Nefnd um notkun íslensku í stafrænni upplýsingatækni

1. Skipun

Með bréfi dagsettu 9. september 2014 skipaði mennta- og menningarmálaráðherra nefnd um notkun íslensku í stafrænni upplýsingatækni. Í skipunarbréfinu segir að nefndin eigi „[...] að gera áætlun um aðgerðir er miði að því að gera íslensku gjaldgenga í stafrænni upplýsingatækni og stuðla að notkun hennar á þeim vettvangi. Áætlunin feli í sér tímasett yfirlit um aðgerðir og áfanga, kostnaðarmat og fjármögnun.“

Í nefndina voru skipuð Eiríkur Rögnvaldsson, prófessor við Háskóla Íslands, Sigrún Helgadóttir, verkefnisstjóri hjá Stofnun Árna Magnússonar í íslenskum fræðum, og Hrafn Loftsson, dósent við Háskólann í Reykjavík, formaður nefndarinnar. Mælst var til að nefndin legði fram áætlun sína í síðasta lagi 1. janúar 2015.

2. Samantekt

Árið 2009 samþykkti Alþingi tillögur Íslenskrar málnefndar að íslenskri málstefnu, *Íslenska til alls*, sem opinbera stefnu í málefnum íslenskrar tungu. Í málstefnunni er m.a. lögð áhersla á að „íslensk tunga verði nothæf – og notuð – á öllum þeim sviðum upplýsingatækninnar sem varða daglegt líf alls almennings“. Til að svo megi verða þarf að gera verulegt átak sem ekki þolir bið. Nauðsynleg gögn og búnaður verða ekki til af sjálfu sér og vegna smæðar markaðarins er útilokað að fyrirtæki leggi í þann mikla kostnað sem felst í uppbyggingu íslenskrar máltækni. Einnig er mikilvægt að breyta lögum um notkun höfundarréttarvarins stafræns efnis til að auðvelda rannsóknir sem ekki eru gerðar í hagnaðarskyni.

Nefndin leggur til að fjárfest verði í íslenskri máltækni með sérstakri langtímaáætlun til 10 ára sem styrki bæði doktorsnema og einstök tækniþróunar- og innviðaverkefni og fylgi þannig fordæmi nokkurra Evrópuþjóða sem við berum okkur gjarnan saman við. Nefndin áætlar að það þurfi um einn milljarð króna til að byggja nauðsynlegan grunn þannig að í lok áætlunarinnar verði íslenskan komin í flokk nágrannatungumála þegar litið er til stuðnings við máltækni. Þetta er vissulega mikið fé, en nefndin telur að á þessu sviði eigi þjóðin ekkert val, sé raunverulegur vilji til þess að gera Íslendingum kleift að halda áfram að nota íslensku á öllum sviðum þjóðlífsins.

Rökstuðningur fyrir þessum tillögum fer hér á eftir, ásamt nánari útfærslu þeirra. Í 3. kafla er saga íslenskrar máltækni stuttlega rakin, í 4. kafla eru tekin dæmi af markáætlunum grannþjóða á sviði máltækni, í 5. kafla er gerð grein fyrir þeim verkefnum sem nefndin telur að eigi að njóta forgangs, og í 6. kafla eru settar fram tillögur nefndarinnar um aðgerðir, ásamt grófri tímaáætlun og kostnaðarmati.

3. Íslensk máltækni

Máltækni er þýðing á enska heitinu *Language Technology*. Orðið vísar til samvinnu tungumáls og tölvutækni í hagnýtum tilgangi, samvinnu sem beinist að því að þróa kerfi sem geta unnið með og skilið náttúruleg tungumál og stuðlað að notkun þeirra í samskiptum manns og tölvu. Þessi samvinna getur bæði falist í notkun tölvutækninnar í þágu tungumálsins og í notkun tungumálsins innan tölvutækninnar.

Upphaf máltækni á Íslandi

Máltækni er ung grein á Íslandi. Haustið 1998 skipaði menntamálaráðherra starfshóp til að gera úttekt á stöðu máltækni á Íslandi og gera tillögur um eflingu greinarinnar. Niðurstaða starfshópsins var sett fram í skýrslunni *Tungutækni – skýrsla starfshóps*[1] sem gefin var út í apríl 1999. Í kjölfar skýrslunnar setti menntamálaráðherra af stað svokallaða „tungutækniáætlun“ til að styrkja stofnanir og fyrirtæki til að byggja upp grunngögn og búnað fyrir máltækni. Til verkefnisins var varið 133 m.kr. á árunum 2000–2004 en það var einungis 1/8 af því fjármagni sem starfshópurinn taldi þurfa (u.þ.b. milljarður). Fyrir þetta fé voru þó byggð upp mikilvæg gagnasöfn sem nýtast í áframhaldandi máltæknistarfi en vegna þess að ekkert framhald varð á fjárveitingum hefur einungis að mjög litlu leyti verið þróaður máltæknibúnaður út frá þessum gögnum. Meðal gagnasafna sem voru unnin fyrir styrk úr tungutækniáætluninni má nefna fyrsta áfanga *Beygingarlýsingar íslensks nútímamáls* (BÍN[2]) og *Markaða íslenska málheild* (MÍM[3]) sem er safn um 25 milljón orða úr margvíslegum textum sem eru greindir málfræðilega. Eitt af erfiðustu og tímafrekustu verkþáttum þess verkefnis var að afla leyfa frá rétthöfum höfundarréttarvarinna texta til þess að hafa þá í málheildinni.

