SAMSPIL TUNGU OG TÆKNI

Afrakstur tungutækniverkefnis
menntamálaráðuneytisins

MENNTAMÁLARÁÐUNEYTIÐ

Menntamálaráðuneytið : rit 17
Nóvember 2004

Útgefandi: Menntamálaráðuneytið
Sölvhólsgötu 4, 150 Reykjavík
Sími: 545 9500
Bréfasími: 562 3068
Netfang: postur@mrn.stjr.is
Veffang: www.menntamalaraduneyti.is

Hönnun og umbrot: PR [pje err]
Prentun: Svansprent

ISBN 9979-777-19-2

Þorgerður Katrín Gunnarsdóttir

Ávarp menntamálaráðherra

Íslendingar hafa borið gæfu til að rækta sitt móðurmál. Tökuorð í íslensku eru miklum mun færri en í öðrum Norðurlandamálum sem er meðal annars að þakka virkri nýyrðasmíði. Á tímum alþjóðavæðingar og hraðrar þróunar í upplýsinga- og samskiptatækni er rík ástæða til að vinna áfram að því að tryggja stöðu íslensks máls til framtíðar. Þar þarf sameiginlega viðleitni stjórnvalda og stofnana, fræðimanna og fyrirtækja, áhugamanna og almennings.

Í heimi þar sem samskipti við tölvur munu í æ ríkari mæli fara fram með töluðu máli hefur tungutækniverkefni menntamálaráðuneytisins unnið að því að gera íslenskuna að lifandi tungumáli. Í riti þessu er rakinn árangur þeirra verkefna sem unnið hefur verið að. Frá upphafi hefur verið lögð áhersla á að leiða saman fyrirtæki og opinberar stofnanir til að vinna að hagnýtum verkefnum í tungutækni en jafnframt leitast við að leggja traustan fræðilegan grunn með rannsóknum. Tel ég að vel hafi tekist til og er það ekki síst að þakka góðu samstarfi opinberra stofnana og einkaaðila. Vil ég þakka verkefnisstjórn um tungutækni og öllum þeim sem unnið hafa að þessu verkefni fyrir vel unnin störf.

Nú þegar tungutækniverkefninu lýkur mun menntamálaráðuneytið styðja við samstarfsnet háskóla og fyrirtækja sem munu sjá um að viðhalda og miðla þeirri þekkingu og þeim gögnum sem þróuð hafa verið. Það er mat þeirra sem stýrt hafa verkefninu að lagður hafi verið góður grunnur að íslenskri tungutækni sem öflug rannsóknar- og þróunarstarfsemi geti byggt á í framtíðinni. Menntamálaráðuneytið mun áfram gera kröfu til þess að íslenskan verði sjálfsagður hluti af heimi upplýsinga og tækni og að komandi kynslóðir geti notið þess auðs sem í tungunni felst.

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

Listi yfir verkefni sem styrkt hafa verið.

Verkefni

Beygingakerfi – Málfræðigreinikerfi – Málfræðipúki
Umsækjandi: Friðrik Skúlason ehf.
Verkefnisstjóri: Maren Albertsdóttir
Verkefnin styrkt um 10 milljónir króna

Beygingarlýsing íslensks nútímamáls
Orðabók Háskólans
Edda hf. – miðlun og útgáfa
Verkefnisstjóri: Kristín Bjarnadóttir
Verkefnið styrkt um 10 milljónir króna

Endurbætt orðskiptiforrit Púkans
Umsækjandi: Friðrik Skúlason ehf.
Verkefnisstjóri: Maren Albertsdóttir
Verkefnið styrkt um 0,5 milljónir króna

Endurbætt tillögugerðar-forrit Púkans
Umsækjandi: Friðrik Skúlason ehf.
Verkefnisstjóri: Maren Albertsdóttir
Verkefnið styrkt um 1 milljón króna

Endurforritun orðabanka Íslenskrar málstöðvar
Íslensk málstöð
Verkefnisstjóri: Ari Páll Kristinsson
Verkefnið styrkt um 4,87 milljónir króna

Hagnýt notkun tungutækni í símtölvunarlausnum
Grunnur-Gagnalausnir ehf.
Verkefnisstjóri: Björn Jónsson
Verkefnið styrkt um 2,5 milljónir króna

