Rögnvaldur Ólafsson

Tungutækniverkefni menntamálaráðuneytisins

Rögnvaldur Ólafsson

Tungutækniverkefni menntamálaráðuneytisins

Tungutækniverkefnið hófst haustið 1998 að frumkvæði Björns Bjarnasonar, þáverandi menntamálaráðherra. Þá fékk hann Rögnvald Ólafsson eðlisfræðing til þess að kanna hver væri staða íslenskrar tungu í upplýsingaþjóðfélaginu. Rögnvaldur fékk til liðs við sig Eirík Rögnvaldsson, prófessor í íslensku við Háskóla Íslands, og Þorgeir Sigurðsson, rafmagnsverkfræðing og íslenskufræðing, hjá Staðlaráði Íslands. Árangurinn af þeirra starfi birtist í skýrslunni: Tungutækni – skýrsla starfshóps (http://www.tungutaekni.is/news/Skyrsla.pdf) sem gefin var út í apríl 1999 af menntamálaráðuneytinu. Í skýrslunni kom fram að þörf væri fyrir átak á sviði tungutækni til þess að tryggja stöðu íslenskrar tungu í upplýsingaþjóðfélaginu. Slíkt átak þyrfti að gera með stuðningi hins opinbera og það mundi borga sig til lengri tíma litið. Átakið þyrfti að gera á fjórum sviðum:

- Byggja upp sameiginleg gagnasöfn, málsöfn, sem geti nýst fyrirtækjum sem hráefni í afurðir
- Hagnýtar rannsóknir á sviði tungutækni þyrfti að styrkja
- Fyrirtæki ætti að styrkja til þess að þróa afurðir tungutækni
- Menntun á sviði tungutækni og málvísinda yrði að efla

Verkefnisstjórn um upplýsingasamfélagið fjallaði um skýrsluna og hún var síðan lögð fyrir ríkisstjórnina og í framhaldi af því lagði þáverandi menntamálaráðherra, Björn Bjarnason, til að fé yrði veitt til þessara mála. Á fjáraukalögum 2000 voru 40 milljónir króna veittar til tungutækni og á fjárlögum 2001 64,5 milljónir króna. Alls voru því 104,5 milljónir króna til ráðstöfunar á árinu 2001. Á árunum 2003 og 2004 voru síðan veittar 28,5 milljónir króna til verkefnisins þannig að alls hafa verið veittar 133 milljónir króna til þess.

Í stuttu máli má segja að tilgangur *Tungutækniverkefnisins* sé að koma fótum undir tungutækni á Íslandi. Í því felst að byggja upp þekkingu á viðfangsefninu og þá gagnagrunna sem þarf til þess að hægt sé að nýta íslenskt mál, bæði ritað og mælt, í nýjustu samskipta- og tölvutækni.

Þegar *Tungutækniverkefnið* fór af stað var lítil sem engin þekking hér á landi á þessu sviði. Smátt og smátt byggðist þekkingin upp og nú í lok árs 2004 hefur ákveðnum áföngum þegar verið náð og aðrir eru í sjónmáli. Þegar *Tungutækniverkefninu* lýkur í árslok 2004 á tungutæknin að vera komin á það stig að hún þurfi ekki lengur sérstakan stuðning heldur geti hún sótt um styrki í hið almenna styrkjakerfi. Síðustu verkum sem *Tungutækniverkefnið* hefur styrkt mun þó ekki ljúka fyrr en á árinu 2005 og einu ekki fyrr en 2007.

Verkefnisstjórn

Ráðherra skipaði í upphafi verksins verkefnisstjórn sem skyldi vera honum til ráðuneytis um verkefnið. Hún er nú þannig skipuð: Ari Arnalds verkfræðingur sem er formaður verkefnisstjórnarinnar; Bjarki A. Brynjarsson verkfræðingur; Höskuldur Þráinsson prófessor í íslensku og Erla Skúladóttir lögfræðingur. Erla Skúladóttir tók við af Kristínu Haraldsdóttur lögfræðingi sem sat í stjórninni fyrstu árin. Verkefnisstjóri er Rögnvaldur Ólafsson eðlisfræðingur og hefur hann frá upphafi verið eini starfsmaður verkefnisins. Rekstur verkefnisins, kynningar, ráðstefnur, útgáfur, styrkveitingar og annað slíkt hefur verið í höndum verkefnisstjóra og stjórnar.