Máltæknisetur

Þegar tungutækniáætluninni lauk var ákveðið að koma á fót samstarfi þeirra stofnana sem mest unnu við rannsóknir, þróun og kennslu í máltækni. Vettvangurinn kallast *Máltæknisetur*[4] og er samstarfsvettvangur þriggja stofnana um rannsóknir, þróun og kennslu í máltækni: Málvísindastofnunar Háskóla Íslands (HÍ), tölvunarfræðideildar Háskólans í Reykjavík (HR) og orðfræðisviðs Stofnunar Árna Magnússonar í íslenskum fræðum (áður Orðabók Háskólans). Þegar setrið var sett á fót árið 2005 var gert ráð fyrir því að það fengi afgang af fjárveitingu tungutækniáætlunarinnar en þeir fjármunir voru uppurnir þegar til kom. Vísindamenn Máltækniseturs (sem nefndarmenn tilheyra) hafa komið að vinnu við langflest rannsóknar- og þróunarverkefni á sviði íslenskrar máltækni frá 2005–2014. Styrkir hafa fengist fyrir nokkur þessara verkefna, t.d. hjá Rannsóknasjóði fyrir verkefni sem tengjast málfræðilegri mörkun og þáttun texta. Máltæknisetur hefur annars ekki fengið neinar sérstakar fjárveitingar en starfsemi þess byggist á vinnu starfsmanna aðildarstofnana. Einnig má nefna að vísindamenn sem tengjast Máltæknisetri tóku þátt í gerð nýs talgervils sem Blindrafélagið hafði forgöngu um að gerður væri (Karl og Dóra[5]) og nýs talgreinis í samstarfi við fyrirtækið Google. Í talgreinisverkefninu stóðu Máltæknisetur og tækni- og verkfræðideild Háskólans í Reykjavík fyrir söfnun raddsýna sem Google notaði til þess að gera talgreini sem má nota í snjalltækjum sem byggja á Android-stýrikerfinu og í Google-leitarvélinni. Raddsýnin eru opin öðrum sem vilja búa til talgreini en Google á talgreininn sjálfan.

Mörg þeirra máltækniverkefna sem Máltæknisetur hefur unnið að eru eðli málsins samkvæmt fyrst og fremst hagnýt fremur en eiginleg rannsóknarverkefni og því er erfitt að sækja um

styrk til þeirra hjá Rannsóknasjóði. En þrátt fyrir ótvírætt hagnýtt gildi verkefnanna gerir smæð markaðarins það að verkum að erfitt er að sýna fram á fjárhagslegan ávinning af þeim og það vegur þungt í mati umsókna hjá Tækniþróunarsjóði. Einn meginvandi þeirra sem sinna íslenskri máltækni er því sá að þróunarverkefni af því tagi sem brýnast er að sinna falla illa að íslenska sjóðakerfinu – þau eru ekki nógu vísindaleg fyrir Rannsóknasjóð og ekki nógu arðvænleg fyrir Tækniþróunarsjóð.

Íslensk málstefna

Hinn 12. mars 2009 samþykkti Alþingi tillögur Íslenskrar málnefndar að íslenskri málstefnu, *Íslenska til alls*[6] sem opinbera stefnu í málefnum íslenskrar tungu. Í málstefnunni er m.a. lögð áhersla á að „íslensk tunga verði nothæf – og notuð – á öllum þeim sviðum upplýsingatækninnar sem varða daglegt líf alls almennings“. Í málstefnunni kemur jafnframt eftirfarandi fram:

> *Það kostar jafnmikið að byggja upp málleg gagnasöfn og máltæknibúnað fyrir tungumál sem 300 þúsund manns tala og fyrir tungumál milljónaþjóða og því er ekki von að fyrirtæki sjái sér hag í því að leggja í mikinn kostnað við að þróa og aðlaga slíkan búnað fyrir íslensku. Að óbreyttu munum við því dragast hægt en örugglega aftur úr á þessu sviði og þar sem tölvutæknin sækir ört á í umhverfi okkar má búast við að enskan yfirtaki fleiri og fleiri þætti daglegs lífs. Þá getur verið skammt í að íslenskan verði eingöngu heimilismál sem unga kynslóðin sér ekki tilgang í að læra almennilega vegna þess að hún er ekki nothæf í nýrri tækni og öðru sem er nýtt og spennandi, á sviðum þar sem nýsköpun af ýmsu tagi á sér stað eða þar sem ný atvinnutækifæri bjóðast.*

Í málstefnunni eru eftirfarandi aðgerðir m.a. lagðar til:

- Að hugbúnaður til að lagfæra og leiðrétta íslenskt málfar verði gerður og kominn í notkun innan þriggja ára.
- Að nothæf þýðingarforrit milli íslensku og valinna erlendra mála, a.m.k. ensku, verði gerð innan fimm ára.
- Að íslenskur talgervill og talgreinir sem gerðir voru á vegum tungutækniátaks menntamálaráðuneytisins verði endurbættir og lagaðir að nýjustu tækni.

Nú, rúmum 5 árum eftir að tillögur Íslenskrar málnefndar voru samþykktar á Alþingi, er endurbættur talgervill orðinn að veruleika. Íslenskur talgreinir fyrir Android hefur verið þróaður í samstarfi við Google en er ekki nýtanlegur í öðrum kerfum. Fyrsti vísir að málfars-leiðréttingu með hugbúnaðinum *Skramba* er í þróun en nothæf þýðingarforrit milli íslensku og erlendra mála hafa ekki verið þróuð.