Hjal
Landssími Íslands hf., Hex hugbúnaður ehf. Nýherji hf., Grunnur-Gagnalausnir ehf., Háskóli Íslands
Verkefnisstjóri: Sæmundur Þorsteinsson
Verkefnið styrkt um 14,82 milljónir króna

Málfræðilegur markari fyrir íslensku
Orðabók Háskólans
Málgreiningarhópurinn
Verkefnisstjóri: Sigrún Helgadóttir
Verkefnið styrkt um 6 milljónir króna

Mörkuð íslensk málheild
Orðabók Háskólans
Verkefnisstjóri: Sigrún Helgadóttir
Verkefnið styrkt um 18,5 milljónir króna

Talkennsl og texti í tal
Nýherji hf.
Verkefnisstjóri: Helgi Örn Viggósson
Verkefnið styrkt um 5 milljónir króna

Maren Albertsdóttir og
Stefán Einar Stefánsson

Beygingar- og málfræðigreinikerfi

Maren Albertsdóttir og Stefán Einar Stefánsson

Beygingar- og málfræðigreinikerfi

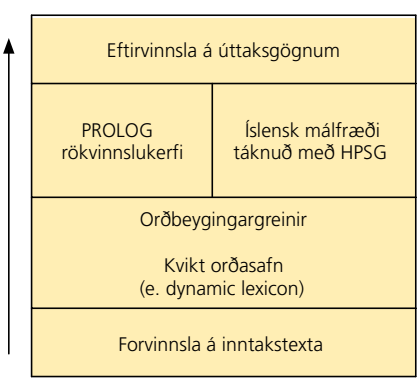
Markmið verkefnisins er að þróa kerfi sem tekur inn setningar á íslensku og skilar upplýsingum um byggingu þeirra og eiginleika einstakra hluta þeirra. Málfræðigreinikerfið nýtir sér beygingargreinikerfi sem tekur inn einstök orð og skilar út upplýsingum um orðflokka og beygingarmyndir. Kerfið skilar því upplýsingum um íslenskar setningar út frá beygingar- og setningafræði.

Verkefnið byggist á þekktum aðferðum sem hafa verið notaðar til að þróa sambærileg kerfi víða um heim fyrir ólík tungumál. Markmiðið er að byggja kerfið eingöngu á beygingar- og setningafræði en geyma merkingarfræði að eins miklu leyti og hægt er. Mörk setninga- og merkingarfræði eru oft óljós en sneitt er framhjá hreinum merkingarfræðilegum atriðum. Kerfið býður samt sem áður upp á að það verði þróað enn frekar út frá merkingarfræði í framtíðinni.

Verkefnið stendur enn yfir og er unnið af Stefáni Einari Stefánssyni og Maren Albertsdóttur ásamt Friðriki Skúlasyni. Eiríkur Rögnvaldsson, umsjónarmaður meistaranáms í tungutækni við Háskóla Íslands, hefur veitt sérfræðiaðstoð sína á verktímanum. Verkefnið hófst formlega í september 2002 og stendur enn. Í lok nóvember 2004 hafði um 36 mannmánuðum verið varið í verkefnið og eru verklok áætluð í maí 2005.

Undirstöðueiningar málfræðigreinikerfisins eru:

- Beygingargreinikerfi eða kvikt orðasafn (e. Dynamic lexicon) sem vinnur með föstu orðasafni
- Íslensk málfræði skilgreind í HPSG (e. Head Driven Phrase Structure Grammar)
- Prolog-rökvinnslukerfi



Undirstöðueiningar

Kvikt orðasafn

Í málfræðigreinikerfinu er stuðst við svokallað kvikt orðasafn (e. dynamic lexicon). Þetta þýðir að í stað þess að geyma langan orðalista með beygingarupplýsingum (auk annarra málfræðiupplýsinga) fyrir hvert orð þá eru þessar upplýsingar sóttar eftir þörfum úr beygingargreininum sem aflar þeirra með reiknifræðilegum hætti. Kostir þessarar aðferðar eru miklir en einn sá helsti er að minnisstærð orðasafnsins minnkar verulega. Þetta á sér í lagi við í íslensku þar sem sama orðið getur haft mjög margar myndir.