Hvað er tungutækni?

Tungutækni er sú tækni sem meðferð tungumálsins í tölvum og hugbúnaði byggist á. Þar er um að ræða að koma rituðu og mæltu máli inn og út úr tölvum og að meðhöndla það í tölvum og hugbúnaði. Til tungutækni teljast t.d. vélrænar þýðingar milli tungumála, leiðrétting á texta o.s.frv.

Greinin er nátengd tölvutækni og tölvuverkfræði og styðst jafnframt oft við gervigreind. Hún byggist einnig á þekkingu á málvísindum og tungumálinu en á það reynir t.d. í villupúkum. Tungutæknin styðst einnig við ýmislegt úr sálfræði, skynjunarfræði og hljóðfræði, eins og hvernig fólk skilur tal og hvernig fólk myndar hljóð og orð. Til dæmis verða talgervlar ekki áheyrilegir án þess að beitt sé þekkingu á hljóðfræði og framburði. Tungutæknin er því það sem kallast þverfagleg grein.

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

Verkefnin

Á árinu 2002 styrkti *Tungutækniverkefnið* ýmis verkefni sem miðuðust að því að styrkja grunninn undir tungutækni. Í fyrsta lagi eru þetta verkefni sem tengjast texta og meðferð hans. Meðal verkefnanna eru:

– Beygingarlýsing 170 þúsund íslenskra orða þar sem skráðar eru allar beygingarmyndir orðanna

– Markari sem er hugbúnaður sem greinir orð í íslenskum texta í orðflokka og hugbúnaður til leiðréttingar á villum í málfræði

– Málfræðiverkefni þar sem setningar eru greindar í orðflokka

– Endurbætur á Púka Friðriks Skúlasonar sem leiðréttir stafsetningu og skiptir orðum milli lína

Þessum verkefnum er nú að mestu lokið. Verkefnunum um beygingarlýsingu og markarann er lokið og má sjá og nýta árangurinn á heimasíðu Orðabókar Háskólans (www.lexis.hi.is). Þessi verk eru nýjung á Íslandi og nauðsynlegur grunnur fyrir áframhaldandi vinnu við tungutækni en nýtast einnig á margan annan hátt. Hjá Friðriki Skúlasyni ehf. er verið að vinna að verkefni um leiðréttingu á málfræði og mun því ljúka á árinu 2005. Verkefninu um leiðréttingar á stafsetningu lauk í desember 2003 og nú er kominn á markað nýr Púki 2003 sem byggist á niðurstöðum þess. Lýsingar á verkefnunum má finna á öðrum stað í þessu riti.

Önnur tegund verkefna sem *Tungutækniverkefnið* hefur hrundið af stað tengist tölvutali og tölvuheyrn en það er tækni sem er í vaxandi mæli notuð í ýmiss konar tækjabúnaði. Til þess að hægt sé að nýta íslenskt mál í þeirri tækni þannig að íslenska standi þar jafnfætis öðrum tungumálum er nauðsynlegt að byggja upp grunnþekkingu á þessu sviði. Snemma styrkti *Tungutækniverkefnið* tvö verkefni um tal. Síðan var í lok árs 2002 styrkt verkefni sem nefnt er Hjal. Það fjallar um talgreiningu, það er að segja að gera tölvum kleift að skilja talað mál. Í verkefninu var þróaður og byggður svonefndur stakorðagreinir, sem skynjar og skilur einstök orð í tali fólks. Í Hjali tóku þátt Landssími Íslands hf., Hex ehf., Nýherji hf., Háskóli Íslands og Grunnur Gagnalausnir hf. Til þess að vinna þetta verkefni þurfti m.a. greiningu á málhljóðum í íslensku tali, að safna talsýnum, skrá þau og búa til orðasafn. Öll þessi verkefni eru mikilvæg undirstaða undir notkun íslensks tals í tækjabúnaði eins og sjálfvirkum svarþjónustum í símkerfum.

