META-NET hvítbókaröðin

Tungumál í evrópsku upplýsingasamfélagi

– Íslenska–

Gerð þessarar hvítbókar var kostuð af Sjöundu rammaáætlun Evrópusambandsins og Stefnumótunaráætlun Evrópusambandsins í upplýsinga- og samskiptatækni samkvæmt samningum við T4ME (samningur 249119), CESAR (samningur 271022), METANET4U (samningur 270893) og META-NORD (samningur 270899).

Formáli

Þessi hvítbók er hluti af ritröð til kynningar á máltækni og möguleikum hennar. Henni er einkum beint til fólks sem starfar i menntageiranum, á fjölmiðlum, í stjórnmálum – og í raun til málsamfélaga í heild.

Aðgengi að máltækni og notkun hennar er mjög mismunandi milli tungumála í Evrópu. Þar af leiðir að aðgerðir sem nauðsynlegar eru til að styðja rannsóknir og þróunarstarf í máltækni eru einnig ólíkar milli mála. Það fer eftir ýmsu hvaða aðgerða er þörf, svo sem stærð málsamfélagsins og hversu flókið tungumálið er.

META-NET, öndvegisnet fjármagnað af Evrópusambandinu, hefur lagt mat á núverandi stöðu í málföngum og máltækni. Þessi greining tók til hinna 23 opinberu mála Evrópusambandsins auk annarra mikilvægra þjóðtungna og svæðisbundinna tungumála í álfunni. Niðurstöður þessarar greiningar benda til að í öllum málunum séu umtalsverðar eyður í rannsóknum. Nákvæmari greining sérfræðinga og mat á núverandi stöðu mun hjálpa til við að hámarka árangur viðbótarrannsókna og lágmarka áhættu.

META-NET tengir saman 47 rannsóknarsetur frá 31 landi. Þau vinna með hagsmunaaðilum úr viðskiptalífinu, frá opinberum stofnunum, iðnaðinum, rannsóknarstofnunum, hugbúnaðarfyrirtækjum og evrópskum háskólum. Þessir aðilar eru í sameiningu að þróa heildstæða tæknisýn og móta útfærða rannsóknarstefnu sem sýnir hvernig unnt verður að nýta máltækni á öllum sviðum árið 2020.

Hafið samband

Máltæknisetur
Hugvísindastofnun Háskóla Íslands
Aðalbyggingu
101 Reykjavík
Ísland

http://maltaeknisetur.is

Höfundar

Eiríkur Rögnvaldsson, prófessor, Háskóla Íslands
Dr. Kristín M. Jóhannsdóttir, Háskóla Íslands
Steinþór Steingrímsson, MSc, Háskóla Íslands
Sigrún Helgadóttir, MSc, Stofnun Árna Magnússonar í íslenskum fræðum
Dr. Hrafn Loftsson, Háskólanum í Reykjavík

Efnisyfirlit

Yfirlit

Máltækni reisir brýr í þágu framtíðar Evrópu

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 sama tungumál.

Evrópumarkaður hefur grundvallarþýðingu fyrir íslenskt efnahagslíf. Árið 2010 fór 81,8% af útflutningi Íslendinga til landa Evrópska efnahagssvæðisins og 5,9% til annarra Evrópulanda.[i] En tungumálaþröskuldar geta komið í veg fyrir viðskipti, sérstaklega hjá smáum og meðalstórum fyrirtækjum sem hafa ekki fjármagn til að bregðast við aðstæðum. Eini valkosturinn í margmála Evrópu af þessu tagi væri að veita einu tungumáli ráðandi stöðu þannig að það útrýmdi að lokum öllum öðrum tungumálum. Þetta er óhugsandi.

Hin hefðbundna leið til að komast yfir tungumálaþröskulda er að læra erlend mál. En án tæknilegs stuðnings er vald á 23 opinberum málum Evrópusambandsins og um 60 öðrum Evrópumálum óyfirstíganleg hindrun fyrir Evrópubúa, sem og fyrir efnahag álfunnar, stjórnmálaumræðu og framfarir í vísindum.

Lausnin er sú að koma upp grundvallarstuðningstækni. Það mun verða allri Evrópu til mikilla hagsbóta, ekki aðeins á hinum sameiginlega evrópska markaði heldur einnig í viðskiptum við önnur lönd, sérstaklega vaxandi hagkerfi. Til að ná þessu marki og varðveita jafnframt menningarlega og mállega fjölbreytni Evrópu er nauðsynlegt að gera kerfisbundna greiningu á mállegum sérkennum allra Evrópumála, svo og á máltæknilegum stuðningi við þau. Máltækni mun á endanum mynda einstaka brú milli evrópskra tungumála.

Máltækni sem lykill handa framtíðinni

Sá búnaður til vélþýðinga og vinnslu talmáls sem nú er á markaðnum stendur ekki enn undir þessum metnaðarfullu kröfum. Ráðandi aðilar á markaðnum eru einkafyrirtæki rekin í hagnaðarskyni með höfuðstöðvar í Norður-Ameríku. Þegar á síðari hluta áttunda áratugar síðustu aldar gerði Evrópusambandið sér grein fyrir mikilvægi máltækni fyrir einingu Evrópu, og veitti fé til fyrstu rannsóknarverkefna sinna á þessu sviði, svo sem EUROTRA. Um sama leyti settu ýmsar þjóðir á fót innlend verkefni sem skiluðu verðmætum niðurstöðum en leiddu ekki til samstilltra evrópskra aðgerða. Öfugt við þessar mjög afmörkuðu aðgerðir til fjármögnunar hafa önnur margmála samfélög eins og Indland (22 opinber tungumál) og Suður-Afríka (11 opinber tungumál) nýlega komið á fót innlendum langtímaverkefnum til málrannsókna og tækniþróunar.

Helstu fyrirtæki á sviði máltækni um þessar mundir reiða sig á ónákvæmar tölfræðiaðferðir sem nýta ekki dýpri málvísindalega þekkingu og aðferðafræði. Setningar eru t.d. þýddar sjálfvirkt með því að máta nýja setningu við þúsundir setninga þýddra af mennskum þýðendum. Gæði útkomunnar fara að miklu leyti eftir magni og gæðum dæmasafnsins. Þótt sjálfvirk þýðing einfaldra texta á málum þar sem til er nægilegt textamagn geti skilað nothæfum niðurstöðum eru slíkar grunnar tölfræðiaðferðir

dæmdar til að bregðast í tungumálum með lítið af gögnum til að byggja á eða þegar formgerð setninga er flókin.

Evrópusambandið hefur þess vegna ákveðið að fjármagna verkefni eins og EuroMatrix og EuroMatrixPlus (frá 2006) og iTranslate4 (frá 2010) sem sinna bæði grunnrannsóknum og hagnýtum rannsóknum og framleiða gögn til að standa undir hágæða máltæknilausnum fyrir öll Evrópumál. Ef við viljum byggja upp máltæknibúnað sem skilar góðum árangri fyrir öll Evrópumál er eina leiðin að greina grundvallarformgerð tungumálanna nákvæmlega.

Evrópskar rannsóknir á þessu sviði hafa þegar skilað verulegum árangri. Þýðingarþjónusta Evrópusambandsins notar t.d. opna þýðingarhugbúnaðinn MOSES sem var aðallega þróaður innan evrópskra rannsóknarverkefna. En í stað þess að byggja á niðurstöðum rannsóknarverkefna sinna hefur Evrópa haft tilhneigingu til að styðja einangruð rannsóknarverkefni sem hafa mun takmarkaðri markaðsáhrif. Efnahagslegt gildi jafnvel fyrstu verkefnanna á þessu sviði má marka af fjölda sprotafyrirtækja sem frá þeim eru runnin.

Máltækni hjálpar til við að sameina Evrópu

Út frá þeirri reynslu sem fengist hefur virðist svo sem 'blönduð' máltækni nútímans, þar sem djúp málgreining er samtvinnuð tölfræðilegum aðferðum, geti brúað bilið milli allra Evrópumála – og til annarra mála. Eins og kemur fram í þessari hvítbókaröð er gífurlega mismunandi milli Evrópulanda hversu vel tungumál þeirra eru stödd í rannsóknum og máltæknilausnum. Þessi skýrsla sýnir að það er einungis á sviði grundvallarbúnaðar og málfanga svo sem málfræðimörkunar, setningafræðilegrar þáttunar, málheilda og trjábanka sem staða íslenskunnar er viðunandi. Á flóknari sviðum eins og í merkingargreiningu setninga og texta, samræðukerfum, upplýsingaheimt, málmyndun, gerð samantektar, merkingargreindra málheilda, o.s.frv., er ekkert til fyrir íslensku. Því er ljóst að mikið starf er óunnið við að tryggja framtíð íslenskunnar sem fullgilds þátttakanda í evrópsku upplýsingasamfélagi nútímans – og framtíðarinnar.

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 tungumála og reisa brýr milli tungumála í Evrópu. Þetta krefst þess að allir hagsmunaaðilar – í stjórnmálum, rannsóknum, viðskiptum, og samfélaginu öllu – sameini krafta sína í þágu framtíðar.

Þessi hvítbókaröð tengist öðrum markvissum aðgerðum sem META-NET stendur að (sjá yfirlit þeirra í viðbæti). Nýjustu upplýsingar eins og gildandi útgáfu framtíðarsýnar[ii] META-NET og útfærða rannsóknarstefnu (Strategic Research Agenda, SRA) er að finna á vefsetri META-NET: http://www.meta-net.eu.

Hættur sem steðja að tungumálinu – ögrun fyrir máltækni

Við verðum um þessar mundir vitni að stafrænni byltingu sem hefur gífurleg áhrif á samskipti og samfélag. Nýleg þróun í stafrænni upplýsinga- og samskiptatækni er stundum borin saman við það þegar Gutenberg fann upp prentverkið. Hvað getur sú samlíking sagt okkur um framtíð evrópsks upplýsingasamfélags og sérstaklega tungumála okkar?

Við verðum um þessar mundir vitni að stafrænni byltingu sem hefur sambærileg áhrif og uppfinning prentverksins á sínum tíma

Eftir uppfinningu Gutenbergs voru stigin tímamótaskref í samskiptum og þekkingarskiptum með verkum eins og t.d. þýðingu Lúthers á Biblíunni yfir á þjóðtunguna. Á þeim öldum sem síðan eru liðnar hafa verið þróaðar menningarbundnar aðferðir til að sinna betur málvinnslu og þekkingarskiptum:

- Stöðlun stafsetningar og málfræði helstu tungumála skapaði möguleika á hraðri útbreiðslu nýrra vísindalegra og vitsmunalegra hugmynda;
- þróun opinberra tungumála gerði fólki kleift að ræða saman innan ákveðinna (oft pólitískra) landamerkja;
- tungumálakennsla og þýðingar milli mála gerðu það mögulegt að eiga samskipti þvert á tungumál;
- ritstjórnarreglur og bókfræðileg viðmið tryggðu gæði prentaðs efnis og aðgengi að því;
- tilkoma mismunandi fjölmiðla, svo sem dagblaða, útvarps, sjónvarps, bóka o.fl. fullnægði mismunandi samskiptaþörfum.

Á síðustu tuttugu árum hefur upplýsingatæknin átt sinn þátt í því að greiða fyrir mörgum ferlum og gera þau sjálfvirk:

- ritvinnslu- og umbrotskerfi hafa komið í stað vélritunar og setningar;
- Microsoft PowerPoint hefur komið í staðinn fyrir glærur og myndvarpa;
- með tölvupósti eru skjöl send og tekið á móti þeim mun hraðar en með bréfasíma;
- Skype býður upp á ódýr netsímtöl og skapar vettvang fyrir fjarfundi;
- snið hljóð- og myndbandaskráa gerir auðvelt að skiptast á margmiðlunarefni;
- leitarvélar greiða notendum aðgang að vefsíðum með leit byggðri á lykilorðum;
- netþjónusta eins og Google Translate skilar nokkurn veginn réttum þýðingum hratt;
- félagsmiðlar eins og Facebook, Twitter, og Google+ greiða fyrir samskiptum, samvinnu og deilingu upplýsinga.

Þrátt fyrir gagnsemi slíkra tóla og búnaðar dugir þetta ekki til að standa undir sjálfbæru margmála evrópsku samfélagi fyrir alla, með frjálsu flæði upplýsinga og varnings.

Tungumálaþröskuldar standa í vegi fyrir evrópsku upplýsingasamfélagi

Við getum ekki vitað nákvæmlega hvernig upplýsingasamfélag framtíðarinnar mun líta út. En miklar líkur eru á því að bylting í samskiptatækni muni 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

Í alþjóðasamfélagi viðskipta og upplýsinga tengjast sífellt fleiri tungumál og málnotendur sífellt hraðar með hjálp nýrra miðla.

skilning og aðgang að deilanlegri þekkingu. Í alþjóðasamfélagi viðskipta og upplýsinga tengjast sífellt fleiri tungumál og málnotendur sífellt hraðar með hjálp nýrra miðla. Vinsældir félagsmiðla (Wikipedia, Facebook, Twitter, YouTube og nú nýlega Google+) eru einungis toppurinn á ísjakanum.

Nú á dögum getum við flutt margra gígabæta texta um heiminn þveran og endilangan á örfáum sekúndum áður en við áttum okkur á því að hann er á máli sem við skiljum ekki. Samkvæmt nýrri skýrslu frá framkvæmdastjórn Evrópusambandsins keyptu 57% evrópskra netnotenda vörur og þjónustu með því að nota tungumál önnur en móðurmál sitt. (Enska er algengasta erlenda tungumálið á þessu sviði en þar á eftir koma franska, þýska og spænska.) 55% notenda lesa erlent mál sér til gagns en aðeins 35% nota annað tungumál til þess að skrifa tölvupóst eða gera athugasemdir á vefnum.[iii] Fyrir nokkrum árum var enska tungumál netsins – megnið af því efni sem þar var að finna var skrifað á ensku – en þetta hefur nú gerbreyst. Algjör sprenging hefur orðið í textamagni á öðrum Evrópumálum á netinu (og sama gildir um tungumál Asíu og Mið-Austurlanda).

Það sætir furðu að hin altæka stafræna gjá sem munur tungumála skapar skuli ekki hafa fengið mikla athygli í opinberri umfjöllun; samt sem áður vekur hún mjög brýna spurningu: Hvaða Evrópumál munu dafna í netvæddu upplýsinga- og þekkingarsamfélagi, og hver eru dæmd til að hverfa?

Tungumál okkar í hættu

Þótt prentverkið hraðaði deilingu upplýsinga í Evrópu olli það því einnig að mörg evrópsk tungumál liðu undir lok. Textar á svæðisbundnum málum og minnihlutamálum komust sjaldan á prent og því voru tungumál eins og korníska og dalmatíska eingöngu notuð í töluðu máli sem takmarkaði notkunarsvið þeirra. Mun netið hafa sambærileg áhrif á tungumál okkar?