Kvika orðasafnið inniheldur um 30.000 einingar sem það getur notað til að mynda ný orð út frá íslenskum beygingar- og orðmyndunarreglum. Orðasafnið er því í raun óendanlegt. Með kvika orðasafninu er notað fast orðasafn til að kljást við undantekningar og sértilfelli.

Orðasöfn sambærilegra kerfa eru oftast byggð upp á annan hátt. Beygingarfræðileg atriði eru oft leyst í málfræðihlutanum ásamt setninga- og merkingarfræði auk þess sem meira er stuðst við föst orðasöfn. Kvika orðasafnið sem notast er við í verkefninu vinnur hins vegar sína vinnu óháð málfræðikerfinu sem slíku. Það er mjög öflugt, nákvæmt og hraðvirkt sem gerir kerfið í heild skilvirkara og auðveldara í viðhaldi og vexti.

HPSG

HPSG (e. Head Driven Phrase Structure Grammar) er málfræði sem auðvelt er að umrita á form sem tölvur skilja. HPSG vinnur með málfræði út frá setninga- og merkingarfræði og sækir ýmislegt í aðrar þekktar málfræðikenningar. Að auki mynda tölvunarfræði, stærðfræði og fleiri skyldar greinar fræðilegan grundvöll undir HPSG.

Sérhvert orð, sem og stærri liðir, fá úthlutað sérstökum ramma í málfræðinni þar sem allar beygingar-, setninga- og merkingarfræðilegar (ef einhverjar eru) upplýsingar koma fram. Heimur málfræðinnar er skilgreindur með nokkurskonar erfðastigveldi sem segir til um af hvaða tagi ákveðinn hlutur er, hvernig hann tengist öðrum hlutum í kerfinu og hvaða skorður hann setur. Hvert orð fær merkimiða frá orðasafninu, einn frá beygingargreinikerfi og annan frá málfræðinni sem segir til um setningafræðilegt hlutverk þess. Reglur, frumsendur og skorður sjá síðan um að binda saman orðin.

Málfræðifyrirbæri í kerfinu er sett fram í römmum sem tákna í raun skorður. Málfræðin samanstendur því af safni af skorðum. Sem dæmi má nefna að sagnorð setja skorður á umhverfi sitt (nærliggjandi orð og/eða liði) þar sem þau vilja t.d. stjórna í hvaða falli frumlag og fyllilíðir eru sem og af hvaða tagi orðin eru, dæmi: ‘Mig (ÞF) langar’ en ekki *’Mér (ÞGF) langar’.

Rökvinnsluvélin

Rökvinnsluvélin reiknar út allar mögulegar greiningar á setningunni sem hún tekur inn samkvæmt þeim skorðum sem málfræðinsetur (e. All-paths parsing). Allar greiningar eru gefnar, sama hversu líklegar þær eru. Útreikningarnir sem vélin framkvæmir eru þar af leiðandi mjög þungir. Kerfið gefur möguleika á að takmarka þær greiningar sem koma til greina t.d. með því að setja inn tölfræðisíu sem velur aðeins „líklegustu" greiningarnar.

Allt þróunarumhverfi hefur verið þróað innan fyrirtækisins.

Afrakstur

Markmiðið með verkefninu var ekki að þróa ákveðið notendakerfi heldur fyrst og fremst að til yrði grunnkerfi sem nýttist í frekari þróunarvinnu á sviði máltækni, gerð kennsluefnis og/eða -forrita eða í hverskyns rannsóknum á íslensku máli. Málfræðigreinikerfið er undirstöðukerfi við þróun í framtíðinni á ýmiskonar tölvubúnaði á sviði máltækni. Kerfið gefur svokallaða djúpgreiningu á texta sem er mjög nákvæm beygingar- og setningafræðileg greining. Aðferðafræði, eins og notuð er í kerfinu, getur nýst við gerð leiðréttingarforrita, talgreina, talgervla og þýðingarvéla.