Máltækni er lykiltækni til þess að tryggja jafnan aðgang allra þegna að upplýsingum þar sem hún gerir fólki kleift að nota móðurmál sitt í samskiptum við tölvur og upplýsingakerfi. Máltækni getur einnig stutt við viðleitni til þess að tryggja grundvallarmannréttindi. Í annarri málsgrein 6. gr. laga um Ríkisútvarpið, fjölmiðil í almannaþágu, nr. 23 frá 20. mars 2013, er

kveðið á um að Ríkisútvarpið skuli veita heyrnarskertum aðgang að fjölmiðlaþjónustu í almannaþágu með textun á fréttum og öðru sjónvarpsefni, með textavarpi, útsendingum á táknmáli og/eða öðrum miðlunarleiðum er henta í þessu skyni og eru í samræmi við tæknilega möguleika á hverjum tíma. Ríkisútvarpinu hefur reynst erfitt að uppfylla þessa skyldu. Með því að nýta góðan talgreini mætti nálgast þetta markmið án mikils tilkostnaðar. Þetta er eitt dæmi um að nýting máltækni geti bætt grundvallarmannréttindi.

META–NORD

Á árunum 2011–2013 tók Máltæknisetur þátt í verkefninu *META-NORD*[7], sem var samstarfsverkefni Norðurlanda og Eystrasaltslanda og hluti af META-NET sem tekur til allra ríkja Evrópusambandsins og tengdra ríkja. Verkefnin voru styrkt af 7. rammaáætlun Evrópusambandsins og stefnumótunaráætlun sambandsins á sviði upplýsingatækni (*ICT Policy Support Programme*). Markmið verkefnisins var að skapa tæknilegar forsendur fyrir margmála upplýsingasamfélagi í Evrópu þar sem allir geti notað móðurmál sitt við öflun og úrvinnslu hvers kyns upplýsinga. Einn þáttur verkefnisins fólst í því að semja skýrslur um stöðu tungumála og máltækni í einstökum Evrópulöndum. Í íslensku skýrslunni, *Íslensk tunga á stafrænni öld*[8], kemur fram að máltæknistuðningur við íslensku er í flokknum „lítill sem enginn stuðningur“ og þar er íslenskan í hópi með írsku, lettnesku, litháísku og maltnesku. Af 30 Evróputungumálum stendur íslenskan næstverst að vígi þegar litið er til stuðnings við máltækni.

Íslenska í tölvuheiminum

Árið 2010 skipaði mennta- og menningarmálaráðherra nefnd til að fylgja eftir stefnu um notendaviðmót í íslenskum skólum og gera áætlun um aðrar aðgerðir sem tilgreindar eru í íslenskri málstefnu um íslensku í tölvuheiminum. Nefndin sendi frá sér skýrsluna *Íslenska í tölvuheiminum*[9] árið 2012 sem m.a. inniheldur tillögur á sviði máltækni. Þar segir m.a.:

> *Nefndin telur að það sé mjög mikilvægt að bæta stöðu íslenskrar máltækni og stefna að því að íslenska komist upp úr neðsta flokki í a.m.k. tveimur þeirra fjögurra þátta sem miðað er við í áðurnefndri skýrslu [innskot: META-NORD skýrslan] á næstu þremur árum. Til að svo megi verða þarf ríkið að styðja myndarlega við þróunarstarf á sviði máltækni. Árið 1999 áætlaði starfshópur um máltækni að það myndi kosta u.þ.b. einn milljarð króna á þágildandi verðlagi að gera íslenska máltækni sjálfbæra. En enda þótt máltækniáætlun menntamálaráðuneytisins 2000–2004 hafi verið árangursrík og haft mikil áhrif á þróun íslenskar máltækni verður að hafa í huga að ráðstöfunarfé hennar var aðeins um 1/8 af því sem starfshópur um máltækni taldi þurfa. Því þarf ekki að koma á óvart að íslensk máltækni standi enn mjög veikt.*

Nám í máltækni

Árið 2002 setti HÍ á laggirnar MA-nám í tungutækni. Vegna skorts á fjármagni og kennurum var ekki hægt að halda náminu úti lengur en til ársins 2004. Haustið 2007 hófu HÍ og HR síðan samstarf um meistaranám í máltækni. Inntöku nemanda var hætt haustið 2010 (aftur vegna fjárskorts) en hófst á ný haustið 2013. Engir nemendur voru teknir inn haustið 2014 en vonast er til að unnt verði að taka inn nemendur haustið 2015. 7 nemendur (3 úr HR, 4 úr HÍ)

hafa útskrifast með meistaragráðu í máltækni á Íslandi. Aðeins einn Íslendingur hefur útskrifast með doktorsgráðu á sviðinu úr erlendum háskóla, svo vitað sé.

Almannarómur

Nú er unnið að því að stofna sjálfseignarstofnunina *Almannaróm*[10] sem mun standa að smíði máltæknilausna fyrir íslensku. Undirbúningur að stofnun félagsins hófst á árinu 2013. Leitað var til fjölmargra fyrirtækja og félaga með ósk um að leggja fram stofnframlag og verða þar með stofnendur að félaginu. Fyrirtæki og félög geta einnig gerst styrktaraðilar að félaginu síðar. Stofnfundur var haldinn 5. júní 2014 og söfnun stofnfjár lauk í nóvember 2014. Tæknilausnir *Almannaróms* verða opnar öllum stofn- og styrktaraðilum, íslensku atvinnulífi og almenningi til góða.