Hin u.þ.b. 80 tungumál Evrópu eru ein ríkulegustu og mikilvægustu menningarverðmæti álfunnar og grundvallarþáttur í hinni einstöku samfélagsgerð hennar.[iv] Þótt tungumál eins og enska og spænska munu að öllum líkindum halda stöðu sinni á hinu stafræna markaðstorgi sem er að verða til gætu mörg evrópsk tungumál orðið gagnslaus í netvæddu samfélagi. Slík þróun myndi veikja alþjóðlega stöðu Evrópu og stangast á við markmið um jafna samfélagsþátttöku allra Evrópuþegna óháð tungumáli. Í 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.[v]

Hin fjölbreyttu tungumál Evrópu eru ein ríkulegustu og mikilvægustu menningarverðmæti álfunnar.

Máltækni er grundvallarstuðningstækni

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.[vi] Samt sem áður fullnægir þessi upphæð einungis litlum hluta núverandi þarfar og framtíðarþarfa fyrir samskipti milli tungumála. Augljósasta aðferðin til að tryggja breidd og dýpt málnotkunar í Evrópu 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.

Máltækni hjálpar fólki til að vinna saman, stunda viðskipti, deila þekkingu og taka þátt í rökræðum þvert á tungumál.

Stafræn máltækni (sem beinist að öllum myndum ritaðs máls og talsamskipta) gerir fólki kleift að vinna saman, stunda viðskipti, deila þekkingu og taka þátt í félagslegum og pólitískum rökræðum óháð tungumáli og

tölvufærni. Hún er oft hulinn hluti af flóknum hugbúnaði sem við nýtum okkur þegar við:

- finnum upplýsingar með notkun leitarvéla á netinu;
- rýnum stafsetningu og málfræði í ritvinnslukerfi;
- skoðum umsagnir um vörur í netverslun;
- hlustum á talaðar leiðbeiningar leiðsagnarkerfis í bíl;
- þýðum vefsíður með hjálp netþjónustu.

Máltækni felst í ýmsum grundvallarbúnaði sem margvísleg ferli innan stærri hugbúnaðarkerfa byggjast á. Tilgangur hvítbókaraðar META-NET er að skerpa sýn okkar á það hversu þroskuð þessi grunntækni er fyrir hin ýmsu Evrópumál.

Evrópa þarfnast traustrar og ódýrrar máltækni fyrir öll tungumál álfunnar.

Til að viðhalda stöðu sinni í fararbroddi nýsköpunar á heimsvísu þarfnast Evrópa máltækni sem er löguð að öllum evrópskum tungumálum og er traust, ódýr og vel samþættuð helstu hugbúnaðarumhverfum. Án máltækni munum við ekki geta skapað notendum árangursríkan, gagnvirkan, margmiðlunar- og margmála reynsluheim í náinni framtíð.

Tækifæri máltækninnar

Í prentheiminum varð stærsta tæknibyltingin þegar farið var að fjölfalda myndir eða texta með notkun prentvéla. Menn þurftu áfram að fletta upp þekkingaratriðum, lesa, þýða, og taka saman þekkingu. Það þurfti að bíða eftir Edison með upptökur á talmáli – en sú tækni bjó þó einnig aðeins til afrit.

Stafræn máltækni getur nú gert sjálfvirkt allt ferlið við þýðingu, samningu efnis og þekkingarstjórnun fyrir öll evrópsk tungumál. Hún getur einnig raungert þróun eðlilegs stýriviðmóts sem byggt er á máli og tali fyrir heimilisraftæki, vélar, bifreiðar, tölvur og vélmenni. Þróun viðskipta- og iðnaðarverkbúnaðar er enn á frumstigi, en áfangar í rannsóknum og þróun á þessu sviði eru þó farnir að opna mikla möguleika. Til dæmis eru vélþýðingar nú þegar sæmilega nákvæmar á afmörkuðum sviðum og tilraunabúnaður skilar margmála upplýsingum og sinnir þekkingarstjórnun og samningu efnis á mörgum Evrópumálum.

Eins og oftast er með tækni var fyrsti máltæknibúnaðurinn, svo sem raddstýrð notendaviðmót og samræðukerfi, þróaður með mjög sérhæfða notkun í huga og sýnir því oft takmarkaða hæfni. En geysimikil markaðstækifæri er að finna í menntageiranum og skemmtanaiðnaðinum þar sem hægt væri að nýta máltækni í leikjum, menningarminjasetrum, menntandi skemmtun, bókasöfnum, hermun og æfingaáætlunum. Upplýsingaþjónusta í farsíma, hugbúnaður fyrir tölvustutt tungumálanám, fjarnámsumhverfi, sjálfsmatstól og forrit til að uppgötva ritstuld eru fáein dæmi þar sem máltækni getur leikið mikilvægt hlutverk. Vinsældir félagsmiðla eins og Twitter og Facebook benda til þess að þörf sé á háþróaðri máltækni sem getur haldið utan um póst, gert útdrætti úr umræðum, bent á hneigð í skoðunum, greint tilfinningaleg svör, bent á brot á höfundarétti eða haft uppi á misnotkun.

Máltækni hjálpar fólki að sigrast á þeirri 'fötlun' sem felst í málfræðilegum fjölbreytileik.

Í máltækni felast gífurleg tækifæri fyrir evrópskt samstarf. Hún getur hjálpað okkur að takast á við hið flókna málumhverfi í Evrópu – þá staðreynd að mismunandi tungumál lifa eðlilegu samlífi í evrópskum viðskiptum, samtökum og skólum. En þegnarnir þurfa að geta haft samskipti yfir þessi tungumálamörk sem skera hinn sameiginlega evrópska markað þvert og endilangt og með aðstoð máltækni má sigrast á þessari hindrun en styðja um leið óhefta notkun einstakra tungumála. Ef við horfum enn lengra fram í tímann mun nýskapandi margmála evrópsk máltækni verða viðmiðun fyrir aðra í alþjóðasamfélaginu þegar þeir fara að virkja sín

eigin margmála samfélög. Líta má á máltækni sem eins konar ‘stuðningstækni’ sem aðstoðar okkur við að yfirstíga ‘fötlunina’ sem fylgir fjölbreytilegu tungumálaumhverfi og gerir málsamfélögin aðgengilegri hvert öðru.

Að lokum má nefna virkt rannsóknarsvið innan máltækninnar sem er notkun máltækni við björgunaraðgerðir á hamfarasvæðum, þar sem rétt framkvæmd getur skipt sköpum: Í framtíðinni gætu greind vélmenni búin hæfileikum til margmála málnotkunar bjargað mannslífum.

Ögranir sem máltækni stendur frammi fyrir

Þótt töluverðar framfarir hafi orðið í máltækni á síðustu árum er hraði tækniframfara og nýsköpunar í framleiðsluvörum enn of lítill. Sá máltæknibúnaður sem mest er notaður, svo sem málfræði- og stafrýnar ritvinnslukerfa, er venjulega einmála og þar að auki einungis til fyrir fáein tungumál. Þótt vélþýðingar á netinu séu gagnlegar til að fá þokkalega hugmynd um efni skjala glíma þær við alls kyns vandamál þegar þörf er á mjög nákvæmum og fullkomnum þýðingum. Vegna þess hve mannlegt mál er flókið er það bæði langt og dýrt ferli sem krefst langtíma fjármögnunar að skrifa hugbúnað sem líkir eftir mannlegu máli og prófa hann við eðlilegar kringumstæður. Til að halda brautryðjendahlutverki sínu í því að takast á við þær tæknilegu ögranir sem fylgja margmála samfélagi verður Evrópa því að beita nýjum aðferðum til að hraða þróuninni. Hér gæti bæði verið um að ræða framfarir í tölvutækni og aðferðir eins og múgvirkjun.

Núverandi hraði tæknilegra framfara er of lítill.

Máltaka manna og véla

Til að útskýra hvernig tölvur fást við tungumál og hvers vegna það er svo erfitt að forrita þær til þess skulum við líta sem snöggvast á það hvernig við tileinkum okkur móðurmálið og annað mál, og skoða síðan hvernig máltæknikerfin virka.

Mannfólkið lærir tungumál á tvo mismunandi vegu. Ungbörn læra móðurmál sitt með því að hlusta á samskipti foreldra sinna, systkina og annarra fjölskyldumeðlima. Um það bil tveggja ára gömul fara þau að mynda fyrstu orðin og stuttar setningar. Þetta er því aðeins mögulegt að börn hafa meðfæddan hæfileika til máls, og til að herma eftir því sem þau heyra og binda það í kerfi.

Mannfólkið öðlast málkunnáttu á tvo mismunandi vegu: Lærir af dæmum og lærir reglurnar sem liggja þar að baki.

Nám annars máls síðar á ævinni krefst meiri áreynslu, einkum vegna þess að nemandinn er ekki umlukinn málsamfélagi sem hefur málið að móðurmáli. Í skólum eru erlend mál venjulega numin með því að læra málfræðilega formgerð, orðaforða og stafsetningu með mynsturæfingum sem lýsa málfræðilegri kunnáttu í formi óhlutstæðra reglna, tafla og dæma. Nám erlends tungumáls verður erfiðara með aldrinum.

Hinar tvær megingerðir máltæknikerfa ‘nema’ tungumál á svipaðan hátt og mennirnir. Tölfræðilegar (eða gagnaknúnar) aðferðir afla málþekkingar úr gífurlega umfangsmiklum textasöfnum. En þótt nægjanlegt sé að nota texta á einu máli til að þjálfa t.d. stafrýna eru samhliða textar á tveim eða fleiri málum nauðsynlegir þegar kemur að þjálfun vélrænna þýðingarkerfa. Algrím vélræns náms ‘lærir’ þá mynstur sem sýna hvernig orð, orðasambönd og heilar setningar eru þýdd.

Hinar tvær megingerðir máltæknikerfa ‘nema’ tungumál á svipaðan hátt og mennirnir.

Þessi tölfræðilega nálgun getur krafist milljóna setninga og gæði útkomunnar aukast í réttu hlutfalli við magn greinds texta. Þetta er ein ástæða þess að þeir sem reka leitarvélar eru áfjáðir í að safna eins miklu af rituðu efni og hægt er. Stafrýnar í ritvinnslukerfum og netþjónustur eins og Google Search og Google Translate byggjast á tölfræðilegum aðferðum. Meginkostur tölfræðinálgunarinnar er sá að vélin lærir fljótt í

samfelldri röð þjálfunarferla, jafnvel þótt gæðin geti verið með ýmsu móti.

Hin meginaðferðin í máltækni og vélþýðingum er að smíða reglukerfi. Þá þurfa sérfræðingar á sviði málvísinda, tölvumálvísinda og tölvunarfræði fyrst að skrá málfræðigreiningu (þýðingarreglur) og búa til orðalista (orðasöfn). Þetta tekur langan tíma og kostar mikla vinnu. Reglukerfin krefjast einnig sérfræðiþekkingar. Sum helstu reglubyggðu vélþýðingarkerfin hafa verið í stöðugri þróun í meira en tuttugu ár. Meginkosturinn við reglukerfin er að sérfræðingarnir hafa meiri stjórn á málvinnslunni. Þetta gerir það mögulegt að laga kerfisbundið villur í hugbúnaðinum og veita notendum nákvæma endurgjöf, sérstaklega þegar reglukerfin eru notuð í tungumálanámi. En vegna þess hversu kostnaðarsöm þessi vinna er hefur reglubyggð máltækni til þessa einungis verið þróuð fyrir stærstu tungumálin.

Þar sem styrkleikar og veikleikar tölfræðilegu kerfanna og reglubyggðu kerfanna eru á mismunandi sviðum beinast rannsóknir um þessar mundir að blönduðum aðferðum sem tengja þessar tvær gerðir saman. Enn sem komið er hafa slíkar aðferðir þó ekki reynst eins vel í markaðshugbúnaði og á rannsóknarstofunum.

Eins og fram hefur komið í þessum kafla byggist alls kyns búnaður sem notaður er í upplýsingasamfélagi nútímans á máltækni. Í Evrópu á þetta sérstaklega við á sviði viðskipta og upplýsinga vegna þess hversu margmála málumhverfið þar er. En þrátt fyrir að máltækni hafi tekið miklum framförum á síðustu árum eru enn miklir möguleikar á því að auka gæði máltæknikerfa. Hér á eftir verður hlutverki íslenskunnar í evrópsku upplýsingasamfélagi lýst og mat lagt á stöðu máltækni fyrir íslensku.

Íslenska í evrópsku upplýsingasamfélagi

Almenn atriði

Um það bil 330 þúsund manns eiga íslensku að móðurmáli. Flestir búa á Íslandi[vii] en fjölmargir Íslendingar eru búsettir erlendis,[viii] svo sem annars staðar á Norðurlöndunum, á meginlandi Evrópu og í Norður-Ameríku. Þá er íslenska móðurmál sumra Vestur-Íslendinga af annarri og þriðju kynslóð[ix] en þeir eru flestir komnir um og yfir sjötugt. Á síðustu árum hefur innflutningur til landsins aukist til muna og þar með hefur þeim fjölgað sem tala íslensku sem erlent mál þótt sá hópur sé enn tiltölulega lítill.

Íslenska er notuð á öllum stigum stjórnsýslu, í skólakerfinu, viðskiptum og í öllum almennum samskiptum í landinu.

Íslensk tunga er notuð á öllum stigum stjórnsýslu, í skólakerfinu, í viðskiptum og öllum almennum samskiptum í landinu. Þótt ekki sé ákvæði um íslenska tungu í stjórnarskrá lýðveldisins hefur nýlega verið fest í lög að íslenska sé opinbert tungumál landsins.[x]

Lítið er um mállýskur í íslensku og vanalega er talað um smávægileg mállýskutilbrigði í framburði fremur en eiginlegar mállýskur. Lífseigast þessara mállýskutilbrigða er harðmælið þar sem lokhljóð eru fráblásin á milli sérhljóða á norðanverðu landinu en ófráblásin annars staðar, í orðum eins og *æpa*, *vita* og *taka*. Önnur mállýskuafbrigði eru smám saman að láta undan síga, svo sem raddaður framburður *l, m, n* á undan *p, t, k* í orðum eins og *úlpa*, *svampur*, *vanta*; vestfirskur einhljóðaframburður á undan *ng* og *nk* í orðum eins og *söngur*, *banki*, en í máli flestra er þar tvíhljóð; og hinn svokallaði *hv*-framburður þar sem borið er fram önghljóð í upphafi orða eins og *hver* þar sem flestir hafa lokhljóðið *k*.[xi] Á hinn bóginn virðist sem ný mállýskutilbrigði séu að myndast, svo sem tvinnhljóðun á *tj* þar sem *tjald* fer að hljóma eins og það væri *tsjald*.[xii]

Einungis er um minniháttar mállýskuafbrigði að ræða í setningagerð og fæst þeirra eru landshlutabundin. Þó virðast einstaka breytingar vera að gerast, sérstaklega í máli yngra fólks, og má þar nefna hina svokölluðu nýju þolmynd, eins og í *það var barið mig* í stað *ég var barin(n)*, svo og útvíkkaða notkun framvinduhorfs, *vera að*, eins og í *ég er ekki að skilja þetta* og *þeir voru að spila mjög vel*. Slík notkun heyrist varla hjá eldra fólki.