Verkefnið hefur nú þegar skilað margvíslegum tólum og þekkingu á sviði máltækni. Þróaður hefur verið ýmis hugbúnaður og hugbúnaðareiningar í tengslum við verkefnið ásamt málfræðihugbúnaði (reiknivél). Hluti af íslenskri málfræði hefur verið skilgreindur í HPSG. Byggt hefur verið upp öflugt orðasafn sem byggist á beygingar- og setningafræði og samanstendur af kviku og föstu orðasafni. Í vinnuferlinu hefur skapast þekking á smíði víðtæks málfræðikerfis, ákveðnu verklagi og aðferðafræði sem spilar stórt hlutverk í þróun íslenskra máltæknitóla.

Verkefnið hefur þróast í að verða meira rannsóknarverkefni en búist var við í upphafi. Þess má geta að unnið er að verkefnum víða um heim sem byggjast á sömu aðferðum m.a. fyrir ensku, þýsku og japönsku. Mestöll þróunar- og grunnvinna fer fram í háskólum en er styrkt af opinberum aðilum og fyrirtækjum sem síðan hagnýta hana. Hér á landi hafa hvorki verið stundaðar sambærilegar rannsóknir á sviði máltækni né þeim hluta sem snýr að tölvunarfræði og tæknilegri útfærslu. Að því leytinu blasir allt annað landslag við þeim sem vilja byggja upp máltæknitól fyrir íslensku en víða annarsstaðar. Það er von þátttakenda að sú þekking og reynsla sem hefur skapast og áunnist á þessum tíma skili sér áfram með einhverjum hætti.

Kristín Bjarnadóttir

Beygingarlýsing íslensks nútímamáls

Kristín Bjarnadóttir

Beygingarlýsing íslensks nútímamáls

Útgáfa 1.0

Orðabók Háskólans og Edda hf. sóttu um styrk til verkefnisstjórnar menntamálaráðuneytisins í tungutækni árið 2002 til að vinna að beygingarlýsingu íslensks nútímamáls. Gengið var frá samningi um verkið 23. ágúst 2002 og því lauk 13. febrúar 2004 þegar menntamálaráðuneytinu var afhentur geisladiskur með útgáfu 1.0 af beygingarlýsingunni, alls 173.389 beygingardæmi á formi xml-skráa.

Markmiðið með verkefninu var að koma upp beygingarlýsingu á tölvutæku formi til nota í ýmiss konar tungutækniverkefni en ítarleg beygingarlýsing er grundvöllur að vélrænni greiningu á íslenskum textum, nauðsynlegur undanfari orðflokkagreiningar og setningagreiningar. Beygingarlýsingin nýtist t.d. við mörkun texta, við gerð leitarvéla, leiðréttingar- og þýðingarforrita, auk þess að vera forsenda skilvirkrar orðabókargerðar og heimildasöfnunar um tungumálið.

Beygingarlýsingin er þegar notuð í ýmsum verkefnum Orðabókar Háskólans, t.d. ISLEX, sem er samnorrænt verkefni um íslensk-skandinavískar orðabækur, Markaðri íslenskri málheild og við leit í Textasafni Orðabókarinnar. Þá er hún notuð í verkefninu Icelandic Online, sem er samstarfsverkefni Háskóla Íslands og Háskólans í Wisconsin.

Orðabók Háskólans sá að öllu leyti um vinnuna við beygingarlýsinguna og lagði til húsnæði, alla aðstöðu og aðgang að gögnum. Edda hf. lagði til beygingarlýsingu þá sem unnin var fyrir tölvuútgáfu *Íslenskrar orðabókar* (2002). Starfsmenn voru Kristín Bjarnadóttir (verkefnisstjóri), Þórdís Úlfarsdóttir, Auður Þórunn Rögnvaldsdóttir, Ragnhildur Hrönn Sigurðardóttir og Aðalsteinn Eyþórsson.

Orðaforðinn í beygingarlýsingunni er fenginn úr 3. útgáfu *Íslenskrar orðabókar* og úr söfnum Orðabókarinnar. Helstu heimildir við rannsóknir á einstökum orðum og beygingarflokkum eru Ritmálsskrá og Textasafn Orðabókarinnar, auk handbóka, greina og ritgerða um íslenska málfræði.

Útgáfa 2.0

Beygingarlýsing íslensks nútímamáls, útgáfa 2.0, hefur nú verið birt á vefsíðu Orðabókarinnar og er hún unnin upp úr útgáfu 1.0 með talsverðum endurbótum, leiðréttingum og nokkrum viðbótum.