Markmið sjálfseignarstofnunarinnar eru tvíþætt, annars vegar að auka samkeppnisfærni íslenskra fyrirtækja og hins vegar að auka mannréttindi og bæta samfélagið. *Almannarómur* mun vinna að markmiðum sínum með því að tryggja að íslenskan standi jafnfætis öðrum tungumálum í tækniheiminum með því að skapa og þróa íslensk máltæknitól eins og talgreini, vélrænar þýðingar, fyrirspurnarkerfi, samræðukerfi og kerfi fyrir stafsetningar- og málfars-leiðréttingar.

4. Markáætlanir í máltækni

Nokkrar Evrópuþjóðir sem við berum okkur gjarnan saman við (t.d. Bretland, Eistland, Finnland, Holland, Noregur og Svíþjóð) hafa sett á laggirnar sérstakar markáætlanir til eflingar einstakra sviða innan máltækni. Ástæðan er sú að mörg máltækniverkefni eru þess eðlis að þau falla illa að verksviði rannsóknarsjóða og annarra samkeppnissjóða, eins og þegar er getið. Nefndin skoðaði sérstaklega þrjár áætlanir: Í fyrsta lagi máltækniáætlun Eista, í öðru lagi menntunaráætlun Svía og Finna á sviði máltækni og í þriðja lagi ný lög í Bretlandi um notkun höfundarréttarvarins stafræns efnis.

Eistland

Í Eistlandi búa um 1,3 milljónir íbúa. Af þeim þjóðum sem tóku þátt í META-NORD-verkefninu 2011–2013 eru aðstæður í Eistlandi því einna líkastar aðstæðum á Íslandi. Aðgerðaráætlun Eista í máltækni[11] birtist í samkeppnissjóði sem styrkir eingöngu verkefni á sviði máltækni. Áætlunin var upphaflega til fimm ára en var síðan framlengd um sjö ár. Fyrri hlutinn stóð frá 2006 – 2010 og nam stuðningurinn jafnvirði 512 m.kr. Seinni hluti stendur frá 2011–2017 og áætlaður stuðningur nemur 1.136 m.kr. á því tímabili.

Áætluninni er skipt í fimm hluta og hverjum hluta er lýst með mælanlegum markmiðum. Fyrsti hlutinn snýst um rannsóknir og þróunarvinnu til þess að búa til frumgerðir af hugbúnaðartólum. Annar hluti áætlunarinnar snýr að gerð málsafna bæði með texta og tali. Í þriðja lagi er rætt um eistneskt máltæknisetur sem á að hafa það hlutverk að gera aðgengileg þau málföng (máltól og málsöfn) sem eru þegar til og safna saman og skrá ný. Í fjórða lagi er gert ráð fyrir vinnu við að tengja saman hin ýmsu tól og málsöfn. Í fimmta lagi er gert ráð fyrir að stjórn máltækniverkefnisins eða opinberir aðilar geti pantað tiltekin verkefni. Fyrir fyrstu tvo flokkana er auglýst eftir umsóknum árlega og mega þau verkefni standa í fjögur ár.

Auglýst er sjaldnar fyrir verkefni sem falla í fjórða og fimmta flokk og mega þau verkefni standa í tvö ár.

Gert er ráð fyrir að öllum málföngum sem gerð eru í verkefninu fylgi notendaleyfi. Mælt er með að notuð séu sem opnust leyfi sem geri notendum kleift að nýta málföngin í verkefnum sem ekki gefa tekjur og að ekki sé heimilt að framselja notkunarleyfið. Eistneska máltæknisetrið á að hafa umsjón með málföngum sem búin eru til í verkefninu, veita aðgang að þeim og heimildir til þess að nota þau.

Skipaður er stýrihópur sem í eiga sæti fulltrúar máltæknisérfræðinga, fólk frá eistneska menntamálaráðuneytinu, iðnaðarráðuneytinu og aðrir fræðimenn. Stýrihópurinn felur eistneska máltæknisetrinu umsjón með verkefninu.

Markmið verkefnisins eru að í lok verkefnistímans árið 2017 verði staðan þessi: Hugbúnaður fyrir málvinnslu verði notaður víða í þjóðfélaginu. Unnt verði að nota talgervingu og talgreiningu á tilteknum sviðum og hjálpartól fyrir ritun texta verði mikið notuð. Einnig verði til kerfi sem leyfi samtal manns og tölvu og til verði frumgerðir af kerfum fyrir vélrænar þýðingar. Eistar skipi sér í flokk þjóða þar sem máltækni er háþróuð.

Svíþjóð

Aðgerðaráætlun Svía í máltækni sneri að doktorsmenntun á sviðinu. Stofnaður var sérstakur skóli *Graduate School of Language Technology* (GSLT[12]) sem hlaut fjárhagslegan stuðning frá ríkinu í 10 ár, frá 2001–2009. Fjármagn til skólans var 200 m.kr. á ári. Sérhver doktorsnemi í máltækni fékk styrk sem nam frá 350.000 kr. – 450.000 kr. á mánuði í fjögur ár. GSLT var samstarf 8 sænskra háskóla en Gautaborgarháskóli var umsjónaraðili. Markmiðið með GSLT var að búa til breiðan þverfaglegan vettvang fyrir framhaldsmenntun í máltækni, auka gæðin á máltækninámi og útskrifa fólk með hæfni til að takast á við verkefni á sviðinu. Í lok áætlunarinnar höfðu rúmlega 40 nemendur útskrifast með doktorsgráðu í máltækni.