Íslenskuna sem töluð er í Vesturheimi má telja sérstaka mállýsku (eða mállýskur) enda hefur orðaforði þar þróast öðruvísi en á Íslandi. Þar má meðal annars nefna vestur-íslensku orðin *telefón* og *kar* (sbr. e. *telephone* og *car*) fyrir *sími* og *bíll*. Þá hafa orðmyndir og framburðaratriði stirðnað eða jafnvel aukist í vestur-íslensku en horfið að mestu eða öllu á Íslandi. Sem dæmi má nefna flámælið sem enn lifir góðu lífi meðal Vestur-Íslendinga.

Sérkenni íslenskrar tungu

Íslenska er norður-germanskt tungumál sem myndar vestur-norrænu málaættina ásamt færeysku og nýnorsku. Það er svokallað FSA-tungumál (eðlileg orðaröð frumlag-umsögn-andlag) og hefur sögnina jafnan í öðru (eða fyrsta) sæti setningar. Vegna ríkulegs beygingakerfis er orðaröð hins vegar tiltölulega frjáls; ákveðin orð geta staðið á ýmsum stöðum án þess að merking breytist. Eftirfarandi setningar hafa t.d. sömu merkingu þrátt fyrir að röð frumlags og andlags hafi verið snúið við.

Íslenska er FSA-tungumál þar sem sögnin er jafnan í öðru (eða fyrsta) sæti setningar en orðaröð þó tiltölulega frjáls.

- Hundurinn (nf.) beit köttinn (þf.).
- Köttinn (þf.) beit hundurinn (nf.).

Íslenska er meðal tiltölulega fárra tungumála þar sem frumlag setningar getur staðið í öðrum föllum en nefnifalli – oftast nær þágufalli en einnig þolfalli (og í nokkrum tilfellum eignarfalli). Í eftirfarandi setningum er

Frumlag setningar getur staðið í öðrum föllum en nefnifalli.

t.d. fornafnið í fyrstu persónu eintölu alltaf frumlagið, þrátt fyrir að standa í þremur mismunandi föllum:

- Ég (nf.) las bókina.
- Mig (þf.) vantar bókina.
- Mér (þgf.) líkar bókin.

Íslenskan er beygingamál og hefur fjögur föll, þrjú kyn og tvær tölur í nafnorðum, fornöfnum, lýsingarorðum og ákveðna (viðskeytta) greininum. Enginn óákveðinn greinir er notaður í málinu. Auk þessa beygjast lýsingarorð bæði veikt (ákveðið) og sterkt (óákveðið). Sagnir beygjast eftir persónu, tölu, tíð, hætti og mynd. Sagt er að íslenskan sé bræðingsmál sem þýðir að einstök ending er oft notuð fyrir fleiri en eina beygingarformdeild. Fjöldi beygingarflokka flækir svo kerfið enn, þannig að margar mismunandi endingar geta staðið fyrir sömu málfræðiformdeild eða formdeildasamsetningu, allt eftir því hver stofninn er.

Orðaforðinn er að mestu norrænn (germanskur) að uppruna þótt fjölmörg tökuorð hafi slæðst inn í málið á þeim ellefu öldum sem liðið hafa síðan land byggðist. Eftir kristnitöku árið 1000 voru t.d. fjölmörg orð tekin úr latínu og við siðaskiptin árið 1550 jukust áhrif frá þýsku með þýðingum á trúarritum og sálmum. Þá var Ísland undir danskri stjórn frá 1380 til 1944 og áhrif danskrar tungu frá þessum tíma eru augljós. Ýmis dönsk orð voru tekin inn í málið og mörg þeirra urðu hluti af íslensku. Þar má m.a. nefna orð eins og *gardínur* (*gardin* á dönsku) og *viskustykki* (*viskestykke* á dönsku).

Orðmyndun í íslensku er mjög virk.

Það er opinber stefna að ný orð skuli smíða úr íslenskum efnivið í stað þess að fá lánuð orð og hugtök úr erlendum málum. Þar sem margs konar hljóðavíxl eru algeng í íslensku má nota þau til þess að mynda nýtt orð af öðru, svo sem *leysni* af *lausn,* og einnig eru hin fjölmörgu viðskeyti málsins notuð til þess að mynda nýtt orð af rótum sem þegar eru til í málinu, svo sem *disk-lingur* af orðinu *diskur*. Algengast er þó að mynda ný orð með samsetningu tveggja eða fleiri sjálfstæðra orða, rétt eins og í *stafsetningar-orða-bók* og *umhverfis-mála-ráðu-neyti*. Þetta gerir tungumálið bæði lifandi og gagnsætt.

Framburður íslensku er tiltölulega gagnsær og að mestu hægt að segja fyrir um hann út frá stafsetningunni. Sá sem kann þær reglur sem gilda um vensl stafsetningar og framburðar ætti því að geta borið fram ný orð sem verða á vegi hans vandræðalaust, svo framarlega sem hann greinir réttilega orðhlutaskil en þau geta haft áhrif á framburð sumra orða. Reglur um áherslu orða eru einnig mjög einfaldar þar sem aðaláherslan fellur alltaf á fyrsta atkvæði og aukaáhersla kemur svo vanalega á annað hvert atkvæði eftir það, þótt það eigi ekki alltaf við í samsettum orðum.

Ritmálið byggist á latneska stafrófinu en þó eru notaðir í íslensku nokkrir stafir sem ekki þekkjast t.d. í ensku. Þetta eru stafirnir Þ/þ (einungis notaður í íslensku þótt upprunann megi rekja til fornensku), Ð/ð (einnig notaður í færeysku), Æ/æ (einnig notaður í norsku, dönsku og færeysku) og Ö/ö (einnig notaður í sænsku, finnsku, eistnesku, þýsku og ungversku). Að auki eru notaðir í íslensku sex broddstafi fyrir ákveðna sérhljóða: Á/á, É/é, Í/í, Ó/ó, Ú/ú og Ý/ý.

Ritaða málið hefur breyst tiltölulega lítið frá fornnorrænu sem gerir Íslendingum það kleift með nokkurri þjálfun að lesa fornískenska texta. Meginbreytingar á stafsetningu á undanförnum áratugum hafa verið niðurfelling setunnar (sem þó er enn notuð í fáeinum eiginnöfnum og ættarnöfnum eins og *Zóphónías* og *Haralz*) og upptaka *é* í stað *je*.

Nýleg þróun

Frá hernámi Breta og síðar Bandaríkjamanna í heimstyrjöldinni síðari hefur íslenskan orðið fyrir mun sterkari áhrifum frá ensku en dönsku og þau áhrif hafa aukist að mun við innreið tónlistar, kvikmynda og sjónvarpsefnis frá Bretlandi og Bandaríkjunum. Vöxtur netsins hefur einnig aukið áhrif ensku á íslensku, enda eru um 95% þjóðarinnar netvædd.

Áhrif frá ensku eru augljósust í fjölda tökuorða úr ensku í íslensku en fæst þessara orða er þó að finna í orðabókum og þau sjást sjaldan á prenti. Þau eru að auki oft litin hornauga af málræktarmönnum. Notkun þeirra einskorðast því að mestu við talað mál og að auki má finna þau í óopinberum og persónulegum skrifum, svo sem í tölvupósti, á bloggsíðum o.s.frv.

Ensk áhrif á málkerfið virðast þó óveruleg. Mörg tökuorðanna sem notuð eru hversdagslega fá íslenskar endingar þótt nokkur þeirra beygist ekki. Þar má nefna *næs* (úr e. *nice*), *kúl* (úr e. *cool*), o.s.frv. Stundum er því haldið fram að sumar breytingar í setningagerð og hljóðkerfi íslenskunnar, svo sem hið útvíkkaða framvinduhorf og tvinnhljóðunin á *tj* sem áður eru nefndar, megi rekja til enskra áhrifa, en um það er þó deilt.

Á undanförnum árum hefur mikið verið rætt um svokallað umdæmistap á Íslandi eins og í mörgum öðrum löndum. Íslenskur vinnumarkaður hefur orðið sífellt alþjóðlegri á síðustu árum – íslensk fyrirtæki starfa erlendis og erlend fyrirtæki starfa á Íslandi. Ensk tunga er því hluti af daglegu starfi þessara fyrirtækja og fundir og bréfleg samskipti fara iðulega fram á ensku. Þá er það orðið algengt að ársskýrslur þessara fyrirtækja, vefsíður og annað efni, séu að hluta eða öllu á ensku. Einnig virðist það vera hálfgerð tíska að íslensk fyrirtæki beri enskt nafn, ýmist eingöngu eða að hluta. Þannig má sjá nöfn eins og *Icelandair, Actavis, Baugur Group* og *Stoðir Invest.*[xiii]

Annað svið atvinnulífsins þar sem ensk tunga er áberandi er upplýsingatækni, en um hana verður betur rætt í öðrum aðalkafla þessa bæklings.

Íslensk málrækt

Í íslenskri málrækt hefur áhersla löngum verið lögð á bæði varðveislu og eflingu íslenskrar tungu. Þetta má sjá greinilega á þeirri vinnu sem lögð hefur verið í uppbyggingu orðaforðans með starfsemi ýmissa íðorðanefnda. Þær eru vanalega skipaðar sjálfboðaliðum úr ýmsum fræði- og atvinnugreinum en málræktarsvið Stofnunar Árna Magnússonar í íslenskum fræðum styður við starf þeirra. Íslensk málnefnd var stofnuð 1964[xiv] en meginhlutverk hennar er að vera stjórnvöldum, og þá einkum mennta- og menningarmálaráðuneytinu, til ráðgjafar um íslenska tungu og íslenska málstefnu auk þess að semja árlega ályktun um stöðu tungunnar. Íslensk málnefnd ber ábyrgð á þeim stafsetningarreglum sem auglýstar eru af menntamálaráðuneytinu og notaðar eru í skólakerfinu. Nefndin beitti sér fyrir stofnun Málræktarsjóðs en hlutverk hans er að „beita sér fyrir og styðja hverskonar starfsemi til eflingar íslenskri tungu og varðveislu hennar“.[xv]

Íslensk málnefnd er stjórnvöldum til ráðgjafar um íslenska tungu og íslenska málstefnu.

Stundum er sagt að allir Íslendingar séu málfræðingar. Bændur og sjómenn, hjúkrunarfræðingar og kennarar hringja í útvarpsstöðvar og Stofnun Árna Magnússonar í íslenskum fræðum og ræða hnökra á málinu og kvarta undan málvillum. Fólk hefur einlægar áhyggjur af stöðu tungunnar í landinu og heilmiklar umræður fara fram um það hvernig best sé að varðveita tungumálið og jafnvel hvort sú varðveisla sé ómaksins verð. Þó líta flestir Íslendingar á tungumálið sem kjarna íslenskrar menningar og íslenskrar sjálfsmyndar og því hefur mikið starf verið unnið í þeim tilgangi að varðveita tungumálið sem best.

Miðstöð íslenskrar málræktar er í *Stofnun Árna Magnússonar í íslenskum fræðum* en meginhlutverk stofnunarinnar er að „vinna að rannsóknum í

íslenskum fræðum og skyldum fræðigreinum, einkum á sviði íslenskrar tungu og bókmennta, að miðla þekkingu á þeim fræðum og varðveita og efla þau söfn sem henni eru falin eða hún á".[xvi] Stofnunin skiptist í nokkrar deildir sem sinna mismunandi þáttum íslensks tungumáls, bókmennta og menningar, svo sem málrækt, orðfræði, máltækni, nafn- og örnefnafræði, handritafræði, þjóðsögum og alþjóðlegum tengslum.

Ríkisútvarpið hefur löngum leikið stórt hlutverk í varðveislu tungunnar, bæði vegna eigin málstefnu og vinsælla útvarpsþátta um tunguna og orðaforðann.

Ríkisútvarpið hefur löngum leikið stórt hlutverk í varðveislu tungunnar, ekki aðeins vegna eigin málstefnu heldur einnig vegna vinsælla útvarpsþátta áður fyrr, eins og *Íslensks máls* og *Daglegs máls* þar sem málfræðingar ræddu um tunguna og orðaforðann, og *Orð skulu standa,* þar sem tvö lið kepptust um að finna rétta merkingu sjaldgæfra orða og hugtaka. Almennt gegna fjölmiðlarnir mikilvægu hlutverki í verndun íslenskrar tungu.

Tuttugu og tvær útvarpsstöðvar eru í landinu og talað mál í þeim öllum er að mestu leyti á íslensku þótt enskan sé yfirgnæfandi í tónlistinni sem leikin er. Að auki eru í landinu tíu sjónvarpsstöðvar og þótt meiri hluti þess efnis sem sjónvarpað er sé á erlendum tungumálum er staða íslenskunnar sterk.[xvii] Allt erlent sjónvarpsefni er textað á íslensku ─ fyrir utan sumt barnaefni sem er talsett ─ og þegar um beinar útsendingar er að ræða frá erlendum stórviðburðum segir íslenskur þulur vanalega frá því helsta sem er að gerast.[xviii]

Dagur íslenskrar tungu hefur verið haldinn hátíðlegur á fæðingardegi Jónasar Hallgrímssonar, 16. nóvember, síðan 1996 og er honum ætlað að efla umræður um íslenska tungu.[xix]

Íslenska í menntakerfinu

Íslensk tunga er mikilvægur þáttur í skólakerfinu og nemendur í 1.-4. bekk grunnskóla verja að lágmarki 1.120 mínútum á viku í íslenskt mál og bókmenntir. Í 5.-7. bekk hefur þessum tímum fækkað í 680 mínútur á viku og síðan 630 mínútur á viku í 8.-10. bekk en það er töluvert minna en aðrar Norðurlandaþjóðir verja í móðurmálskennslu.[xx] Í framhaldsskóla er einnig minni tíma varið til móðurmálskennslu en annars staðar á Norðurlöndunum, eða að lágmarki 20 einingum af þeim 200 sem krafist er til stúdentsprófs.[xxi]

Í PISA-könnununum sem gerðar hafa verið frá árinu 2000 fór lesskilningur íslenskra ungmenna, sérstaklega drengja, stöðugt minnkandi. Í könnuninni 2009 hafði ástandið hins vegar batnað nokkuð og Ísland var þar í ellefta sæti og í svipaðri stöðu og aðrar Norðurlandaþjóðirnar að Finnum frátöldum.[xxii]

Háskóli Íslands er eini háskólinn þar sem hægt er að taka doktorspróf í íslensku en meistarapróf í málinu er hægt að taka frá Manitobaháskóla í Kanada auk Háskóla Íslands. Þó nokkrir háskólar víða um heim bjóða upp á B.A.-próf í íslensku.