Heildarorðaforðinn í útgáfu 2.0 skiptist svo:

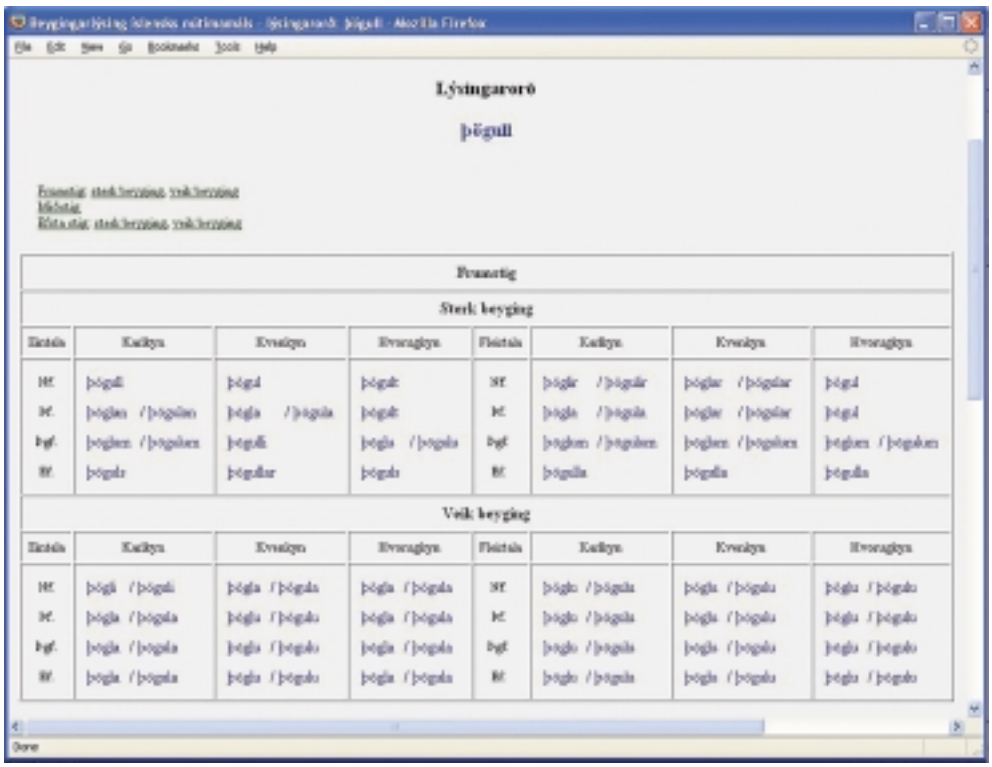
Skipting beygingardæma eftir orðflokkum:

	Orð	Orðmyndir	Meðalfjöldi orðmynda
Nafnorð	137.300	1.960.700	14,3
Lýsingarorð	26.300	2.334.900	88,8
Sagnir	7.600	538.200	70,8
Mannanöfn	4.760	19.445	4,1
Annað	70	849	12,1
Alls	176.030	4.854.094	27,6

Hér er beygingardæmi nafnorðsins hnífur, eins og það birtist á vefsíðu Orðabókarinnar.

Beygingardæmin

Takmarkið við birtingu beygingardæmanna er að einskorða efnið við raunverulegar myndir hvers orðs, þ.e. að sýna afbrigði þar sem það á við (t.d. í þágufalli eintölu af nafnorðinu *hnífur: hníf/hnífi*) og eyður þar sem beygingarmyndir koma aldrei fyrir (t.d. er ekki til miðmynd af sögninni *auðvelda*). Án afbrigða eru beygingarmyndir nafnorðs 16, þ.e. fjögur föll í eintölu og fleirtölu, án greinis og með greini. Beygingarmyndir sagnar í persónuhætti eru 48, auk boðháttar og lýsingarhátta, en alls geta beygingarmyndir hverrar sagnar orðið 106. Beygingarmyndir lýsingarorðs sem tekur stigbreytingu eru allt að 120, án afbrigða.



Beygingarmyndum getur fjölgað talsvert þegar afbrigði eru sýnd, eins og sjá má af hluta beygingardæmis lýsingarorðsins þögull.