Hjali lauk snemma árs 2004. Að verkefninu loknu má segja að helstu grunnatriði talskynjunar séu komin í viðunandi horf. Verkefnið gekk mjög vel og áhugavert og vel heppnað samstarf var þar á milli öflugra íslenskra fyrirtækja, Háskóla Íslands og erlends fyrirtækis, Scansoft, sem sérhæfir sig í gerð talgreina.

Eins og fyrr var sagt er greinirinn sem unninn var í Hjali svonefndur stakorðagreinir sem greinir einstök orð. Miklu meiri vinna er að búa til talgreini sem greinir samfelldan texta. Slíkur greinir mun þó að verulegu leyti byggjast á þeirri vinnu sem unnin var í Hjali.

Í framhaldi af Hjali hafa þegar verið búnar til

símaþjónustur þar sem hægt er að hringja og biðja um upplýsingar í almennu mæltu máli, tölvan skilur um hvað er beðið og svarar með aðstoð talgervils.

Talgervill er tæki eða hugbúnaður sem kemur texta frá sér í mæltu máli, gefur tölvum mál ef svo má segja. Í tækni sem nýtir talgreini þarf að jafnaði einnig talgervil því að í samskiptum við fólk þarf bæði að skilja mál þess og tala til þess. Talgervill gerir sjónskertum fært að „lesa" ritað mál sé það á tölvutæku formi. Hann getur einnig nýst þeim sem eiga erfitt með að lesa af einhverjum ástæðum og vilja nýta tölvu til að lesa fyrir sig. Í Hjali hefur verið notaður íslenskur talgervill frá Infovox sem kallaður er Snorri. Sjónskertir nota eldri og ófullkomnari talgervil frá sama fyrirtæki. Þessir talgervlar eru ekki nægilega góðir fyrir almenn not. Í framhaldi af Hjal-verkefninu var fyrir tilstuðlan *Tungutækniverkefnisins* unnin undirbúningsvinna að nýjum íslenskum talgervli og standa vonir til þess að nýr talgervill verði gerður fljótlega. Fyrsti nemandinn sem útskrifaðist með meistarapróf í tungutækni skrifaði ritgerð sína um íslenskan talgervil og vandamál tengd honum.

Þriðja tegund verkefna sem *Tungutækniverkefnið* hefur komið af stað eru textagrunnar. Markaður textagrunnur, eða málheild, er mjög stórt safn texta þar sem öll orð hafa verið greind í orðflokka. Textinn er valinn á kerfisbundinn hátt úr ýmsum ritum, dagblöðum, bókum, tölvupósti o.s.frv. Slíkur grunnur er nauðsynlegur fyrir flest verkefni í tungutækni og nauðsynlegt er að þjóðin eigi slíkan grunn. Meðal umsókna í desember 2001 var umsókn um að vinna slíkan grunn. Þá var ekki talið tímabært að styrkja hann þar sem hann byggist á betri beygingarlýsingu fyrir íslensku en þá var til og markara sem ekki var heldur til. Slík verkefni voru hins vegar styrkt 2001 og er nú lokið eins og fyrr var sagt. Í framhaldi af því var á árinu 2004 ákveðið að styrkja gerð íslenskrar málheildar. Vinna við hana er nú hafin og er áætlað að í grunninum verði 25 milljónir orða sem verða fullgreind í orðflokka og beygingarmyndir. Þegar þessu verkefni lýkur í júní 2007 má segja að viðunandi grunnur sé kominn að tungutækni hvað varðar texta. Þar sem tungumálið breytist sífellt mun þurfa að halda við textagrunnum og öðrum söfnum.