Aðeins tveir af þeim sjö háskólum sem í landinu eru hafa sérstaka málstefnu þar sem íslenska er tilgreind sem opinbert mál háskólans. Enska er sífellt meira notuð í starfi háskólanna þar sem erlendum kennurum hefur fjölgað og þar að auki stefna allir háskólarnir að því að fjölga erlendum nemendum. Vegna þessa fer námskeiðum sem kennd eru á ensku fjölgandi, sem og doktorsritgerðum skrifuðum á því máli. Þá hefur það aukist að íslenskir fræðimenn skrifi fræðigreinar sínar á ensku og námsefni í skólunum er æ meir á enskri tungu.[xxiii]

Með því að fjölga íslenskutímum í skólum landsins má bæta íslenskukunnáttu nemenda og búa þá þannig betur undir virka þátttöku í íslensku samfélagi

Með því að fjölga íslenskutímum í skólum landsins má bæta íslenskukunnáttu nemenda og búa þá þannig betur undir virka þátttöku í íslensku samfélagi. Máltækni gæti verið hjálpleg í þessu sambandi enda gefur hún möguleika á tölvustuddu tungumálanámi sem gerir nemendum kleift að njóta tungumálsins á skemmtilegan hátt, t.d. með því að tengja orðaforða í

ákveðnum texta við skilgreiningar á orðunum eða við hljóðskrá eða myndband með viðbótarupplýsingum, svo sem framburði orðanna.

Alþjóðlegir þættir

Ísland er lítið land og í raun aðeins örríki í samfélagi þjóðanna, og því eru áhrif íslenskra lista, vísinda og fræða erlendis að vonum aðeins smávægileg. Örfáir íslenskir tónlistarmenn hafa náð vinsældum utan landsins, svo sem *Björk, SigurRós* og *Gus Gus*, en þar sem tónlist þeirra er meira og minna sungin á ensku gerir hún lítið til þess að auka hróður tungumálsins utan landsteinanna. Það sama má segja um velgengni íslenskra rithöfunda erlendis sem hefur kynnt íslenska menningu fyrir öðrum þjóðum en ekki beinlínis íslenska tungu. Hins vegar hafa vinsældir íslenskra tónlistarmanna og rithöfunda, uppgangur – og fall – íslenskra banka og fyrirtækja erlendis, svo og áherslur Íslands á umhverfisvæna orku vakið athygli annarra þjóða á Íslandi og skilað sér í aukinni umfjöllun um landið í erlendum fjölmiðlum og fjölgun ferðamanna til landsins. Íslendingasögurnar, víkingarnir og íslenski hesturinn eru því ekki lengur einu íslensku fjársjóðirnir sem heilla útlendinga.

Áhugi á íslensku á alþjóðavettvangi fer vaxandi.

Íslensk tunga hefur lítil áhrif á önnur tungumál og aðeins örfá íslensk orð hafa ratað sem tökuorð inn í önnur mál. Þar eru langalgengastar orðmyndir dregnar af eiginnafninu *Geysir* sem í mörgum málum tákna goshver. Þannig er það *geyser* í ensku, frönsku, gallísku og ítölsku; *geiser* í afríkaans, basknesku, hollensku, eistnesku, frísnesku og þýsku; *gejser* í dönsku og sænsku; *geysir* á norsku; *geizers* á lettnesku; *geizeris* á litháísku; *gheizer* á rúmensku; *gejzir* á króatísku; *gejzír* á tékknesku, ungversku og slóvakísku; *gejzer* á pólsku; *géiser* á portúgölsku og spænsku; og *gayzer* á tyrknesku. Þá er enska orðið *eider* tökuorð úr íslensku, komið af orðinu *æður*, og íslenska orðið *tölt* er almennt notað erlendis um fimmta gang íslenska hestsins.

Aukinn áhugi á íslenskri tungu og menningu kemur greinilega fram í vaxandi fjölda þeirra nemenda sem stunda íslenskunám, ýmist á Íslandi eða í öðrum löndum. Við Háskóla Íslands jókst fjöldi erlendra nema í íslenskunámi um nærri 100% milli áranna 2005 og 2007 og árið 2008 bauð Háskólinn í fyrsta sinn upp á námsleið í hagnýtri íslensku ætlaða þeim sem vilja læra tungumálið án þess að leggja áherslu á hinn akademíska þátt námsins. Íslenska er nú kennd í um 40 háskólum utan Íslands og styrkir Ísland 18 þeirra fjárhagslega.[xxiv] Þá er boðið upp á sjálfstæð íslenskunámskeið í fjölmörgum löndum, svo sem í fyrrum Íslendingabyggðum Kanada og Bandaríkjanna, og á milli 300 og 400 manns fara daglega inn á heimasíðu *Icelandic Online*.[xxv]

Íslensk tunga er hvergi gjaldgeng í alþjóðlegum samskiptum en því hefur verið haldið fram að staða málsins myndi styrkjast á alþjóðavettvangi ef landið gengi í Evrópusambandið,[xxvi] þar sem íslenska yrði þar með eitt af opinberum tungumálum sambandsins.[xxvii] Einnig er hægt að nýta máltækni til að bregðast við þeirri ógn sem stafar af ensku með því að þróa vélþýðingar og margmála upplýsingaheimt og hjálpa þannig til við að lágmarka það persónulega og efnahagslega óhagræði sem felst í því að hafa ekki ensku að móðurmáli.

Staða íslensku myndi væntanlega styrkjast á alþjóðavettvangi ef landið gengi í Evrópusambandið.

Íslenska á netinu

Í júní 2010 höfðu um það bil 95% þjóðarinnar aðgang að netinu[xxviii] og í aldurshópnum 35-44 ára var hlutfallið allt að 100%. Í byrjun maí 2011 voru 197.000, eða 61,8% þjóðarinnar, skráðir notendur Facebook.[xxix]

Árið 2010 voru 25.000 skráð *.is* lén[xxx] og um það bil 5.600 lén voru á landinu fyrir utan *.is* kerfið.[xxxi] Fjöldi vefsetra er talinn í kringum 7.500 en þar eru þó hvorki taldar bloggsíður innan *.is* léna né vefir á erlendum lénum eins og *blogspot.com* og *wordpress.com*.

Næstum allir Íslendingar nota Netið.

Netið er orðið svo vinsælt að árið 2010 gerðist það í fyrsta sinn að auglýsendur eyddu meiri peningum í auglýsingar á netinu en í prentmiðlunum.[xxxii] Slíkt hefur reyndar ekki enn gerst á Íslandi en virðist þó stefna í þá átt. Af sjö vinsælustu vefjunum á Íslandi eru þrír fréttamiðlar (*mbl.is, visir.is, pressan.is*). Netið hefur einnig að miklu leyti tekið við af símaskránni þar sem upplýsingasíðan *ja.is* er fimmta mest notaða síða landsins. Aðrar vinsælar síður eru *Google*, *Facebook* og *YouTube*.[xxxiii] Bæði Google og Facebook bjóða nú upp á íslenskt notendaviðmóti.

Vöxtur Netsins skiptir miklu máli fyrir máltækni.

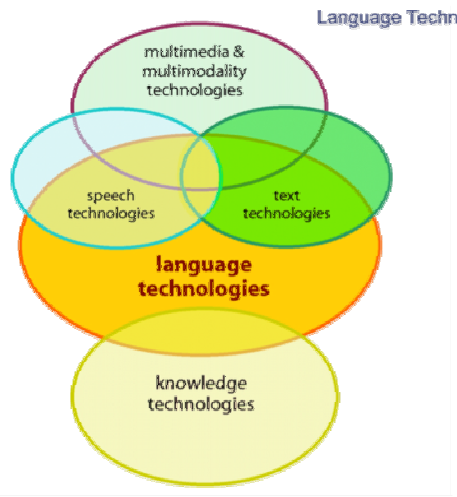
Vöxtur netsins er mikilvægur fyrir máltækni að tvennu leyti. Annars vegar er fjöldi texta á stafrænu formi algjör gullnáma þegar kemur að greiningu á notkun tungumála, og þá sérstaklega þegar safna þarf tölfræðilegum upplýsingum. Hins vegar býður netið upp á fjöldann allan af notkunarsviðum fyrir máltækni.

Hægt er að nýta máltækni til að þýða vefsíður sjálfvirkt milli tungumála.

Leitarvélar eru án efa mest notaði hugbúnaðurinn á netinu en þær nýta margs konar sjálfvirka málvinnslu eins og við munum sjá í síðari hluta þessa bæklings. Þar er um að ræða margbrotna máltækni sem er breytileg eftir tungumálum. Í íslensku þarf til dæmis að taka tillit til mismunandi beygingarendinga nafnorða, lýsingarorða og sagna, svo og mismunandi stofna, eins og í orðunum *svartur* og *svört*. Notendur netsins geta einnig nýtt máltækni á annan hátt, svo sem með sjálfvirkum þýðingum vefsíðna á mörg tungumál. Þegar litið er á gríðarlegan kostnað við mennska þýðingu þessa efnis vekur furðu hversu lítið hefur verið gert til að þróa slíkan þýðingarbúnað. Ástæðuna má ef til vill rekja til þess hversu margslungin íslensk tunga er í raun, svo og hversu fjölbreytta tækni þarf til að smíða dæmigerðan máltæknibúnað.

Í næsta kafla er að finna yfirlit um máltækni og helstu afurðir hennar en einnig er kynnt mat á stöðu máltækni fyrir íslensku.

Máltæknilegur stuðningur við íslensku



Undir máltækni falla m.a. hugbúnaðarkerfi sem hönnuð eru til þess að vinna með mannlegt mál. Tungumál eru bæði rituð og töluð en þótt talmálið hafi þróast á undan og sé þannig eðlilegasta form mállegra samskipta eru margbrotnar upplýsingar og mestöll mannleg þekking geymd og miðlað í rituðu máli. Til að vinna með og framleiða þessi mismunandi form tungumálsins höfum við annars vegar taltækni og hins vegar textatækni, en báðar nýta orðasöfn, málfræðireglur og merkingarfræði. Þetta þýðir að máltækni tengir tungumálið við mismunandi form þekkingar, óháð því hvernig henni er miðlað (í tali eða texta). Myndin hér hægra megin sýnir hvernig hinar ýmsu greinar máltækni tengjast. Í öllum samskiptum tengjum við tungumálið öðrum samskiptaháttum og upplýsingamiðlum – tali getur fylgt látbragð og andlitstjáning. Stafrænir textar tengjast myndum og hljóði. Í kvikmyndum getur komið fram bæði talað og ritað mál. Tal- og textatækni skarast því og fléttast saman við margs konar aðra tækni sem greiðir fyrir úrvinnslu fjölhátta samskipta og margmiðlunargagna.

Hér á eftir verður fjallað um meginverksvið máltækni, þ.e. málrýni, vefleit, taltækni og vélþýðingar. Undir þetta fellur verkbúnaður og grundvallartækni eins og:

- stafrýni
- ritstoð
- tölvustutt tungumálanám
- upplýsingaheimt
- útdráttur upplýsinga
- samantekt texta
- spurningasvörun
- talkennsl
- talgerving

Áður en þessum notkunarsviðum og búnaði verða gerð skil munum við lýsa stuttlega högun dæmigerðs máltæknikerfis.

Högun máltæknibúnaðar

Í dæmigerðum hugbúnaði til málvinnslu felast nokkrar einingar sem endurspegla mismunandi þætti tungumálsins. Myndin hér til hægri sýnir mjög einfaldaða byggingu ritvinnslukerfis. Þrjár fyrstu einingarnar snúa að gerð og merkingu ílagstextans:

1. Forvinnsla: hreinsun gagna, afnám sniðs, greining ílagstungumáls, *þ* e.t.v. skipt út fyrir *th* í íslensku, o.s.frv.
2. Málfræðigreining: sögnin fundin svo og andlög hennar, enn fremur ákvæðisorð, o.s.frv.: setningagerð greind.
3. Merkingargreining: einræðing orða (fundið út hver er merking orðsins í tilteknu samhengi); greining endurvísunar (t.d. hvaða fornafn vísar til hvaða nafnorðs í setningunni) og staðgengla; og merking setningarinnar sýnd á þann hátt að tölva geti lesið hana.

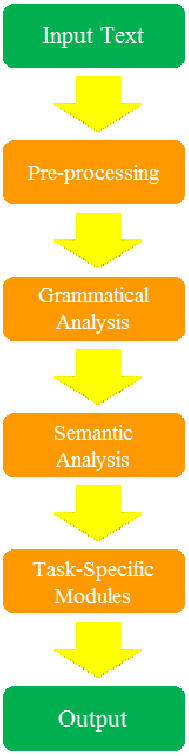


Figure 2: A Typical Text Processing Application Architecture

Eftir greiningu textans geta verkbundnar einingar séð um ýmsar aðrar aðgerðir, svo sem sjálfvirka samantekt ílagstexta og uppflettingu í gagnagrunni. Þetta er einfölduð lýsing á uppbyggingu verkbúnaðarins en gefur þó innsýni í það hversu flókinn máltæknibúnaður er.

Að lokinni kynningu á helstu verksviðum máltækninnar verður gefið stutt yfirlit yfir núverandi stöðu máltæknirannsókna og máltæknimenntunar, og

að lokum drepið á rannsóknaverkefni sem ýmist er lokið eða eru í gangi. Síðan verður gerð grein fyrir mati sérfræðinga á stöðu helstu máltæknitóla og málfanga út frá ýmsum mælikvörðum, s.s. aðgengi, þroska og gæðum. Heildarstaða máltækni fyrir íslensku er að lokum dregin saman í töflu.

Helstu verksvið

Í þessum kafla verður fjallað um mikilvægustu máltæknitól og málföng, og gefið yfirlit yfir máltækni á Íslandi. Máltæknitól og málföng sem eru undirstrikuð í textanum er einnig að finna í töflu í lok þessa kafla.

Máltæknitól og málföng sem eru undirstrikuð í textanum er einnig að finna í töflu í lok þessa kafla.

Málrýni

Flestir sem hafa unnið með ritvinnslukerfi eins og Microsoft Word vita að í því er stafrýnir sem bendir á stafsetningavillur og stingur upp á leiðréttingum. Fyrstu stafrýnarnir báru orðin í textanum saman við safn rétt ritaðra orða. Nú er þessi hugbúnaður mun þróaðri. Með því að nota sérhæfð algrím til textagreiningar má greina villur í beygingu (svo sem ranga eignarfallsendingu) og setningagerð, eins og þegar sögnina vantar eða þegar ósamræmi er á milli sagnar og frumlags (t.d. *ég *skrifar bréf*). Hins vegar munu fæstir stafrýnar finna villur í eftirfarandi dæmum:

Ég var um þetta leiti á næsta leyti.