Framtíðarverkefnið

Í tengslum við vinnu við markaða íslenska málheild er fyrirhugað að bæta talsverðu efni við beygingarlýsinguna, t.d. örnefnum og fyrirtækja- og stofnanaheitum, auk viðbóta við almennan orðaforða. Þá berast ábendingar um viðbætur og leiðréttingar frá notendum vefsíðunnar og er þeim tekið með þökkum. Eðli málsins samkvæmt lýkur verkefni af þessu tagi aldrei þar sem fengist er við lifandi mál.

Vegna tíðra endurbóta er mikilvægt að þeir sem nota beygingarlýsinguna hafi samband við Orðabók Háskólans til að nálgast nýjustu útgáfu.

Friðrik Skúlason ehf.

Endurbætt tillögugerðar- og orðskiptiforrit Púka

Friðrik Skúlason ehf.

Endurbætt tillögugerðar- og orðskiptiforrit Púka

Með endurbótum á tillögugerðar- og orðskiptiforriti Púka var leitast við að bæta Púka og festa hann þannig í sessi sem öflugt íslenskt tungutæknitól fyrir almenna notendur. Endurbæturnar skiluðu sér inn í nýjustu útgáfu forritsins, Púka 2003. Verkefnin, sem tóku um fimm mannmánuði, voru unnin af Friðriki Skúlasyni.

Sérstaða Púka sem leiðréttingarforrits fyrir íslensku felst í sveigjanleika og nákvæmni við villuleit. Ólíkt öðrum stafsetningarforritum notar Púkinn mjög öfluga aðferð þar sem hann leiðir út orðmyndir frá grunnmynd orða, með öðrum orðum hann kann íslenskar beygingar- og orðmyndunarreglur. Auk þess geymir Púki ýmsar nánari upplýsingar um orðin sem varða m.a. merkingarfræði og nýtast við villuleit. Hefðbundin stafsetningarforrit sem byggjast á stórum uppflettiorðasöfnum eða tölfræðilegum aðferðum greina ekki á milli orðanna á þennan hátt sem eykur hættuna á að þeim yfirsjáist villur og þekki ekki rétt mynduð orð.

Sú vitneskja sem Púki býr yfir gerir það að verkum að hann yfirfer textann með mun betri árangri en almennt tíðkast, tillögugerð hans er nákvæmari og hann býður auk þess upp á þann möguleika að beygja orð, skipta orðum á milli lína og finna samheiti orða.

Púki hefur styrkt stöðu sína enn frekar með þeim endurbótum sem voru gerðar fyrir tilstilli tungutæknisjóðs.

Endurbætt tillögugerðarforrit Púkans

Púki er leiðréttingarforrit Friðriks Skúlasonar ehf. Púki skoðar eitt og eitt orð í senn og athugar hvort það sé rétt stafsett. Þegar hann finnur rangt orð inni í texta eða orð sem hann þekkir ekki staðnæmist hann við það og kemur með tillögur að réttu orði. Mörg orð eru tví- eða margræð og því getur reynst erfitt í sumum tilfellum að láta Púka stinga upp á því orði sem notandi ætlaði að skrifa. Ástæðan er sú að Púki leggur til orð út frá útliti orðanna en getur ekki metið efni eða innihald textans.

Tillögugerð Púka leyfði öll rétt mynduð samsett orð í íslensku óháð því hvort þau voru í raun notuð eða ekki. Þetta gerði það

að verkum að tillögugerðin stakk oft upp á orðum sem þóttu langsótt. Endurbætur á tillögugerð Púka fólust annarsvegar í að fækka bulltillögum sem Púki kom með þegar hann fann rangt stafsett orð og hinsvegar að bæta orðum við tillögugerðina þar sem ástæða þótti til.

Við endurbætur á tillögugerð var m.a. notast við textabanka sem byggist á orðasafni Morgunblaðsins frá síðustu þremur árum (u.þ.b. fjórar milljónir setninga – 450 MB). Endurbætur á Púka, sem varða bæði tillögugerð og orðskiptiforrit, hafa tvöfaldað orðasafn hans.

