfunnar, stjórnmálaumræðu og framfarir í vísindum. Lausnin er sú að koma upp stuðningstækni á borð við máltækni. Það mun verða allri Evrópu til mikilla hagsbóta, ekki aðeins á hinum sameiginlega evrópska markaði heldur einnig í viðskiptum við önnur lönd. Til að ná þessu marki og varðveita jafnframt menningarlega og mállega fjölbreytni Evrópu er nauðsynlegt að gera kerfisbundna greiningu á möllegum sérkennum allra Evrópumála, svo og á máltæknilegum stuðningi við þau.

Miklar líkur eru á því að í framtíðinni munu byltingar í samskiptatækni skapa nýja tegund tengsla milli fólks sem talar mismunandi tungumál. Þetta setur aukinn þrýsting á fólk að læra ný tungumál og þó sérstaklega á hönnuði að búa til nýjan tæknibúnað sem tryggi gagnkvæman skilning og aðgang að deilanlegri þekkingu. Hin u.þ.b. 80 tungumál Evrópu eru ein ríkulegustu og mikilvægustu menningarverðmæti álfunnar og grundvallarþáttur í samfélagsgerð hennar en ef ekkert væri að gert. gætu mörg evrópsk tungumál orðið gagnslítil í netvæddu samfélagi framtíðarinnar. Slík þróun myndi veikja alþjóðlega stöðu Evrópu og stangast á við markmið um samfélagsþátttöku allra Evrópubúa á jafnréttisgrundvelli, óháð tungumáli. Í nýlegri skýrslu UNESCO um fjöltyngi er lögð áhersla á að tungumál séu ómissandi tæki til þess að njóta grundvallarmannréttinda, svo sem tjáningarfrelsis, menntunar og þátttöku í samfélaginu.

Áður fyrr beindust aðgerðir til að vernda og varðveita tungumál einkum að tungumálakennslu og þýðingum. Giskað hefur verið á að evrópski markaðurinn á sviði þýðinga, túlkunar, staðfærslu hugbúnaðar og alþjóðavæðingar vefsetra hafi velt 8,4

milljörðum evra árið 2008 og er talinn munu vaxa um tíu prósent á ári. Samt sem áður fullnægir þessi upphæð einungis litlum hluta núverandi þarfar og framtíðarþarfa fyrir samskipti milli tungumála. Augljósasta aðferðin til að tryggja notkun allra Evrópumála í samfélagi framtíðarinnar er að nota viðeigandi tækni, rétt eins og við notum tæknina til að leysa þarfir okkar í samgöngum, orku og stuðningi við fatlaða, svo að eitthvað sé nefnt. Upplýsingaþjónusta í farsíma, hugbúnaður fyrir tölvustutt tungumálanám, fjarnámsumhverfi, sjálfsmatstól og forrit til að uppgötva ritstuld eru dæmi um svið þar sem máltækni getur leikið mikilvægt hlutverk.

Í máltækni felast einnig gífurlegir möguleikar fyrir evrópskt samstarf því að hún getur hjálpað okkur að takast á við hið flókna málumhverfi í álfunni. Til að öðlast yfirsýn yfir stöðu máltækni í Evrópu hefur META-NET beitt sér fyrir gerð skýrslna, svokallaðra hvítbóka, um 30 Evrópumál. Þar er gerð grein fyrir málsamfélaginu og hlutverki málsins í því; máltæknirannsóknum og máltækniiðnaði í landinu; hlutverki máltækniafurða og máltækniþjónustu í landinu; og lagalegum atriðum varðandi máltækni, s.s. höfundarréttarmálum. Skýrslurnar verða bæði á ensku og því máli sem hver skýrsla fjallar um. Skýrslan um íslensku er nú tilbúin og aðgengileg, bæði á íslensku og ensku, á vefsetri META-NORD: vefir.hi.is/metanord. Í byrjun næsta árs verða skýrslurnar síðan prentaðar og þeim dreift þannig.

Hver skýrsla skiptist í þrjá meginkafla. Fyrsti kaflinn er sameiginlegur öllum skýrslunum og gerir grein fyrir mikilvægi máltækni fyrir upplýsingaþjóðfélag nútímans og framtíðarinnar. Annar kafli er sérstakur fyrir hvert tungumál og lýsir málsamfélaginu, sérkennum tungumálsins, nýlegri þróun þess, málrækt, stöðu þess í menntakerfinu, alþjóðlegri stöðu þess, o.fl. Í þriðja kafla er gerð grein fyrir helstu tegundum máltæknibúnaðar og stöðu tungumálsins hvað varðar þann búnað. Í lok kaflans eru töflur þar sem reynt er að meta málföng og máltæknibúnað sem til er fyrir málið.