Hún segir að móðir sýn hafi aðra sín á málið.

Hann þótti hafa stirt stöðu sína.

Til þess að hægt sé að fást við slíkar villur þarf að greina samhengi textans, t.d. þegar ákveða skal hvort íslenskt lýsingarorð eigi að vera með einu n-i (kvenkyn) eða tveim (karlkyni), eins og í eftirfarandi dæmi:

Hann er farinn.

Hún er farin.

Greining slíkra villna byggist ýmist á sérstakri málfræðilýsingu fyrir hvert tungumál, sem mikinn tíma og sérþekkingu þarf til að fella inn í hugbúnaðinn, eða á tölfræðilegu mállíkani. Slíkt líkan reiknar líkurnar á því að tiltekið orð birtist í ákveðnu umhverfi (t.d. eftir því hvaða orð fara á undan og á eftir). Til dæmis er *hann er farinn* líkleg orðaruna en *hún er farinn* er það ekki. Tölfræðilegu mállíkani af þessu tagi má koma upp á sjálfvirkan hátt með því að nota mikið af (réttum) málgögnum (málheild). Báðar aðferðirnar hafa einkum verið þróaðar fyrir ensk málföng og það er ekki auðvelt að yfirfæra þær á íslensku sem hefur sveigjanlegri orðaröð, ótakmarkaða möguleika á samsetningu orða og ríkulegra beygingarkerfi.

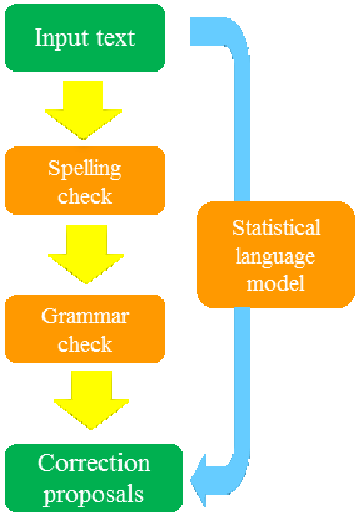


Figure 3: Language Checking (left: rule-based; right: statistical)

Málrýni er ekki bundin við ritvinnslukerfi; hún er líka notuð í ritstoðarkerfum, þ.e. hugbúnaðarumhverfi til að skrifa handbækur og önnur rit eru skrifuð samkvæmt ákveðnum stöðlum fyrir flókna upplýsingatækni, heilbrigðisgeirann, verkfræði og fleira. Af ótta við kvartanir og skaðabótakröfur viðskiptavina vegna rangrar notkunar sem rekja má til illskiljanlegra leiðbeininga leggja fyrirtæki sífellt meiri áherslu á gæði tæknilegra leiðbeininga, á sama tíma og þau stefna á alþjóðlegan markað (með þýðingum og staðfærslu). Framfarir í málvinnslu hafa leitt til þróunar á ritstoðarbúnaði sem aðstoðar höfunda tæknilegra leiðbeininga við að velja orð og setningagerð sem samræmast iðnaðarreglum og skorðum fyrirtækja á notkun íðorða.

Málrýni er ekki bundin við ritvinnslukerfi; hún er líka notuð í ritstoðarkerfum.

Stafrýnir hefur verið til fyrir íslensku frá því seint á níunda áratugnum þegar Friðrik Skúlason ehf. (Frisk Software) þróaði stafsetningaforritið

Púka. Forritið hefur síðan verið uppfært og endurbætt. Það er til fyrir MS Office og er mikið notað. Aðrir stafrýnar hafa einnig verið hannaðir. Árið 2002 þróaði hollenska fyrirtækið Polderland stafrýni fyrir MS Office og einnig er til stafrýnir í opnum hugbúnaði fyrir GNU/Linux forrit sem byggð eru á Aspell. Þessi forrit skoða eingöngu stök orð og ráða því ekki við margar algengar stafsetningarvillur. Frumgerð að samhengisháðum stafrýni hefur verið felld inn í LanguageTool[xxxiv] og vinnur með Open Office. Sá stafrýnir gæti hugsanlega myndað grunninn að málfræðirýni, en slíkt forrit er ekki til fyrir íslensku.

Fyrir utan stafrýna og ritstoð er málrýning einnig mikilvæg fyrir tölvustutt tungumálanám og henni er einnig beitt við sjálfvirka leiðréttingu á fyrirspurnum sem sendar eru vefleitarvélum eins og tillögukerfi Google *‚Áttirðu við:‘*

Vefleit

Leit á vefnum, svo og á innri netum og í stafrænum bókasöfnum, er væntanlega það svið þar sem máltækni er mest notuð nú á dögum, en er þó fremur skammt á veg komin. Leitarvélin Google, sem komið var á laggirnar 1998, er nú notuð í 80% allra vefleita í heiminum.[xxxv] Síðan 2004 hefur sögnin *gúg(g)la* verið notuð í íslensku þótt hún hafi ekki enn komist í prentaðar orðabækur. Hvorki leitarviðmót Google né framsetning niðurstaðna hefur tekið grundvallarbreytingum frá fyrstu útgáfu. Í nýjustu útgáfu býður Google reyndar upp á leiðréttingar á ranglega stafsettum orðum og hefur nú bætt við merkingarlegum leitarmöguleikum sem geta bætt nákvæmni leitarinnar með því að greina merkingu orða í samhengi leitarorðsins.[xxxvi] Velgengni Google sýnir að með stóru gagnasafni og skilvirkum aðferðum við að lykla gögnin getur tölfræðileg aðferð skilað viðunandi niðurstöðum.

Þegar um flóknari upplýsingaleit er að ræða er nauðsynlegt að nýta dýpri málfræðiþekkingu til textatúlkunar. Tilraunir með orðaföng eins og tölvutæk samheitasöfn og verufræðileg málföng (s.s. WordNet fyrir ensku og GermaNet fyrir þýsku) hafa sýnt verulega bættan árangur í að finna síður þar sem samheiti við leitarorðið koma fyrir, svo sem *hagnaður*, *arður*, *gróði* og *ábati* eða jafnvel fjarskyldari orð.

Næsta kynslóð leitarvéla verður að vera útbúin mun þróaðri máltækni, einkum til að ráða við leitartexta í formi spurningar eða annars konar setningar í stað einstakra leitarorða. Til að bregðast við fyrirspurninni „Láttu mig fá lista yfir öll fyrirtæki sem voru yfirtekin af öðrum fyrirtækjum síðustu fimm árin“ þarf máltæknikerfið að greina setningagerð og merkingu fyrirspurnarinnar og hafa atriðisorðaskrá til að kalla fram viðeigandi skjöl á fljótvirkan hátt. Til að unnt sé að gefa viðunandi svar þarf að beita setningalegri þáttun til greiningar á málfræðilegri formgerð setningarinnar og greina að verið sé að leita að fyrirtækjum sem hafa verið yfirtekin en ekki þeim fyrirtækjum sem tóku yfir önnur fyrirtæki. Þá þarf að skilgreina sambandið *síðustu fimm ár* svo hægt sé að ákvarða við hvaða ár er átt. Að lokum þarf að máta leitarfyrirspurnina við ógrynni af óskipulögðum gögnum svo finna megi upplýsingarnar sem leitað er að. Þetta er kallað *upplýsingaheimt* og felur í sér leit að skjölum og vægisröðun þeirra. Til þess að hægt sé að búa til lista yfir fyrirtæki þarf kerfið einnig að þekkja ákveðinn orðastreng í skjali sem nafn fyrirtækis, en það ferli kallast *nafnakennsl*.

Enn meiri áskorun felst í því að máta leitarfyrirspurnina við skjöl á öðrum tungumálum. Þvermála upplýsingaheimt felur í sér sjálfvirka þýðingu leitarfyrirspurnar yfir á öll möguleg tungumál og síðan þýðingu niðurstaðnanna aftur yfir á markmálið.

Nú er gögn í auknum mæli að finna á öðru sniði en sem texta og því er orðin til þörf á þjónustu sem gefur kost á margmiðlunarupplýsingaheimt

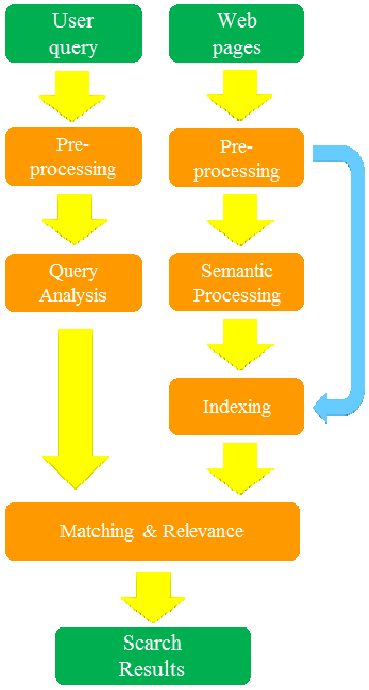


Figure 4: Web Search Architecture

Næsta kynslóð leitarvéla verður að vera útbúin mun þróaðri máltækni.

með því að leita að myndum, hljóði eða myndböndum. Þegar um er að ræða hljóð- og myndbandsskrár þarf sérstök talkennslaeining að breyta tali í texta (eða hljóðritun) sem síðan er hægt að máta við leitarfyrirspurnina.

Í beygingarmálum eins og íslensku er mikilvægt að hægt sé að leita að öllum beygingarmyndum orðs í einu í stað þess að þurfa að leita að hverri mynd sérstaklega. Þetta má gera með aðstoð gagnagrunnsins Beygingarlýsing íslensks nútímamáls, BÍN, sem þróaður hefur verið á Stofnun Árna Magnússonar í íslenskum fræðum. Gagnagrunnurinn hefur að geyma um það bil 280.000 beygingardæmi með meira en 5.8 milljónum beygingarmynda. Hver færsla inniheldur nefnimyndina, orðmyndina, orðflokkinn og beygingarþætti nafnorða, sérnafna, lýsingarorða, sagna og atviksorða.

Fyrir nokkrum árum þróaði einkafyrirtækið Spurl leitarvélina *Emblu* sem nýtti þennan gagnagrunn. Sama algrím er notað við leit í íslensku símaskránni og á nokkrum öðrum síðum. Google leitarvélin er nú búin svipuðum hæfileikum, en þó ekki eins margþættum. Engin sérhæfð leitarvél fyrir íslensku er til eins og er, fyrir utan íslenska viðmótið í Google, og ekki er verið að vinna að slíkri leitarvél.

Taltækni

Taltækni er notuð til að smíða viðmót sem gerir notandanum kleift að tala við tölvuna í stað þess að nota tölvuskjáinn, lyklaborð og mús. Nú á dögum nýta fyrirtæki raddstýrð notendaviðmót í ýmiss konar sjálfvirkri og hálfsjálfvirkri símaþjónustu við viðskiptavini, starfsmenn eða viðskiptafélaga. Helstu atvinnugreinar sem nýta slík raddstýrð viðmót eru bankarstarfsemi, birgjar, almenningssamgöngur og fjarskiptafyrirtæki. Taltækni má t.d. einnig nota í viðmóti leiðsögutækja í bílum og í stað myndræns viðmóts og snertiskjáa sem notendaviðmót í snjallsímum.

Taltækni er notuð til að smíða viðmót sem gerir notandanum kleift að tala við tölvuna í stað þess að nota tölvuskjáinn, lyklaborð og mús.

Taltækni byggist á ferns konar grundvallartækni:

1. Sjálfvirk talkennsl ákvarða hvaða orð notandinn segir í tiltekinni segð.
2. Málskilningur greinir setningafræðilega formgerð segðarinnar og túlkar hana út frá viðkomandi kerfi.
3. Samræðustjóri ákvarðar hvað þarf að gera út frá ílagi notandans og virkni kerfisins.
4. Talgerving breytir svari kerfisins í hljóð sem notandinn nemur.

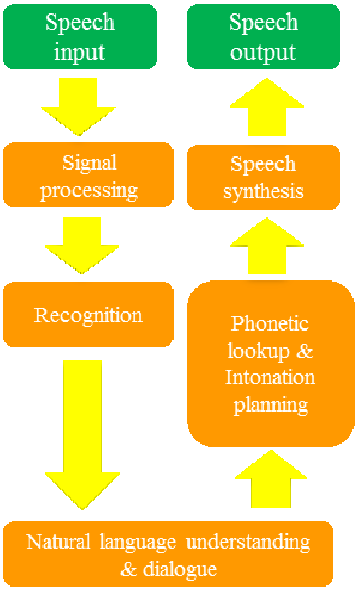


Figure 5: Simple Speech-based Dialogue Architecture

Eitt erfiðasta viðfangsefni talkennslabúnaðar er að greina rétt þau orð sem notandinn segir. Því þarf annaðhvort að takmarka hugsanlegar segðir notandans við afmarkað mengi lykilorða eða byggja upp mállíkön sem ná yfir stóran hluta segða í eðlilegu máli. Með vélrænum námsaðferðum er líka hægt að koma upp mállíkönum á sjálfvirkan hátt úr talmálsheildum, stórum söfnum hljóðskráa með textaumritun. Takmörkun leyfilegra segða leiðir venjulega til þvingaðrar notkunar á talviðmótinu og getur haft þau áhrif að notendur taki því ekki vel; en smíði viðamikils mállíkans, finstilling þess og viðhald eykur kostnaðinn við kerfið verulega. Raddstýrð notendaviðmót sem nýta mállíkan og gefa notandanum sveigjanleika í því hvernig hann ber fram erindi sitt í byrjun — t.d. heilsa með *Hvað get ég gert fyrir þig?*— eru yfirleitt sjálfvirk og fá jákvæðari viðbrögð notenda.

Fyrirtæki nota yfirleitt upptökur með lestri atvinnumanna til að mynda frálag talviðmótsins. Í stöðluðum segðum þar sem orðalagið er ekki háð tilteknu samhengi eða ákveðnum notanda getur þetta verið fullkomlega nóg til að notandinn sé sáttur. En þegar segðirnar eru breytilegar getur tónfallið orðið óeðlilegt vegna þess að bútar úr mismunandi hljóðskrám eru tengdir saman. Talgervlar eru sífellt að verða betri í því að skila breytilegum segðum sem hljóma eðlilega, en þó má enn bæta þá.