Helstu verkþættir:

1. Orðasafn bætt: Púki þekkti ekki viðkomandi orð og kom því með bullorð sem tillögur. Þetta var lagfært með því að bæta orðaforða Púka.

2. Bullorðum fækkað: Púki samþykkti ýmis orð sem eru í raun leyfileg út frá íslenskum orðmyndunarreglum en eru samt sem áður bullorð eða jafnvel erlend. Þetta var lagfært með því að taka um 50 eins til tveggja stafa orð t.d. 'il', 'te' 'á' og 'ari' og leyfa tillögugerðinni ekki að nota þau í seinni hluta samsettra orða þar sem þau víkka tillögugerðina of mikið.

3. Notkun viðskeyta þrengd: Viðskeyti sem finnast aðeins í ákveðnum orðum voru yfirfarin til að koma í veg fyrir að þau yrðu leyfð í öllum samsetningum.

4. Algeng orð sett inn sem ein heild: Um eitt þúsund algengustu samsettu orðin úr textabankanum voru sett inn sem ein heild. Ástæðan er sú að tillögugerðin notar fyrst orð sem eru í einu lagi. Ef hún finnur ekkert slíkt notar hún þau orð sem eru möguleg úr tveimur hlutum o.s.frv. Með því að setja algeng orð inn í einu lagi tekur tillögugerðin þessi orð fram yfir samsett orð sem minnkar líkurnar á bulltillögum.

5. Fleiri möguleikar prófaðir: Tillögugerðin prófar fleiri möguleika en áður. Ástæðan er sú að Púki var uppbyggður á þann hátt að hann gerði ráð fyrir aðeins einni villu í hverju orði. Nú prófar hann mun fleiri möguleika sem víkkar tillögugerðina til muna.

Endurbætt orðskiptiforrit Púkans

Orðskiptiforrit Púka var bætt með það að markmiði að Púki gæti skipt öllum orðum rétt og veldi alltaf líklegustu skiptinguna hverju sinni.

Helsti galli orðskiptiforritsins var að það skipti ekki orðum sem voru geymd heil í orðasafninu. Einnig skipti hann í einstaka tilfellum orðum vitlaust miðað við hefðbundna skiptingu, þá aðallega samsettum orðum. Sum orð eru tví- eða margræð og hægt að skipta þeim á fleiri en einn hátt. Sumir möguleikar eru þó ansi langsóttir þótt skiptingin sé möguleg, þessum tilfellum var fækkað til muna.

Helstu verkþættir:

1. Skipting sett í heil samsett orð: Orðasafnið var yfirfarið m.t.t. samsettra orða sem eru geymd heil og settar inn viðeigandi skiptingar. Áður skipti Púki einungis orðum sem voru samsett. Ef samsett orð var geymt heilt í orðasafni skipti hann því ekki.

2. Skiptimerki sett inn: Viðeigandi skiptimerki voru sett inn í ákveðin samsett orð. Orð með viðskeyttum greini voru t.d. geymd heil áður og því skipti Púki þeim ekki en gerir nú. Skiptingar á samsetningum hafa forgang í Púka. Skiptingar á atkvæðum eru ekki leyfilegar. Púki er byggður inn í enska Púkann sem leyfir ekki val þarna á milli eða hvorttveggja. Skiptingar á samsetningum eru teknar fram yfir atkvæðaskiptingu þar sem langflestir notendur Púka vilja skipta á þann hátt í ritvinnslu.

3. Líklegasta skiptingin valin: Yfirferð á þeim orðum sem eru tví- eða margræð og gefa möguleika á tveimur eða fleiri mismunandi skiptingum er lokið þar sem líklegasta skiptingin fyrir hvert orð var valin þar sem möguleiki var á.

Dæmi: ‘sjóðs-láni’ – ‘sjóð-sláni’
‘fisk-afli’ –‘fis-kafli’

Púki getur ekki skoðað setningar í heild heldur aðeins eitt orð í einu. Þó er unnt er að ná góðum árangri í skiptingum sem varða tví- eða margræðni. Í einhverjum tilvikum er engan veginn hægt að gera upp á milli möguleika. Sumt af því mun Púki ráða við í framtíðinni, þ.e. þau atriði sem setningafræðin ræður við, dæmi: ‘heims-enda’ eða ‘heim-senda’. Önnur vandamál í orðskiptingum sem varða merkingarfræði og samhengi munu seint verða leyst.