Meginniðurstöður fyrir íslensku eru þær að íslenska standi þokkalega hvað varðar einföldustu grunnforsendur máltækninnar í búnaði og málföngum, svo sem textagreiningu og málheildum. Einnig eru til einstöku gögn og búnaður með takmarkaða virkni á sviðum eins og talgervingu, talkennslum, vélrænum þýðingum, talmálsheildum, hliðstæðum málheildum og orðagögnum. Háþróaður máltæknibúnaður og málföng, svo sem til textatúlkunar og málmyndunar, er ekki til. Því er ljóst að mikið starf er óunnið við að tryggja framtíð íslenskunnar sem fullgilds þátttakanda í evrópsku upplýsingasamfélagi nútímans – og framtíðarinnar.

Í lok skýrslnanna er að finna samanburð á stöðunni í þeim 30 tungumálum sem þær taka til. Þar eru skoðuð fjögur svið; talvinnsla, vélrænar þýðingar, textagreining og málföng. Tungumálum er skipt í 4-5 klasa fyrir hvert svið, eftir því hversu vel þau eru stödd á sviði máltækni fyrir það svið. Í þessum samanburði kemur fram geysimikill innbyrðis munur á Evrópumálum. Þótt ágætur hugbúnaður og málföng sé til fyrir sum lönd og verksvið eru grundvallareyður á þessum sviðum í öðrum málum – einkum þeim sem töluð eru í litlum málsamfélögum. Mörg tungumál skortir grunntækni til textagreiningar og nauðsynleg málföng til að þróa slíka tækni. Önnur hafa grundvallarbúnað og málföng en hafa ekki burði til að ráðast í merkingarlega vinnslu.

Íslenska er í lægsta klasanum hvað varðar öll tól og málföng sem um ræðir. Hún er þar á sömu slóðum og önnur tungumál sem fáir tala, svo sem írska, lettneska, litháíska og maltneska. Athyglisvert er að eistneska, sem aðeins um milljón manna hefur að

móðurmáli, stendur aðeins betur á sumum sviðum, enda stóð ríkisstjórnin þar fyrir verulegu átaki til að efla eistneska máltækni. Öll þessi tungumál eru langt að baki stórþjóðamálum eins og t.d. þýsku og frönsku. En jafnvel málföng og máltæknitól fyrir þau tungumál ná hvorki sömu gæðum né yfirgripi og hliðstæð föng og tól fyrir ensku, sem er í fararbroddi á nær öllum sviðum máltækninnar. Þó eru enn fjölmargar eyður í enskum málföngum hvað varðar hágæða búnað.

Um síðustu aldamót var íslensk máltækni varla til. Þetta breyttist eftir 1999, þegar sérstakur starfshópur skilaði skýrslu um máltækni til menntamálaráðherra. Í þessari skýrslu voru gerðar tillögur um ýmsar aðgerðir til að koma íslenskri máltækni á laggirnar. Starfshópurinn áætlaði að það myndi kosta u.þ.b. einn milljarð króna að gera íslenska máltækni sjálfbæra. Þegar því marki væri náð ætti markaðurinn að geta tekið við þar eð hann hefði aðgang að opnum málföngum sem hefði verið komið upp á vegum máltækniáætlunar ríkisstjórnarinnar og yrðu afhent á jafnréttisgrundvelli til allra sem hygðust nýta þau í markaðsvörum.

Það verður að benda á að heildarfjármagnið sem veitt var til máltækniáætlunarinnar frá 2000-2004 var aðeins um 1/8 af þeirri upphæð sem áðurnefndur starfshópur taldi að þyrfti til. Það þarf því ekki að koma á óvart að íslensk máltækni er enn á bernskuskeiði. 330.000 málnotendur eru ekki nægilegur fjöldi til að standa undir kostnaðarsamri þróun á nýjum vörum. Um þessar mundir vinna nánast engin íslensk fyrirtæki að máltækni vegna þess að þau sjá enga hagnaðarvon í henni. Því er ákaflega mikilvægt að halda áfram opinberum stuðningi við íslenska máltækni enn um sinn.

Fyrir lítið málsamfélag og lítið rannsóknarumhverfi eins og það íslenska er samvinna lífsnauðsyn – ekki bara innanlands heldur einnig alþjóðleg. Þess er að vænta að þátttaka Íslands í META-NORD og META-NET muni gera mögulegt að þróa, staðla og gera aðgengileg ýmis mikilvæg málföng og stuðla þannig að vexti og viðgangi íslenskrar máltækni. Langtímamarkmið META-NET er að innleiða hágæða máltækni fyrir öll tungumál þannig að menningarleg fjölbreytni stuðli að eflingu pólitískrar og efnahagslegrar einingar. Tæknin mun brjóta múra milli evrópskra tungumála og reisa brýr milli þeirra í staðinn. Þetta krefst þess að allir hagsmunaaðilar – í stjórnmálum, rannsóknum, viðskiptum, og samfélaginu öllu – sameini krafta sína í þágu framtíðar.