Viðmót á taltæknimarkaðnum hafa verið stöðluð umtalsvert á undanförnum áratug að því er snýr að hinum ýmsu tæknieiningum þeirra. Einnig hefur orðið veruleg markaðssamþjöppun fyrirtækja í talkennslum og talgervingu. Á innanlandsmarkaði í G20-löndunum (efnahagslega sterkum og fjölmennum löndum) hafa fimm alþjóðleg fyrirtæki verið ríkjandi, og í Evrópu einkum tvö – Nuance (bandarískt) og Loquendo (ítalskt). Árið 2011 tilkynnti Nuance um yfirtöku Loquendo þannig að markaðssamþjöppunin heldur enn áfram.

Tveir talgervlar fyrir íslensku hafa verið settir á markað. Formendabyggður talgervill var upphaflega gerður í kringum 1990 og betrumbættur í kringum 2000. Talgervillinn var að mestu notaður af blindum og sjónskertum en þótti ekki nógu fullkominn til notkunar í kerfum og verkbúnaði fyrir almennan markað.

Árið 2005 var búinn til nýr talgervill í samvinnu Háskóla Íslands, Símans og Hex software, sem er ekki lengur starfandi. Talgervillinn byggðist á tækni Nuance sem sá um þjálfun hans. Hann hefur hingað til ekki verið notaður í verkbúnaði fyrir almennan markað og sumum notendum finnst raddgæðin ekki fullnægjandi. Þar sem gæði tiltækra talgervla þykja ekki nógu mikil hefur aðalnotandi þeirra, Blindrafélagið, ákveðið að þróa nýjan talgervil í samvinnu við Háskóla Íslands, Háskólann í Reykjavík og pólska hugbúnaðarfyrirtækið Ivona. Gangi allt samkvæmt áætlun verður kerfið tilbúið árið 2012.[xxxvii]

Stakorðagreinir var þróaður fyrir íslensku árið 2003. Hann skilaði góðum árangri í greiningu, eða um 97% nákvæmni. Þá hefur íslenskur stúdent við Tokyo Institute of Technology hannað frumgerð af kerfi fyrir sjálfvirk orðaflaumskennsl í íslensku. Kerfið náði 67,5% orðanákvæmni.[xxxviii] Hvorugt þessara kerfa hefur verið notað í verkbúnað fyrir almennan markað. Um mitt ár 2011 hófu Háskólinn í Reykjavík og Máltæknisetur samvinnu við Google um undirbúning að smíði talþekkjara fyrir íslensku.[xxxix]

Miklar breytingar má sjá framundan vegna útbreiðslu snjallsíma sem nýs vettvangs fyrir tengsl fyrirtækja og viðskiptavina, í viðbót við venjulega síma, vefinn og tölvupóst. Þessar breytingar munu einnig hafa áhrif á nýtingu taltækninnar. Notkun á raddstýrðu viðmóti venjulegra síma mun fara minnkandi en mikilvægi talaðs máls sem notendavæns samskiptamáta við snjallsíma er sífellt að aukast. Það sem knýr þessa þróun er einkum aukin nákvæmni í talkennslum óháðum mælanda í þeim upplestrarkerfum sem þegar eru í boði sem miðlæg þjónusta fyrir notendur snjallsíma.

Vélþýðingar

Einfaldasta gerð vélrænna þýðinga felst í því að skipta út orðum í öðru málinu fyrir orð úr hinu málinu.

Hugmyndina að því að nota tölvur til að þýða mannleg mál má rekja til ársins 1946 og var henni fylgt eftir með töluverðu fjármagni til rannsókna á sjötta áratugnum og aftur á þeim níunda. Samt sem áður hafa vélþýðingar ekki náð að uppfylla þau fyrirheit um sjálfvirkar þýðingar milli tungumála sem þær gáfu á upphafsárunum.

Einfaldasta gerð vélþýðinga felst í því að skipta út orðum í öðru málinu fyrir orð úr hinu málinu. Þetta getur verið gagnlegt á efnissviðum þar sem notað er mjög afmarkað og formúlukennt mál, svo sem í veðurfregnum. En til þess að þýðing á máli sem er ekki eins staðlað verði viðunandi þarf að fella stærri textaeiningar (orðasambönd, málsgreinar, jafnvel heilar efnisgreinar) sem nákvæmast að samsvörunum þeirra í markmálinu. Helstu vandkvæðin felast í því að mannlegt mál er margrætt. Margræðni skapar vanda á mörgum sviðum, svo sem við einræðingu merkingar á orðasviðinu (*villa* er bæði ‘mistök’ og ‘veglegt hús’), og fallstjórn á setningafræðisviðinu, eins og í:

The woman saw the car and her husband, too

*[Konan sá bílinn og **maðurinn hennar** líka.]*

*[Konan sá bílinn og **manninn sinn** líka.]*

Ein leið til að búa til vélþýðingarkerfi er að nota málfræðilegar reglur. Þegar þýtt er á milli náskyldra tungumála getur aðferð beinna umskipta verið fýsileg, eins og í dæminu hér að ofan. En reglubyggð kerfi (byggð á málfræðilegri þekkingu) greina oft ílagstextann og skapa táknbyggt millistig sem texti markmálsins er síðan leiddur af. Velgengni þessarar aðferðar er undir því komin að til sé yfirgripsmikið orðasafn með beygingarlegum, setningafræðilegum og merkingarlegum upplýsingum, ásamt stóru safni málfræðireglna sem þjálfaðir málfræðingar hafa smíðað vandvirknislega. Það er langt og þar með dýrt ferli að koma þessum forsendum upp.

Á síðari hluta níunda áratugarins þegar tölvur urðu öflugri og ódýrari jókst áhugi á að nýta tölfræðileg líkön í vélþýðingum. Tölfræðileg líkön byggjast á greiningu tvímála málheilda, svo sem Europarl hliðstæðu málheildarinnar, sem hefur að geyma þingskjöl Evrópuþingsins á 11 Evrópumálum. Ef nóg er af gögnum virka tölfræðilegar vélþýðingar nægilega vel til þess að fá nokkurn veginn rétta merkingu texta á erlendu tungumáli með því að skoða samhliða texta og greina líkleg orðamynstur. Ólíkt þekkingarknúnum kerfum skila tölfræðilegar (eða gagnaknúnar) vélþýðingar oft málfræðilega röngu frálagi. Kosturinn við gagnaknúnar vélþýðingar er sá að þær eru ekki eins mannaflsfrekar, og einnig geta þær ráðið við ýmis málleg sérkenni (s.s. málshætti og orðtök) sem fara forgörðum í þekkingarstýrðu kerfunum.

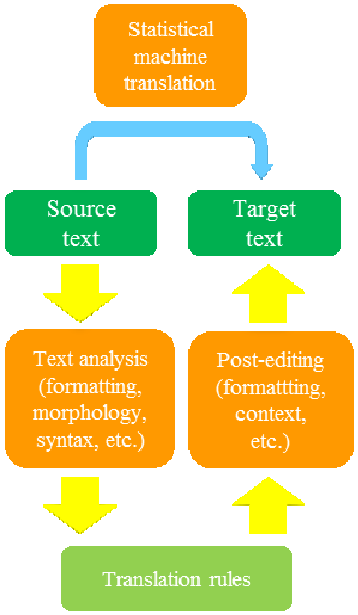


Figure 6: Machine translation (top: statistical; bottom: rule-based)

Styrkleikar og veikleikar þekkingarknúinna og gagnaknúinna vélþýðinga eru á mismunandi sviðum og því einbeita vísindamenn sér núorðið að blönduðum aðferðum sem sameina aðferðafræði beggja. Ein aðferðin er sú að nota bæði þekkingarknúið og gagnaknúið kerfi og láta svo valeiningu ákveða hvert sé besta frálag hverrar setningar. Þegar um er að ræða lengri setningar en 12 orð verða niðurstöðurnar þó sjaldnast fullkomnar. Betri aðferð er að sameina bestu hluta hverrar setningar úr mörgum frálögum; þetta getur verið tiltölulega flókið þar sem samsvaranir mismunandi möguleika eru ekki alltaf augljósar og því þarf að aðlaga þær.

Vélþýðingar milli íslensku og annarra mála eru mjög snúnar.

Vélþýðingar milli íslensku og annarra mála eru mjög snúnar. Vegna fjölbreyttra möguleika til smíði samsettra orða er oft erfitt að greina orð og hafa nægilega yfirgripsmikið orðasafn; frjáls orðaröð og sagnaragnir skapa ýmis vandamál í greiningu, og auðugt beygingarkerfi veldur vandkvæðum við að merkja rétt öll beygingaratriði s.s. kyn, fall, tölu, hátt, tíð, o.s.frv.

Vélþýðingar fyrir íslensku hafa lítið þróast. Stefán Briem, sjálfstætt starfandi fræðimaður, hefur unnið að vélþýðingum síðan snemma á níunda áratugnum og hefur hannað vélþýðingarkerfi fyrir íslensku. Árið 2008 opnaði hann á vefnum ókeypis þjónustu sem býður upp á þýðingar milli íslensku og þriggja annarra tungumála (ensku, dönsku og esperantó).[xl] Hrafn Loftsson, kennari við Háskólann í Reykjavík, og samstarfsmenn hans hafa hannað reglubyggt grófþýðingakerfi úr íslensku á ensku, grundvallað á Apertium-verkvangnum.[xli] Forútgáfa er nú á vefnum.[xlii] Google Translate hefur gefið kost á þýðingum úr og á íslensku síðan 2009. Gæðin voru heldur lítil í byrjun en hafa aukist.

Enn má auka gæði vélþýðingarkerfa verulega. Helstu vandkvæðin felast í aðlögun málfanganna að tilteknum efnissviðum eða notendahópum, og samþættingu tækninnar við vinnuferli sem nú þegar eru búin íðorðagrunni

og þýðingarminni. Annað vandamál er að flest núverandi kerfi eru miðuð við ensku og sinna einungis þýðingum milli íslensku og örfárra annarra mála. Þetta leiðir til árekstra í þýðingarflæðinu og þvingar notendur vélrænna þýðinga til að læra á mismunandi orðakótunartól fyrir mismunandi kerfi.

Matskeppnir nýtast vel til að bera saman gæði vélþýðingarkerfa, mismunandi aðferðafræði og frammistöðu þeirra gagnvart mismunandi tungumálapörum. Taflan hér á eftir, sem unnin var innan Euromatrix+ verkefnis Evrópusambandsins, sýnir útkomu allra para milli 22 af 23 opinberum tungumálum Evrópusambandsins. (Írska var ekki með í samanburðinum.) Niðurstöðum er raðað samkvæmt BLEU einkunnakvarða, þar sem hærri einkunn fæst fyrir betri þýðingu.[xliii] (Mennskur þýðandi myndi ná um 80 stigum.)

Bestu niðurstöðurnar (í grænum og bláum lit) fengust fyrir tungumál sem njóta góðs af umfangsmiklum samhæfðum rannsóknaráætlunum, sem og af tilvist margra samhliða málheilda (t.d. enska, franska, hollenska, spænska og þýska). Þau tungumál sem verr koma út eru merkt með rauðu. Þau skortir annaðhvort slíkar rannsóknaráætlanir eða eru eðlisólík öðrum tungumálum (t.d. ungverska, maltneska og finnska).

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Staða vélþýðinga fyrir mismunandi tungumálapör samkvæmt Euromatrix+ verkefninu

Önnur verksvið

Gerð máltæknibúnaðar felur oft í sér fjölda undirverkþátta sem ekki eru alltaf sýnilegir notendunum en gegna þó þýðingarmiklum þjónustuhlutverkum á bak við tjöldin. Þessir verkþættir byggjast allir á mikilvægum rannsóknarefnum sem hafa orðið að sjálfstæðum undirgreinum innan tölvumálvísinda.

Gerð máltæknibúnaðar felur oft í sér fjölda undirverkþátta sem ekki eru alltaf sýnilegir notendunum en gegna þó þýðingarmiklum þjónustuhlutverkum á bak við tjöldin

Spurningasvörun er t.d. virkt rannsóknarsvið og í tengslum við það hafa markaðar málheildir verið byggðar upp og vísindasamkeppnir haldnar. Spurningasvörun felur í sér annað og meira en lykilorðaleit (þar sem leitarvélin svarar með því að skila af sér safni skjala sem gætu varðað efnið) og gerir notendum kleift að spyrja beinskeyttra spurninga sem kerfið svarar á einkvæman hátt. Til dæmis:

> *Spurning: Hversu gamall var Neil Armstrong þegar hann steig fæti á tunglið?*

Svar: 38 ára.

Þótt spurningasvörun sé augljóslega af sömu rót og vefleit er hún nú fyrst og fremst yfirheiti yfir rannsóknarspurningar eins og: hvaða tegundir spurninga eru til og hvernig á að fást við þær; hvernig á að greina og bera saman þau skjöl sem hugsanlega hafa að geyma svarið (veita þau ósamrýmanleg svör?); og hvernig á að veiða afmarkaðar upplýsingar (svarið) út úr skjali á öruggan hátt án þess að hunsa samhengið.

Þetta tengist upplýsingaútdrætti, sviði sem var sérlega vinsælt og áhrifaríkt á tímum tölfræðibyltingarinnar í tölvumálvísindum snemma á tíunda áratug síðustu aldar. Með upplýsingaútdrætti er reynt að bera kennsl á tilteknar upplýsingaeiningar í tilteknum skjalaflokkum, svo sem að greina helstu þátttakendur í yfirtöku fyrirtækja eins og frá þeim er greint í umfjöllun dagblaða. Annað svið sem hefur verið rannsakað er frásagnir af hryðjuverkum. Þar er helsti vandinn að fella textann að sniðmáti sem tilgreinir brotamann, skotmark, tíma, staðsetningu og afleiðingar atviksins. Slík útfylling efnisbundinna sniðmáta er megineinkenni upplýsingaútdráttar og hann er því annað dæmi um tækni á bak við tjöldin sem myndar vel afmarkað rannsóknarsvið sem síðan þarf að fella inn í viðeigandi verkbúnað.

Tvö jaðarsvið sem ýmist geta verið sjálfstæður verkbúnaður eða þjónað sem stoðþættir bak við tjöldin eru samantekt texta og málmyndun. Með samantekt er leitast við að draga meginatriði langs texta saman í stuttu máli og er meðal annars boðið upp á slíkt í Microsoft Word. Þar er einkum stuðst við tölfræðilega aðferð til að greina ‘mikilvæg’ orð í textanum (þ.e. orð sem eru hlutfallslega mun algengari í textanum en í almennri málnotkun) og ákvarða síðan hvaða setningar innihalda flest þessara mikilvægu orða. Þær setningar eru síðan dregnar út úr textanum og settar saman til að mynda samantektina. Í þessari aðferð sem er mjög algeng í búnaði á almennum markaði felst samantektin eingöngu í því að draga setningar úr textanum, og textinn er því skorinn niður í hlutmengi upphaflegra setninga. Önnur aðferð, sem byggist á talsverðum rannsóknum, er sú að mynda nýjar setningar sem ekki koma fyrir í frumtextanum. Þetta krefst dýpri skilnings á textanum og er því mun viðkvæmara. Í flestum tilfellum er textamyndun ekki sjálfstæður búnaður heldur er hún felld inn í viðameiri hugbúnað, svo sem upplýsingakerfi í heilbrigðisþjónustu þar sem upplýsingum um sjúklinga er safnað, þær geymdar og síðan unnið úr þeim. Skýrslugerð er aðeins eitt af mörgum sviðum þar sem samantekt nýtist.