Vélrænar þýðingar eru mjög mikilvægur hluti tungutækni og gætu skipt miklu máli hér á landi. *Tungutækniverkefnið* hefur ekki styrkt þær, einkum þar sem ekki var talið að nauðsynlegur fræðilegur grunnur væri fyrir hendi til þess að ná árangri á þessu sviði. Slík verkefni eru einnig dýr og fjármagn verkefnisins takmarkað. *Tungutækniverkefnið* hefur hins vegar lagt mikilvægan grunn að vinnu við vélrænar þýðingar þar sem verkefni á því sviði þurfa gögn og þekkingu sem unnin hefur verið í verkefnum sem hafa verið styrkt.

Á árinu 2003 styrkti *Tungutækniverkefnið* endurgerð orðasafna Íslenskrar málstöðvar. Árangurinn má sjá á heimasíðu Íslenskrar málstöðvar (www.ismal.hi.is) og verkefninu er lýst á öðrum stað í þessu riti. Orðasöfnin

eru mikilvæg við þýðingar texta á mörgum fagsviðum, m.a. þegar hugbúnaður er þýddur yfir á íslensku. Markmiðið með verkefninu var að létta slík verk og bæta aðgengi að orðasöfnunum.

Ýmis önnur verk hafa verið unnin í verkefninu. Einkum hafa það verið verkefnisstjóri, Rögnvaldur Ólafsson, og formaður verkefnisstjórnar, Ari Arnalds, sem hafa sinnt ýmsum verkum tengdum markmiðum *Tungutækniverkefnisins*. Má þar m.a. nefna ýmiss konar kynningarstarf og fyrirlestrahald, aðstoð við Microsoft í sambandi við gerð leiðréttingarforrits fyrir íslensku, athuganir á möguleikum á þýðingum á viðmóti hugbúnaðar og fleira slíkt.

Menntun

Haustið 2002 hófst meistaranám í tungutækni við Háskóla Íslands fyrir atbeina *Tungutækniverkefnisins*. Námið heyrir undir íslenskuskor heimspekideildar og veitir Eiríkur Rögnvaldsson prófessor því forstöðu. Nemendur úr fyrsta árgangi þess náms unnu við fyrrnefnt Hjal-verkefni og gekk það samstarf nemenda og fyrirtækja mjög vel. Nú þegar hefur einn nemandi útskrifast frá Háskóla Íslands með meistarapróf í tungutækni.

Menntun í tungutækni hefur því komist í mun betra horf fyrir tilstuðlan *Tungutækniverkefnisins.*

Erlent samstarf

Að frumkvæði þáverandi menntamálaráðherra Íslands, Björns Bjarnasonar, setti Norræna ráðherranefndin árið 2000 á stofn norræna verkefnið Nordisk Sprogteknologi. (sjá www.norfa.no undir Sprogteknologi). Verkefnið vinnur að því að auka samstarf norrænu þjóðanna á sviði tungutækni og styrkir norræn samstarfsverkefni á því sviði. Fulltrúi Íslands í stjórn Nordisk Sprogteknologi hefur frá upphafi verið Rögnvaldur Ólafsson, verkefnisstjóri *Tungutækniverkefnisins*.

Fyrir frumkvæði Nordisk Sprogteknologi-verkefnisins var sótt um styrk til NORFA til þess að reka norrænan rannsóknaháskóla í tungutækni. Í lok árs 2003 fékkst styrkur að upphæð 5 milljónir norskra króna til fimm ára og skólinn hefur nú tekið til starfa (http://ngslt.org/). Háskóli Íslands er aðili að þessu verkefni. Næstu ár munu því íslenskir nemendur geta sótt um styrki til þess að stunda nám við Norræna tungutækniskólann og samstarfsskóla hans á Norðurlöndum.

Undanfari Norræna tungutækniskólans var að verkefnið Nordisk Sprogteknologi hefur undanfarin ár styrkt norræna nemendur til þess að nema við skóla á Norðurlöndum. Íslenskir nemendur í tungutækni hafa t.d. verið styrktir undanfarin ár til þess að sækja einstök námskeið við sænska Tungutækniskólann, GSLT (sjá http://www.gslt.hum.gu.se/nslp.html).