Finnland

Árið 1999 skrifaði Dr. Kimmo Koskenniemi, prófessor í máltækni hjá háskólanum í Helsinki, skýrslu[13] fyrir finnska menntamálaráðuneytið um nauðsyn þess að þróa máltækninám í Finnlandi. Í kjölfar skýrslunnar setti ráðuneytið á fót sérstakan sjóð til að styðja við kennslu í máltækni sem leiddi til stofnunar skólans *Graduate School of Language Technology* (KIT[14]) . Um það bil 15 nemendur útskrifuðust með doktorsgráðu í máltækni úr KIT.

Bretland

Í júní 2014 tóku gildi ný lög í Bretlandi um notkun höfundarréttarvarins stafræns efnis í rannsóknum, skólakerfinu, bókasöfnum og skjalasöfnum[15]. Breytingar sem leiða af nýju lögunum eiga að fjarlægja hindranir fyrir texta- og gagnanámi (e. *text and data mining*) í rannsóknum sem eru ekki gerðar í hagnaðarskyni. Í frétt um nýju lögin frá 1. júní 2014[16] er þess getið að þessi notkun textanna sé framlag í þágu framfara í ýmsum greinum vísinda og tækni og má þar fyrst nefna máltækni. Aðrar breytingar á lagaumhverfinu í Bretlandi munu gagnast margvíslegum stofnunum og félögum til þess að nota og varðveita eigið efni. Þess er

jafnframt getið að lagabreytingin muni koma til með að hafa veruleg áhrif á breskt hagkerfi næsta áratug.

Umrædd lagabreyting byggist á skýrslunni *Digital Opportunity: A review of Intellectual Property and Growth* frá 2011 um endurskoðun höfundaréttarlaga með tilliti til möguleika í stafrænni notkun[17].

CLARIN

CLARIN[18] er evrópskt innviðaverkefni sem miðar að því að byggja upp og staðla rafræn málleg gagnasöfn og hugbúnað til að nota við rannsóknir og þróunarverkefni, svo og viðhalda og veita aðgang að þessum gögnum og búnaði. Undirbúningsferli CLARIN stóð yfir árin 2008-2011 og var kostað af Evrópusambandinu. Ísland tók óbeinan þátt í því á lokastigi, þó án þess að fá nokkurn styrk. Upp úr undirbúningsferlinu hefur nú vaxið innviðaklasinn CLARIN ERIC, en ERIC (*European Research Infrastructure Consortium*) er ný tegund alþjóðlegra lögaðila sem ESB kom á fót 2009. Þátttakendur í ERIC eru ríki eða milliríkjastofnanir. Kostnaður við CLARIN ERIC er alfarið greiddur af þátttakendum sjálfum. Stefnan er að öll lönd Evrópska efnahagssvæðisins verði þátttakendur í CLARIN ERIC. Af Norðurlöndunum eru Danmörk og Svíþjóð þegar þátttakendur og Finnland og Noregur á leið inn. Engin skref hafa verið tekin í átt að þátttöku Íslands en þar þurfa stjórnvöld að eiga frumkvæði.

5. Forgangsverkefni

Í skýrslunni *Íslenska í tölvuheiminum* frá 2012, sem vísað er til hér að framan, eru tvö verkefni nefnd sem forgangsverkefni í íslenskri máltækni. Annað er **hugbúnaður til leiðréttingar á íslensku málfari** og hins vegar **íslenskur talgreinir**. Nefndin er sammála þessu en telur rétt að bæta **vélrænum þýðingum** við lista yfir forgangsverkefni enda eru þær einnig nefndar sem forgangsverkefni í Íslenskri málstefnu frá 2009. Hér á eftir fylgir rökstuðningur fyrir vali þessara þriggja verkefna en auk þess leggur nefndin til að lagaumhverfi verði breytt í þá veru að auðvelda notkun stafrænna texta í rannsóknum sem ekki eru gerðar í hagnaðarskyni.

Málfarsleiðrétting

Hugtakið málfarsleiðrétting (e. *grammar checking/correction*) vísar til skoðunar á því hvort málfarsreglum er fylgt, svo sem reglum um gerð setninga og innra samræmi, fallstjórn, samræmi frumlags og sagnar og annað slíkt. Hugbúnaður til leiðréttingar á málfari, t.d. hugbúnaður sem greinir villur í beygingu og setningagerð, er til fyrir mörg nágrannatungumál okkar. Fyrir Microsoft Office eða LibreOffice er t.d. hægt að fá viðbætur sem innihalda málfarsleiðréttingu fyrir dönsku, sænsku og finnsku, að ógleymdum „stóru“ málunum eins og ensku, þýsku og frönsku. Nágrannaþjóðum okkar finnst sjálfsagt að hægt sé að setja upp málfarsleiðréttingu fyrir þeirra tungumál í helstu ritvinnsluforritum – það ætti okkur líka að finnast.