Engin þeirra tóla sem rætt er um í þessum undirkafla eru til fyrir íslensku.

Námsleiðir

Máltækni er mjög þverfaglegt svið þar sem saman kemur sérþekking málfræðinga, tölvunarfræðinga, stærðfræðinga, heimspekinga, sálfræðinga, taugafræðinga og fleiri. Hún hefur því ekki öðlast traustan sess í íslensku háskólaumhverfi. Um síðustu aldamót var ekki boðið upp á neinar námsleiðir eða einstök námskeið í máltækni eða tölvumálvísindum í neinum íslenskum háskóla og engar rannsóknir voru í gangi á þessum sviðum.

Máltækni er mjög þverfaglegt svið sem hefur ekki öðlast traustan sess í íslensku háskólaumhverfi.

Haustið 2002 tók Háskóli Íslands upp þverfaglegt meistaranám í máltækni. Um er að ræða tveggja ára nám (120 ECTS einingar) þar sem forkröfur eru B.A.-próf í tungumálum eða málvísindum eða B.Sc.-próf í tölvunarfræði (eða rafmagns- eða hugbúnaðarverkfræði). Árið 2007 var námið endurskipulagt í samvinnu milli íslenskudeildar Háskóla Íslands og tölvunarfræðideildar Háskólans í Reykjavík. Á meðan Norræni máltækniskólinn (Nordic Graduate School of Language Technology – NGSLT) var og hét, á árunum 2004-2009, gátu nemendur einnig tekið einstök

námskeið við skóla annars staðar á Norðurlöndunum og í Eystrasaltslöndunum.

Vegna skorts á fé og mannafla hefur ekki verið mögulegt að taka nýja nemendur inn í meistaranámið síðan 2009. Hins vegar er reglulega boðið upp á einstök námskeið í máltækni, málvinnslu og gagnamálfræði, bæði við Háskóla Íslands og Háskólann í Reykjavík.

Innlend verkefni og viðfangsefni

Aðeins um 320.000 manns tala íslensku og það er ekki nóg til þess að standa undir kostnaðarsamri þróun nýrra afurða. Það kostar jafn mikið að smíða máltæknibúnað fyrir íslensku og fyrir tungumál sem hundruð milljóna manna tala. Vegna þessa starfa næstum engin máltæknifyrirtæki á almennum markaði á Íslandi. Friðrik Skúlason ehf. hefur þróað og selt stafrýninn *Púka* en vinnur ekki að neinum nýjum framleiðsluvörum á sviði máltækni. Á síðasta áratug unnu Síminn og hugbúnaðafyrirtækið Hex með Háskóla Íslands að smíði bæði stakorðagreinis og talgervils fyrir íslensku. Hvorugt þessara fyrirtækja vinnur lengur að mál- eða taltækni.

Clara er nýlegt fyrirtæki sem þjónustar önnur fyrirtæki sem vilja vita hvað fólki finnst um framleiðsluvörur þeirra og þjónustu. Kerfi Clöru notar merkingargreiningu og sérstaka aðferð við framsetningu gagna til að greina viðhorf fólks á netinu. Verkfæri fyrirtækisins til greiningar á vefsíðum á íslensku kallast *Vaktarinn*.[xliv] Á fyrsta starfsári var það með 1200 notendur ef með eru taldir þeir sem notuðu þjónustuna ókeypis til reynslu. Clara er eina fyrirtækið á Íslandi sem er að þróa máltæknibúnað sem markaðsvöru.

Árið 2000 setti íslenska ríkið af stað sérstakt máltækniátak með það fyrir augum að styðja stofnanir og fyrirtæki í því að búa til grundvallargögn fyrir íslenska máltækni. Þetta frumkvæði leiddi til nokkurra verkefna sem hafa haft mjög mikil áhrif á máltækni á Íslandi. Helstu afurðir máltækniátaksins eru eftirfarandi:[xlv]

- Gagnagrunnur með beygingarlýsingu íslensks nútímamáls
- Málfræðilega mörkuð málheild með 25 milljónum orða
- Þjálfunarsafn fyrir gagnastýrða málfræðilega mörkun
- Talgervill
- Stakorðagreinir
- Betrumbættur stafrýnir

Þegar máltækniátakinu lauk árið 2004 ákváðu fræðimenn frá þremur stofnunum (Háskóla Íslands, Háskólanum í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum) sem höfðu tekið þátt í flestum verkefnum máltækniátaksins að sameinast um stofnun Máltækniseturs með það að markmiði að vinna áfram að verkefnum sem þegar voru komin af stað. Aðalhlutverk Máltækniseturs er að:

- vera upplýsingaveita um íslenska máltækni og reka vefsetur í því skyni (http://maltaeknisetur.is);
- stuðla að samstarfi háskóla, stofnana og fyrirtækja um máltækniverkefni;
- skipuleggja og samhæfa háskólakennslu á sviði máltækni;
- taka þátt í norrænu, evrópsku og alþjóðlegu samstarfi á sviði máltækni;
- standa fyrir og eiga aðild að rannsóknar- og þróunarverkefnum á sviði máltækni;
- halda utan um ýmiss konar hráefni og afurðir á sviði máltækni;

- halda árlega ráðstefnu með þátttöku fræðimanna, fyrirtækja og almennings;
- beita sér fyrir eflingu íslenskrar máltækni á öllum sviðum.

Á undanförnum árum hafa fræðimenn Máltækniseturs átt frumkvæðið að nokkrum nýjum verkefnum sem hafa verið styrkt að hluta til af Rannsóknasjóði og Tækniþróunarsjóði. Mikilvægasta afurð þessa verkefna er opni hugbúnaðurinn IceNLP (málfræðilegi markarinn IceTagger, hluta-þáttarinn IceParser, og lemmunarforritið Lemmald).[xlvi] Árið 2009 fékk Máltæknisetur háan þriggja ára öndvegisstyrk frá Rannís til verkefnisins 'Hagkvæm máltækni utan ensku – íslenska tilraunin'. Innan þessa verkefnis er unnið að þróun máltæknigagna fyrir íslensku.

Eins og hér hefur komið fram hafa margvísleg verkefni leitt til þróunar ýmissa máltæknitóla og málfanga fyrir íslensku. Í næsta kafla er gefið yfirlit yfir núverandi stöðu íslenskrar máltækni.

Aðgengi að máltæknitólum og málföngum

Í eftirfarandi töflu er gefið yfirlit yfir stöðu íslenskrar máltækni og máltæknibúnaðar. Einkunnir máltæknitóla og málfanga eru byggðar á mati helstu sérfræðinga á sviðinu sem gáfu einkunnir á skalanum frá 0 (mjög lágt) til 6 (mjög hátt) út frá sjö viðmiðum.

	Magn	Aðgengi	Gæði	Yfirgrip	Þroski	Sjálfbærni	Aðlögunarhæfni
Máltækni (tól, tækni og verkbúnaður)							
Talkennsl	1	1	1	2	1	0	1
Talgerving	1	1	2	3	2	1	1
Textagreining	2	5,5	4,5	4	4	4	3
Textatúlkun	0,5	0,5	0,5	0,5	0,5	0,5	0,5
Málmyndun	0	0	0	0	0	0	0
Vélþýðingar	1	4	2	2	2	2	3
Málföng (tilföng, gögn og þekkingargrunnar)							
Málheildir	1,5	4	3,5	3	2,5	4,5	3
Talmálsheildir	1	2	2	2	1	2	2
Hliðstæðar málheildir	1	1	2	1	2	2	1
Orðaföng	1	2	3	3	2	2	2
Málfræðilýsing	1	4	3	3	3	3	2

Meginniðurstöður fyrir íslensku eru eftirfarandi:

- Íslenska stendur 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, vélþýðingum, talmálsheildum, hliðstæðum málheildum og orðagögnum.
- Háþróaður máltæknibúnaður og málföng, svo sem til textatúlkunar og málmyndunar, er ekki til.

Um síðustu aldamót var íslensk máltækni varla til. Þetta breyttist eftir 1999, þegar sérstakur starfshópur skilaði skýrslu um máltækni til menntamálaráðherra.[xlvii] 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 (sem þá jafngilti um 10 milljónum evra) 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.[xlviii] 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, en miðað við núverandi efnahagsástand er vart að búast við framlagi á fjárlögum á næstunni.

Samanburður tungumála

Máltæknistuðningur er mjög mismunandi milli málsamfélaga. Til að bera saman stöðuna milli mála er í þessum kafla sett fram mat sem byggist á tveimur verkbúnaðarsviðum (vélþýðingum og talvinnslu), einni gerð undirliggjandi tækni (textagreiningu) og grundvallarmálföngum sem þarf til smíði máltæknibúnaðar.

1. klasi	2. klasi	3. klasi	4. klasi
	Enska, franska, þýska, spænska, ítalska, hollenska, tékkneska, danska, portúgalska, finnska	Baskneska, búlgarska, katalónska, króatíska, eistneska, galisíska, gríska, ungverska, pólska, serbneska, slóvenska, sænska	Íslenska, írska, lettneska, litháíska, maltneska, norska, rúmenska, slóvakíska

Mynd 1: Tungumálaklasar fyrir talvinnslu

1. klasi	2. klasi	3. klasi	4. klasi
	Enska	Spænska, katalónska, þýska, ítalska, franska, tékkne-	Baskneska, búlgarska, króatíska, danska, hollenska, eistneska, finnska, galisíska, gríska, ungverska, íslenska, írska, lettneska, litháíska, maltneska, norska, pólska, portúgalska, rúmenska, serbneska, slóvakíska, slóvenska,

Mynd 2: Tungumálaklasar fyrir vélþýðingar

Lýsing klasa (fyrir talvinnslu og vélþýðingar)

- 1. klasi (afburðagóður máltæknistuðningur): Til er tækni sem hefur yfirburði í gæðum og virkni sem hægt er að nota í svo að segja öllum búnaði sem máli skiptir.
- 2. klasi (góður stuðningur): Til er tækni af viðunandi gæðum og virkni þar sem nýtingin er takmörkuð við ákveðinn búnað eða svið.
- 3. klasi (meðalgóður stuðningur): Til eru frumgerðir úr rannsóknum, fyrstu markaðsvörur eða ókeypis þjónusta af mismunandi gæðum og virkni.
- 4. klasi (lítill eða nær enginn stuðningur): Frá því að vera á hönnunarstigi að hráum frumgerðum af mjög takmörkuðum gæðum og virkni.

1. klasi	2. klasi	3. klasi	4. klasi	5. klasi
	Enska	Tékknesk a, hollenska, þýska	Búlgarska, franska, norska, pólska, portúgalska, spænska, sænska, baskneska, katalónska, danska, finnska, galisíska, ungverska, írska, ítalska, rúmenska	Króatíska, eistneska, gríska, íslenska, lettneska, litháíska, maltneska, serbneska, slóvakíska, slóvenska

Mynd 3: Tungumálaklasar fyrir textagreiningu

1. klasi	2. klasi	3. klasi	4. klasi	5. klasi
	Enska	Þýska, ungverska, sænska, franska, hollenska,	Spænska, norska, portúgalska, danska, pólska, rúmenska, búlgarska, katalónska, baskneska, galisíska, eistneska	Ítalska, finnska, króatíska, slóvenska, íslenska, írska, gríska, lettneska,slóvakíska, serbneska,

Mynd 4: Tungumálaklasar fyrir málföng

Lýsing tungumálaklasa (fyrir textagreiningu og málföng)

- 1. klasi (afburðagóður máltæknistuðningur): Til er tækni/málföng sem er í víðtækri notkun og nær yfir svo að segja allt málkerfið (orðaforða, samsetningar, málfræði, líkingar o.s.frv.).
- 2. klasi (mjög góður stuðningur): Til er tækni/málföng sem er nýtt í ýmsum búnaði og nær yfir mikilvægustu þætti málkerfisins.
- 3. klasi (góður stuðningur): Til er tækni/málföng sem nær yfir nokkurn hluta málkerfisins og er nýtt í búnaði sem er venjulega bundinn á-kveðnum sviðum.

- 4. klasi (meðalgóður stuðningur): Til eru frumgerðir málfanga eða búnaðar úr rannsóknum en af misjöfnum gæðum og yfirgripi.

- 5. klasi (lítill eða nær enginn stuðningur): Frá því að vera á hönnunarstigi að hráum frumgerðum (mjög takmörkuð gæði og yfirgrip, leikfangakerfi).

Töflurnar hér að framan sýna að 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. 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.

Niðurstöður

Í þessari hvítbókaröð hefur verið gerð mikilvæg upphafsatlaga að því að meta máltæknistuðning fyrir 30 Evrópumál, og gera ítarlegan samanburð á þessum málum. Með því að greina eyður, þarfir og skort er evrópskt máltæknisamfélag og aðrir hagsmunaaðilar nú í stöðu til þess að skipuleggja meiri háttar rannsóknar- og þróunaráætlun sem miðast að því að Evrópa verði raunverulega margmála með stuðningi tækninnar.

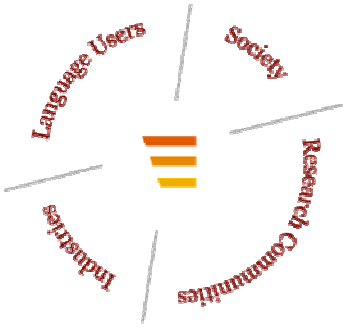
Hér hefur komið fram geysimikill innbyrðis munur á Evrópumálum. Þótt ágætur hugbúnaður og málföng sé til fyrir sum tungumál og verksvið eru grundvallareyður á þessum sviðum í öðrum málum (venjulega þeim ‘smærri’). Mörg tungumál skortir grunntækni til textagreiningar og nauðsynleg málföng til að þróa slíka tækni. Önnur hafa grundvallarbúnað og málföng en hafa ekki burði til að ráðast í merkingarlega vinnslu. Þess vegna er enn þörf á víðtæku átaki til að ná því metnaðarfulla markmiði að koma upp hágæða vélþýðingum milli allra Evrópumála.

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.