Kynning

Kynning á tungutækni og til hvers hún er nýtileg hefur verið allmikil á vegum *Tungutækniverkefnisins*. Þessi tækni er nú að verða þokkalega vel þekkt hér á landi.

Tungutækniverkefnið heldur úti heimasíðu, www.tungutaekni.is, með fréttum og upplýsingum um tungutækni. Gefinn var út kynningarbæklingur um verkefnið bæði á íslensku og ensku. Haldnir hafa verið nokkrir kynningarfundir og tvær ráðstefnur með erlendum fyrirlesurum. Verkefnisstjóri hefur einnig flutt fjölda fyrirlestra um *Tungutækniverkefnið*.

Í byrjun júlí 2001 hélt Guðrún Magnúsdóttir, forstjóri ESTeam í Aþenu, erindi í Odda um vélrænar þýðingar á vegum Tungutækniverkefnisins. Fyrirtæki Guðrúnar sérhæfir sig í að semja hugbúnað fyrir vélrænar þýðingar og hefur gengið mjög vel. Rúmlega hundrað manns sóttu fundinn. Fyrirlesturinn vakti athygli og í kjölfar hans birtust viðtöl við Guðrúnu í fjölmörgum íslenskum fjölmiðlum.

Í nóvember 2001 var haldin ráðstefna um tungutækni á vegum *Tungutækniverkefnisins*. Hún var nefnd Samspil tungu og tækni og var haldin í Salnum í Kópavogi. Tveir erlendir sérfræðingar héldu erindi á ráðstefnunni, Anders Nøklestad, tæknistjóri hjá Háskólanum í Osló, sem er sérfræðingur á sviði gagnagrunna fyrir texta, og nefndist fyrirlestur hans: *The Oslo Corpus – content, tagging, and interface* og Björn Granström frá KTH (Konunglega verkfræðiháskólanum) í Stokkhólmi, sem er sérfræðingur á sviði talaðs máls, en erindi hans hét: *Speech Technology - cooperation between academia and industry in Sweden*. Að auki fluttu fjórir íslenskir sérfræðingar erindi á ráðstefnunni. Ráðstefnuna sátu um 120 manns og tókst hún í alla staði vel.

Síðari ráðstefnan var haldin í lok maí 2003 í Háskóla Íslands. Hún var haldin í samstarfi við norræna tungutækniverkefnið, Nordisk Sprogteknologi, og í tengslum við norræna ráðstefnu um máltölvun, NODALIDA. Þar fluttu fjölmargir íslenskir og erlendir sérfræðingar erindi.

Þriðja ráðstefnan verður haldin 30. nóvember 2004. Þar verða kynnt þau verkefni sem hafa verið styrkt og sá árangur sem náðst hefur í *Tungutækniverkefninu*.

Lokaorð

Segja má að öllum helstu markmiðum *Tungutækniverkefnisins* hafi verið náð eða að þau náist á næstu árum þegar verkefnum sem styrkt hafa verið lýkur. Eins og fyrr sagði hafa um 133 milljónir króna verið veittar til verksins. Kostnaður við verkefnið hefur verið mun minni en ætlað var í upphafi. Kemur þar margt til, m.a. að ýtrustu hagsýni hefur verið gætt, reynt hefur verið að byggja sem mest á reynslu annarra þjóða og síðast en ekki síst hafa fyrirtæki og stofnanir lagt í verkið, fé, tíma og fyrri verk. Þótt mikill árangur hafi orðið og markmið

Tungutækniverkefnisins hafi náðst er enn margt ógert á sviði tungutækni, til dæmis vantar góðan íslenskan talgervil og vélrænar þýðingar á texta. Eigi íslenska að vera tungumál sem nýtir nýja tækni þarf áfram að rannsaka og þróa hvernig málið samhæfist tækni hvers tíma.

Verkefnisstjóri *Tungutækniverkefnisins*,

Rögnvaldur Ólafsson