Eins og áður hefur komið fram hafa ýmis grunntól, eins og málfræðilegur markari og þáttari, og málsöfn, eins og *Beygingarlýsing íslensks nútímamáls* og *Mörkuð íslensk málheild*, verið þróuð undanfarin ár, fyrst hjá Orðabók Háskólans (sem nú er hluti af Stofnun Árna Magnússonar í íslenskum fræðum) og eftir 2005 í samstarfi fræðimanna sem eiga aðild að

Máltæknisetri. Þessi málföng[19] koma að gagni við þróun hugbúnaðar til að leiðrétta málfar og hafa reyndar verið notuð að hluta til í hugbúnaðinum *Skramba*[20] sem telja má fyrsta vísi að hugbúnaði fyrir málfarsleiðréttingu á íslenskum texta.

Íslenskur talgreinir

Talgreining (e. *speech recognition*), vörpun talaðs máls yfir í texta, hefur smám saman orðið eitt mikilvægasta svið máltækni. Ástæðan er m.a. sú að góður talgreinir gefur notendum kost á að eiga samskipti við tölvustýrð tæki með því að nota talað mál í stað músar og lyklaborðs. Tækin sem nýta sér talgreiningu geta spurt notendur spurninga og leyft þeim að svara á þann máta sem er notandanum eðlilegastur, þ.e. með tali. Hægt er að nýta talgreiningu á mörgum sviðum, s.s. í tölvukerfum bíla, í heilbrigðiskerfinu við innlestur læknaskýrslna, í þjónustuverum ýmissa fyrirtækja, í tölvustuddu tungumálanámi, til stuðnings heyrnaskertum, lesblindum og öðrum sem eiga, vegna fötlunar sinnar, erfitt með innslátt texta. Talgreinir er jafnframt forsenda fyrir ýmiss konar nýsköpunarverkefnum sem byggja á samskiptum milli manns og tölvu með notkun talaðs máls.

Eins og að framan greinir hefur fyrirtækið Google, í samstarfi við Máltæknisetur, þróað talgreini fyrir íslensku. Þáttur Máltækniseturs fólst í því að safna íslenskum raddsýnum sem notuð voru til að þjálfa Google-talgreininn. Umrædd raddsýni eru opin og öllum aðgengileg en talgreinirinn sjálfur er hins vegar eign Google. Það er engin vissa fyrir því að fyrirtækið Google muni halda áfram að bæta hann fyrir íslensku og jafnframt er engin leið að nýta hann í öðrum tækjum en þeim sem byggja á Android-stýrikerfinu.

Þess vegna er mjög mikilvægt fyrir Íslendinga að þróa sinn eigin talgreini. Hér er átt við að þróa talgreini sem byggir að hluta til á umræddum raddsýnum en jafnframt þarf að safna fleiri sýnum. Frumkóði (texti forrits) Google-talgreinisins er ekki opinn og erfitt er að nýta hann í öðrum tækjum en þeim sem Google styður. Vegna þessa er mikilvægt að nýr talgreinir verði opinn og aðgengilegur fyrir alla þá sem vilja nýta talgreininn í eigin tölvukerfum.

Vélrænar þýðingar

Í upplýsingasamfélagi nútímans er mikilvægt að geta sett texta fram á ýmsum tungumálum svo að hann gagnist sem flestum. Í vélrænum þýðingum eru tölvur (hugbúnaður) notaðar til að þýða texta úr einu tungumáli, frummáli, yfir á annað tungumál, markmál. Fyrirtækið Google hefur unnið merkilegt starf á sviði vélrænna þýðinga undanfarin ár og gert þýðingarvél[21] sína aðgengilega almenningi. Fyrirtækið byggir sjálfvirkar þýðingar á tölfræðilegum líkönum þar sem stikar líkansins fást með sjálfvirkri greiningu á samhliða málheildum (e. *parallel corpora*), þ.e. textum á tilteknu frummáli og sömu (þýddum) textum á tilteknu markmáli.

Önnur meginaðferð sem notuð er í þýðingarkerfum byggir á málfræðilegum reglum. Til eru frumgerðir af þess konar kerfum sem styðja íslenskar þýðingar, eins og *Tungutorg*[22] og kerfið *Apertium*[23]. Það síðarnefnda byggir á opnum hugbúnaði og opnum gögnum.

Hvorki þýðingarvél Google né gögnin sem hún er þjálfuð á eru aðgengileg öðrum sem vilja þróa eigin þýðingarvélar. Fyrirtæki eða stofnanir sem vilja bjóða upp á sjálfvirkar þýðingar á milli íslensku og annarra tungumála geta vissulega keypt aðgang að *Google Translate API*[24] en hvað gerum við ef Google ákveður einn daginn að leggja niður þýðingarvél sína fyrir

íslensku? Af þessum sökum er mikilvægt fyrir Íslendinga að þróa þýðingarkerfi sem byggir á opnum hugbúnaði og opnum gögnum.

Þýðingarminni (e. *translation memory*) er gagnagrunnur sem geymir „búta“, t.d. einstakar setningar, málsgreinar, fyrirsagnir og titla, sem þegar hafa verið þýddir. Tilgangur þýðingarminnis er að endurnýta og samræma þýðingar. Mikilvægt er að sett verði á fót opið þýðingarminni á Íslandi sem auðveldi þýðingar á nytjatextum.

6. Tillögur

Eins og áður segir er tilgangur nefndarinnar að gera áætlun um aðgerðir er miði að því að gera íslensku gjaldgenga í stafrænni upplýsingatækni og stuðla að notkun hennar á þeim vettgangi. Í 3. kafla var staða og þróun íslenskrar máltækni rakin undanfarin 15 ár. Íslensk máltækni er enn á bernskuskeiði en segja má að tekist hafi að leggja grunn að sviðinu þrátt fyrir afar takmarkað fjármagn. Íslenskan er samt sem áður í neðsta flokki Evrópumála (flokknum „Lítill sem enginn stuðningur“, samkvæmt META-NORD-skýrslunni) m.t.t. stuðnings við máltækni.