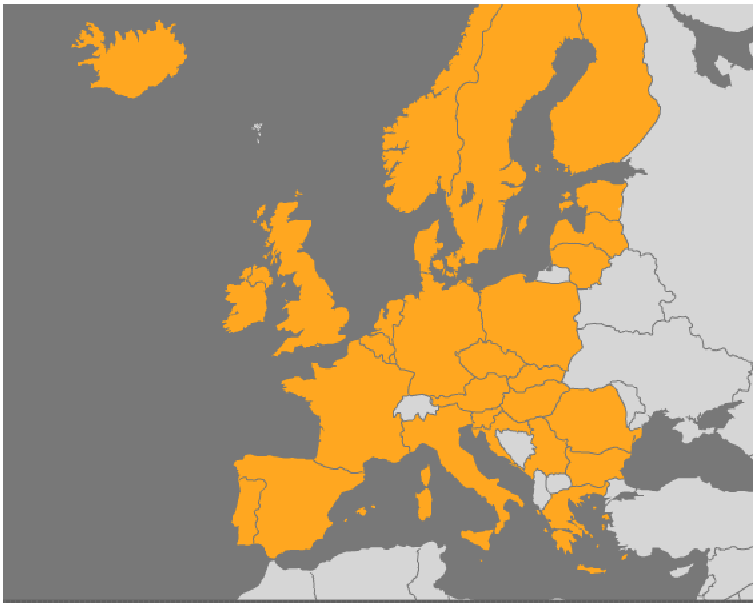
Langtímamarkmið META-NET er að koma upp hágæða máltækni fyrir öll tungumál til að skapa pólitíska og efnahagslega einingu með menningarlegum fjölbreytileika. Tæknin mun hjálpa til við að brjóta múra og reisa brýr milli Evrópumála. Þetta krefst þess að allir hagsmunaaðilar – í stjórnmálum, rannsóknum, viðskiptum, og samfélaginu öllu – sameini krafta sína til framtíðar.

Um META-NET

ooooMETA-NET er öndvegisnet fjármagnað af Evrópusambandinu. Netið samanstendur af 47 þátttakendum frá 31 Evrópulandi. META-NET fóstrar Tæknibandalag um margmála Evrópu (Multilingual Europe Tec nology Alliance, META), sem er sístækkandi samfélag evrópskra fræðimanna og stofnana á sviði máltækni.



The Multilingual Europe Technology Alliance (META)



Lönd sem eiga fulltrúa í META-NET

META-NET vinnur með öðrum frumkvöðlum eins og Common Langu ge Resources and Technology Infrastructure (CLARIN), sem stuðlar að eflingu stafrænna hugvísindarannsókna í Evrópu. META-NET fóstrar tæknilega undirstöðu raunverulegs margmála evrópsks upplýsingasamfélags sem:

- gerir kleift að eiga samskipti og samvinnu þvert á tungumál;
- sér fyrir jöfnum aðgangi að upplýsingum og þekkingu á öllum tungumálum;
- gefur Evrópubúum kost á þróaðri og ódýrri netvæddri upplýsingatækni.

META-NET ýtir undir og kynnir margmála tækni fyrir öll Evrópumál. Sú tækni raungerir sjálfvirkar þýðingar, samningu efnis, upplýsingavinnslu og þekkingarstjórnun fyrir fjölbreyttan búnað og margvísleg efnissvið. Netið stefnir að því að bæta þær aðferðir sem beitt er þannig að samskipti og samvinna þvert á tungumál verði greiðari. Evrópubúar eiga jafnan rétt á upplýsingum og þekkingu óháð því hvaða tungumál þeir tala.

Aðgerðaáætlun

META-NET var sett af stað 1. febrúar 2010 með það að markmiði að efla rannsóknir á sviði máltækni. Netið styður hugsjónina um Evrópu sem einn stafrænan markað og upplýsingarými. META-NET hefur ýtt úr vör ýmsum aðgerðum til að ná markmiðum sínum. META-VISION, META-SHARE og META-RESEARCH eru þrjár aðgerðaáætlanir netsins.

Þrjár aðgerðaáætlanir META-NET

META-VISION fóstrar kvikt og áhrifamikið samfélag hagsmunaaðila sem sameinast um sameiginlega sýn og útfærða rannsóknarstefnu (strat gic research agenda, SRA). Megináherslan í þessu starfi er að byggja upp samræmt og samtengt máltæknisamfélag í Evrópu með því að leiða saman fulltrúa dreifðra og fjölbreyttra hópa hagsmunaaðila. Á fyrsta ári META-NET beindust kynningar á FLaReNet Forum (Spáni), Language Technology Days (Lúxemborg), JIAMCATT 2010 (Lúxemborg), LREC 2010 (Möltu), EAMT 2010 (Frakklandi) og ICT 2010 (Belgíu) að því að ná athygli út á við. Áætlað er að META-NET hafi nú þegar náð sambandi við meira en 2.500 manns á sviði máltækni, LT, til að vinna með þeim að markmiðum sínum og framtíðarsýn. Á META-FORUM 2010 í Brussel kynnti META-NET fyrstu niðurstöður af mótun sinni á framtíðarsýn fyrir meira en 250 þátttakendum. Í nokkrum gagnvirkum fundalotum veittu þátttakendur endurgjöf á þá sýn sem netið hafði fram að færa.

META-SHARE skapar opinn og dreifðan vettvang til að skiptast á gögnum og deila þeim. Jafningjastýrt net gagnabrunna mun vista málleg gögn, tól og vefþjónustu sem er skjalað með hágæða lýsigögnum og skipulagt í stöðluðum flokkum. Auðvelt verður að nálgast gögnin og leita í þeim í heild. Þarna verða bæði opin og ókeypis málföng sem og gögn með takmörkuðum aðgangi sem greiða verður fyrir notkun á. META-SHARE leitast við að afla mállegra gagna, búnaðar og kerfa sem þegar eru til, sem og framleiðsluvara sem eru nýjar eða á vinnslustigi og nauðsynlegar til að byggja upp og meta nýja tækni, vörur og þjónustu. Endurnýting, samtenging og endurvinnsla mállegra gagna gegnir lykilhlutverki í þessu. META-SHARE verður á endanum miðpunktur máltækni-markaðarins fyrir þá sem vinna að þróun, sérfræðinga í staðfærslu, þýðendur og tungumálasérfræðinga frá litlum, meðalstórum og stórum fyrirtækjum. META-SHARE sinnir öllu þróunarferli máltækninnar – frá rannsóknum til nýsköpunar í vörum og þjónustu. Lykilatriði í þessu starfi er að koma META-SHARE á fót sem mikilvægum og verðmætum þætti evrópskra og alþjóðlegra innviða máltæknisamfélagsins.

META-RESEARCH reisir brýr til skyldra tæknisviða. Þetta starf leitast við að nýta framfarir á öðrum sviðum og einbeita sér að nýskapandi rannsóknum sem geta eflt máltækni. Einkum er stefnt að því að færa meiri merkingarfræði inn í vélþýðingar, ná fram bestu deilingu verkþátta í blönduðum vélþýðingum, nýta samhengi við útreikning sjálfvirkra þýðinga og byggja upp reynslugrunn fyrir vélþýðingar. META-RESEARCH vinnur með öðrum sviðum og greinum, svo sem vélrænu námi merkingarvefssamfélaginu. META-RESEARCH leggur áherslu á að safna gögnum, ganga frá gagnasöfnum og skipuleggja málgögn til að nota við mat; að safna saman birgðum tóla og aðferða; og skipuleggja vinnustofur og æfingaferli fyrir þátttakendur í samfélaginu. Þetta starf hefur nú þegar bent ákveðið á þætti í vélþýðingum þar sem merkingarfræði getur endurbætt núverandi aðferðir. Þar að auki hefur starfið getið af sér tillögur um það hvernig eigi að fást við þann vanda að fella merkingarupplýsingar inn í vélþýðingarkerfi. META-RESEARCH er líka að leggja lokahönd á ný málföng til nota í vélþýðingum, Annotated Hybrid Sample MT Corpus,

sem inniheldur gögn um tungumálapörin ensku-þýsku, ensku-spænsku og ensku-tékknesku. META-RESEARCH hefur einnig þróað hugbúnað sem safnar margmála málheildum sem eru huldar á vefnum.

Þátttakendur

Eftirfarandi tafla sýnir þær stofnanir sem taka þátt í META-NET og fulltrúa þeirra.

Land	Stofnun	Fulltrúi
Austurríki	Háskólinn í Vín	Gerhard Budin
Belgía	Háskólinn í Antwerpen	Walter Daelemans
	Háskólinn í Leuven	Dirk van Compernolle
Bretland	Háskólinn í Manchester	Sophia Ananiadou
	Háskólinn í Edinborg	Steve Renals
Búlgaría	Búlgarska vísindaakademían	Svetla Koeva
Danmörk	Háskólinn í Kaupmannahöfn	Bolette Sandford Pedersen og Bente Maegaard
Eistland	Háskólinn í Tartu	Tiit Roosmaa
Finnland	Aalto-háskóli	Timo Honkela
	Háskólinn í Helsinki	Kimmo Koskenniemi og Krister Linden
Frakkland	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Grikkland	Mál- og talvinnslustofnunin, “Athena” R.C.	Stelios Piperidis
Holland	Háskólinn í Utrecht	Jan Odijk
	Háskólinn í Groningen	Gertjan van Noord
Írland	Dublin City-háskólinn	Josef van Genabith
Ísland	Háskóli Íslands	Eirikur Rögnvaldsson
Ítalía	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale “Antonio Zampolli”	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Króatía	Háskólinn í Zagreb	Marko Tadić
Kýpur	Kýpurháskóli	Jack Burston
Lettland	Tilde	Andrejs Vasiljevs
	Stærðfræði- og tölvunarfræðistofnunin, Lettlandsháskóla	Inguna Skadina

Land	Stofnun	Fulltrúi
Litháen	Stofnun litháískrar tungu	Jolanta Zabarskaitė
Lúxemborg	Arax Ltd.	Vartkes Goetcherian
Malta	Möltuháskóli	Mike Rosner
Noregur	Háskólinn í Bergen	Koenraad De Smedt
Portúgal	Háskólinn í Lissabon	Antonio Branco
	Kerfisverkfræði- og tölvustofnunin	Isabel Trancoso
Pólland	Pólska vísindaakademían	Adam Przepiórkowski og Maciej Ogrodniczuk
	Háskólinn í Lodz	Barbai-ra Lewandowska-Tomaszczyk og Piotr Pęzik
Rúmenía	Rúmenska vísindaakademían	Dan Tufis
	Alexandru Ioan Cuza-háskólinn	Dan Cristea
Serbía	Háskólinn í Belgrad	Dusko Vitas, Cvetana Krstev og Ivan Obradovic
	Stofnun Mihailo Pupin	Sanja Vranes
Slóvakía	Slóvakíska visindaakademían	Radovan Garabik
Slóvenía	Jozef Stefan-stofnunin	Marko Grobelnik
Spánn	Barcelona Media	Toni Badia
	Tækniháskólinn í Katalóníu	Asunción Moreno
	Pompeu Fabra-háskólinn	Núria Bel
Svíþjóð	Háskólinn í Gautaborg	Lars Borin
Tékkland	Karlsháskóli í Prag	Jan Hajic
Ungverjaland	Ungverska vísindaakademían	Tamás Váradi
	Tækni- og viðskiptaháskólinn í Búdapest	Géza Németh og Gábor Olaszy
Þýskaland	DFKI	Hans Uszkoreit og Georg Rehm
	RWTH háskólanum í Aachen	Hermann Ney
	Háskólinn í Saarland	Manfred Pinkal

Tilvísanir

[i] https://hagstofa.is/lisalib/getfile.aspx?ItemID=12358

[ii] http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf

[iii] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[iv] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[v] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[vi] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[vii] http://www.hagstofa.is/Hagtolur/Mannfjoldi

[viii] http://visindavefur.hi.is/svar.php?id=53154

[ix] http://www12.statcan.ca/census-recensement/index-eng.cfm

[x] http://www.althingi.is/altext/139/s/0870.html

[xi] http://mallyskur.is/

[xii] http://fraedi.is/tvinnhljod/

[xiii] http://*www.islenskan.is/Islenska_til_alls.pdf*

[xiv] http://www.arnastofnun.is/page/arnastofnun_mal_islenskmalnefnd

[xv] http://www.arnastofnun.is/page/arnastofnun_mal_malraektarsjodur

[xvi] http://www.althingi.is/lagas/139a/2006040.html

[xvii] http://www.hagstofa.is/Hagtolur/Menningarmal/Utvarp

[xviii] http://www.althingi.is/lagas/139a/2000053.html

[xix] http://www.menntamalaraduneyti.is/menningarmal/dit/

[xx] http://www.menntamalaraduneyti.is/utgefid-efni/namskrar/nr/3953

[xxi] http://www.menntamalaraduneyti.is/utgefid-efni/namskrar/nr/3954

[xxii] http://www.namsmat.is/vefur/rannsoknir/PISA_2009/pisa_2009_island.pdf

[xxiii] http://www.islenskan.is/Islenska_til_alls.pdf

[xxiv] http://*www.islenskan.is/Islenska_til_alls.pdf*

[xxv] http://icelandiconline.is/

[xxvi] http://www.visir.is/article/20101001/FRETTIR01/175424536

[xxvii] http://ec.europa.eu/enlargement/press_corner/key-documents/opinion-iceland_2010 _en.htm

[xxviii] http://www.internetworldstats.com/stats4.htm#european

[xxix] http://www.checkfacebook.com/

[xxx] http://www.isnic.is/tolur/index.html

[xxxi] http://www.webhosting.info/registries/country_stats/IS

[xxxii] http://www.absmedia.is/frettir/nr/111000/

[xxxiii] http://www.absmedia.is/frettir/nr/112752/

[xxxiv] http://www.languagetool.org/

[xxxv] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[xxxvi] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic search_capabilities.html

[xxxvii] http://www.blind.is/verkefni/talgervlaverkefnid/

[xxxviii] http://www.hindawi.com/journals/asmp/2008/573832/ref/

[xxxix] http://almannaromur.is

[xl] http://tungutorg.is/

[xli] http://www.apertium.org/

[xlii] http://nlp.cs.ru.is/ApertiumISENWeb/index_en.jsp

[xliii] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA.

[xliv] http://www.vaktarinn.is/

[xlv] http://dspace.utlib.ee/dspace/bitstream/handle/10062/9670/Icelandic%20language%20resources.pdf;jsessionid=A7320810CB6EA717510D0460EADE8C5B?sequence=1

[xlvi] Fáanlegt hjá http://icenlp.sourceforge.net

[xlvii] http://brunnur.stjr.is/mrn/utgafuskra/utgafa.nsf/xsp/.ibmmodres/domino/OpenAttachment/mrn/utgafuskra/utgafa.nsf/F0250A90B6D7F31B002576F00058D4B8/Attachment/tungutaekni.pdf

[xlviii] http://dspace.utlib.ee/dspace/bitstream/handle/10062/9670/Icelandic%20language%20resources.pdf;jsessionid=A7320810CB6EA717510D0460EADE8C5B?sequence=1