Það er einkum tvennt sem stendur íslenskri máltækni fyrir þrifum. Í fyrsta lagi er skortur á fólki með sérfræðiþekkingu á sviðinu. Íslenska mun ekki verða gjaldgeng í upplýsingatækniþjóðfélaginu án fólks með doktorsmenntun í máltækni. Menntunin veitir fólki nauðsynlega þekkingu og þjálfun í þeim aðferðum sem beitt er í máltækni og ýtir undir þróun og nýsköpun á sviðinu. Í öðru lagi er skortur á fjármagni til einstakra verkefna, bæði tækniþróunarverkefna og innviðaverkefna. Nánast engin íslensk fyrirtæki hafa séð sér hag í því að þróa íslenskar máltæknilausnir því að þróun grunntóla og innviða er ekki komin nógu langt til að fyrirtæki leggi í þróun tæknilausna. Eins og fram kemur í Íslenskri málstefnu (3. kafli) kostar jafnmikið að byggja innviði og máltæknilausnir fyrir íslensku og fyrir tungumál milljónaþjóða. Opinber stuðningur við íslenska máltækni er því mjög mikilvægur.

Aðgerðir

Nefndin leggur til að fjárfest verði í íslenskri máltækni með sérstakri langtímaáætlun til 10 ára sem styrki bæði doktorsnema og einstök verkefni og fylgi þannig fordæmi nokkurra Evrópuþjóða sem við berum okkur gjarnan saman við. Markmiðið er að eftir 10 ár verði íslenskan ekki lengur í neðsta flokki Evrópumála þegar litið er til stuðnings við máltækni heldur komin í flokk með hinum Norðurlöndunum (í flokkinn „Brotakenndur stuðningur“, samkvæmt META-NORD-skýrslunni). Stefnt er að því að í lok áætlunarinnar verði búið að útskrifa 5 nemendur með doktorsgráðu á sviðinu. Nýtt fólk sem hefur sérhæft sig í máltækni er lífsnauðsynlegt fyrir sviðið, bæði til að nýliðun vísindamanna í máltækni á Íslandi eigi sér stað og einnig er það forsenda fyrir því að nýsköpunarverkefni verði til.

Áætlunin felur í sér árlegt fjárframlag ríkisins í 10 ár. Um verði að ræða samkeppnissjóð, „Máltæknisjóðinn“, sem Rannsóknamiðstöð Íslands (Rannís) verði falin umsjón með og mat á umsóknum í sjóðinn verði á höndum erlendra aðila. Sjóðurinn verði þrískiptur:

1. Styrkir vegna tækniþróunarverkefna
2. Styrkir vegna innviðaverkefna
3. Styrkir til doktorsnema

Í tækniþróunarverkefnum njóti verkefni á sviði talgreiningar, málfarsleiðréttingar og vélrænna þýðinga forgangs fyrstu árin. Lögð verði áhersla á þróun hugbúnaðar sem byggir á opnum leyfum og opnum gögnum til að tryggja að auðvelt verði að þróa kerfin áfram af komandi kynslóðum. Í innviðaverkefnum er jafnframt mikilvægt að gagnasöfn verði öllum aðgengileg samkvæmt opnum og stöðluðum leyfum. Þá er mikilvægt að leggja fé í þátttöku Íslands í CLARIN ERIC (árgjald er nú um 1.800 þús. kr. fyrir Ísland en hækkar á næstu árum) til að unnt verði að fylgjast með þróuninni í þessum efnum erlendis, taka þátt í henni og hafa gagn af.

Kostnaðarmat

Nefndinni var falið að gera tillögur um verkefni sem ráðist yrði í og meta kostnað við þau. Þar eð hér er um langtímaverkefni að ræða er útilokað á þessu stigi að gera nákvæma fjárhagsáætlun enda þróast tæknin ört og forsendur breytast á ýmsan hátt. Hins vegar er ljóst að aðgerðir sem miða að því að gera íslensku gjaldgenga í stafrænni upplýsingatækni og stuðla að notkun hennar á þeim vettvangi eru gríðarstórt verkefni. Kostnaðurinn við það hleypur á mörg hundruð milljónum króna og nægir þar að vísa í máltækniverkefni Eista sem fjallað er um í 4. kafla, en nefndin hefur einnig haft hliðsjón af máltækniáætlunum Svía og fleiri þjóða.

Nefndin metur það svo að um það bil 990 m.kr. – tæplega einn milljarð króna – þurfi á næstu 10 árum til að ná markmiðum áætlunarinnar. Möguleg skipting kostnaðar á einstök ár og verkefni er sýnd í töflu 1.

Taflan sýnir þá þrískiptingu sem lögð er til í Máltæknisjóðnum, þ.e. styrki til tækniþróunarverkefna, innviðaverkefna og til doktorsnema. Í tækniþróunarverkefnum er fyrst og fremst verið að nýta tiltæka tækni, laga hana að íslensku og íslenskum aðstæðum og beita henni á íslensk gögn og gagnasöfn til að smíða hugbúnað og tól til að vinna með íslensku. Í þessum flokki er gert ráð fyrir að unnið verði að forgangsverkefnum á næstu fjórum árum (2015-2018) en eftir það muni Máltæknisjóðurinn styrkja ýmis önnur tækniþróunarverkefni. Kostnaður við þróun talgreinis er áætlaður 90 m.kr., en þar er stuðst við ítarlega kostnaðaráætlun sem sjálfseignarstofnunin Almannarómur (sjá 3. kafla) hefur gert fyrir umsókn til Tækniþróunarsjóðs (sem ekki hlaut brautargengi). Kostnaður við kerfi fyrir vélrænar þýðingar er áætlaður 100 m.kr. en kostnaður við þróun kerfis fyrir málfarsleiðréttingu er áætlaður 30 m.kr. Sem dæmi um önnur tækniþróunarverkefni má nefna samræðukerfi, fyrirspurnarkerfi, kerfi fyrir málgreiningu og kerfi fyrir merkingargreiningu.

	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	
Tækniþróunarverkefni											**470**
Málfarsleiðrétting	10	20									30
Talgreining	20	30	40								90
Vélrænar þýðingar		20	40	40							100
Önnur tækniþróunarv.				20	35	35	40	40	40	40	250
Innviðaverkefni											**425**
Uppbygging gagnasafna		5	10	25	35	35	35	35	35	35	250
Viðhald gagnasafna	5	5	5	10	15	15	15	15	15	15	115
Önnur innviðaverkefni					10	10	10	10	10	10	60
Doktorsstyrkir	5	10	10	10	10	10	10	10	10	10	**95**
Samtals	**40**	**90**	**105**	**105**	**105**	**105**	**110**	**110**	**110**	**110**	**990**

Tafla 1: Áætlaður kostnaður einstakra verkefna á 10 ára tímabili

Forgangsverkefnunum í aðgerðaráætluninni er hægt er að skipta upp í eftirfarandi áfanga:

- Í lok árs 2016 verður fyrsta útgáfa af alhliða málfarsleiðréttingaforriti aðgengileg í helstu ritvinnslukerfum
- Í lok árs 2017 verður fyrsta útgáfa af íslenskum talgreini nýtanleg hjá fyrirtækjum og stofnunum
- Í lok árs 2018 verður fyrsta útgáfa af íslensku þýðingarkerfi nýtanleg hjá fyrirtækjum og stofnunum

Forsenda fyrir öllum máltæknibúnaði eru ítarleg málleg gagnasöfn þar sem er að finna nákvæmar og flokkaðar upplýsingar um alla þætti tungumálsins – hljóð og hljóðkerfi, beygingar, setningagerð, orðaforða, merkingu o.s.frv. Slík söfn er vitaskuld ekki hægt að fá að láni erlendis frá – þau verður að byggja upp fyrir hvert tungumál. Vegna þess að málið og málnotkunin er sífellt að breytast er nauðsynlegt að vinna stöðugt að viðhaldi þessara safna en mikill kostnaður fylgir uppbyggingu og viðhaldi þeirra. Helstu innviðaverkefni felast í uppbyggingu nýrra mállegra gagnasafna en þar er kostnaður áætlaður 250 m.kr. Þar má nefna merkingargreint orðasafn (WordNet[25]) sem nýtist í vélrænum þýðingum, leitarkerfum, samræðukerfum o.fl., fleiri og stærri málheildir, samhliða málheildir, orðasöfn og þýðingarminni. Einnig þarf að viðhalda og staðla þau gagnasöfn sem þegar hafa verið byggð upp og efla þau og bæta mörkun og þáttun (áætlaður kostnaður 115 m.kr.), því að málleg gagnasöfn úreldast mjög fljótt vegna breytinga á orðaforða og málnotkun.

Kostnaðaráætlunin gerir ráð fyrir 95 m.k.r. styrkjum til doktorsnema í máltækni á næstu 10 árum. Með þessu er gert ráð fyrir að Máltæknisjóðurinn hafi fjármagn til að geta styrkt tvo doktorsnema á hverju ári.

[1] http://www.menntamalaraduneyti.is/media/MRN-pdf_Upplysingar-Utgefid/tungutaekni.pdf
[2] http://bin.arnastofnun.is/
[3] http://mim.arnastofnun.is
[4] http://www.maltaeknisetur.is
[5] http://www.blind.is/verkefni/talgervlaverkefnid/
[6] http://rafhladan.is/handle/10802/4405
[7] http://vefir.hi.is/metanord/
[8] http://www.meta-net.eu/whitepapers/volumes/icelandic
[9] http://www.menntamalaraduneyti.is/ahugavert/nr/7369
[10] http://almannaromur.is/
[11] https://www.keeletehnoloogia.ee/en/national-programme-for-estonian-language-technology-2011-2017
[12] http://www.gslt.hum.gu.se
[13] http://www.ling.helsinki.fi/users/koskenni/kieliteknologia/opm-raportti.html
[14] http://www.ling.helsinki.fi/kit/tutkijakoulu/courses/
[15] http://www.legislation.gov.uk/uksi/2014/1372/pdfs/uksi_20141372_en.pdf
[16] https://www.gov.uk/government/news/new-exceptions-to-copyright-reflect-digital-age
[17] https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth
[18] http://clarin.eu/
[19] http://malfong.is
[20] http://skrambi.arnastofnun.is/
[21] https://translate.google.com/
[22] http:/tungutorg.is
[23] http://www.apertium.org/index.eng.html#translation
[24] https://cloud.google.com/translate/
[25] http://wordnet.princeton.edu/