Ari Páll Kristinsson

Endurforritun orðabanka Íslenskrar málstöðvar

Ari Páll Kristinsson

Endurforritun orðabanka Íslenskrar málstöðvar

Eitt af helstu verkefnum Íslenskrar málstöðvar er að láta íslenskt íðorðastarf til sín taka á ýmsan hátt með margháttuðum stuðningi við orðanefndir og sérfræðinga sem sinna orðaforða sérgreina sinna. Í nóvember 1997 var stigið mikilsvert framfaraspor til að auðvelda útgáfu og birtingu íðorðasafna: orðabanki Íslenskrar málstöðvar var tekinn í gagnið til að birta íðorðasöfn á vefnum. Orðabankinn hefur að geyma sérhæfðan hugbúnað fyrir skráningu og birtingu orðasafna. Meginhlutverk hans er að safna saman íðorðum, þ.e. orðum sem bundin eru tilteknum fræðigreinum og starfsgreinum, á íslensku og fleiri tungumálum, og birta þau þannig að nýtist sem flestum.

Íslenskt íðorðastarf byggist á almennum áhuga á málrækt í samfélaginu. Frumkvæði að gerð íðorðasafna kemur því jafnan frá sérfræðingum sem áhuga hafa á íslensku máli og vilja geta rætt og ritað um sérgrein sína á móðurmálinu. Margar íðorðanefndir hafa verið myndaðar sem sumar hverjar hafa um áratugaskeið staðið að umfangsmikilli útgáfu. Nú eru orðasöfn þeirra tiltæk í orðabankanum og auðvelt að leita í þeim öllum í einu.

Íslensk málstöð er þeim til aðstoðar sem vinna slíkt íðorðastarf, hvort sem um er að ræða orðanefndir eða einstaklinga. Í tengslum við Íslenska málstöð er íðorðastarf unnið á fræðasviðum sem skipta mörgum tugum. Efni orðabankans er því afar fjölbreytilegt eins og sjá má af þessum lista um svið orðasafna í birtingarhluta orðabankans: bílorð, byggingarlist, byggingarverkfræði, eðlisfræði, efnafræði, endurskoðun, erfðafræði, flugorð, fundarorð, gjaldmiðlaheiti, hagfræði, iðjuþjálfun, jarðfræði, landafræði, landupplýsingar, líforðasafn, læknisfræði, matarorð úr jurtaríkinu, málfræði, málmiðnaður, nytjaviðir, ónæmisfræði, plöntuheiti og ættir háplantna, raftækniorð, ríkjaheiti, sjávardýr, sjávarútvegsmál, sjómennsku- og vélfræðiorð, stjórnmálafræði, stjórnsýsluorð, stjörnufræði, timburorð, tölfræði, tölvuorð, umhverfisfræði, uppeldis- og sálarfræði, upplýsingafræði, veðurorð, verkefnastjórnun, þýðingafræði; auk réttritunarorðabókar og nýyrðadagbókar Íslenskrar málstöðvar.

Mikil gróska varð í íðorðastarfi eftir að orðabanki Íslenskrar málstöðvar var opnaður. Þá gerðu einnig vart við sig ýmsir sem áhuga

höfðu á íðorðastarfi í sérgrein sinni og áttu jafnvel orðasafn í fórum sínum. Margir hafa sýnt orðabankanum áhuga og hafa boðið fram íðorðasöfn án þess að leitað hafi verið eftir því sérstaklega. Ljóst er að orðabankinn hefur átt mikinn þátt í því að kynna íðorðastarf og veitt ýmsum möguleika á því að koma orðaforðanum á framfæri.

Íslensk málstöð gerir samning við höfunda orðasafna um að þeir fái endurgjaldslausan aðgang að skráningarkerfi orðabankans til þess að vinna að söfnum sínum. Sem mótframlag höfunda kemur svo að málstöðin má birta orðasöfnin í birtingarhluta orðabankans á Netinu þegar þau teljast tilbúin. Höfundar hafa eftir sem áður allan rétt til verka sinna og geta birt þau hvar og hvenær sem þeir vilja.

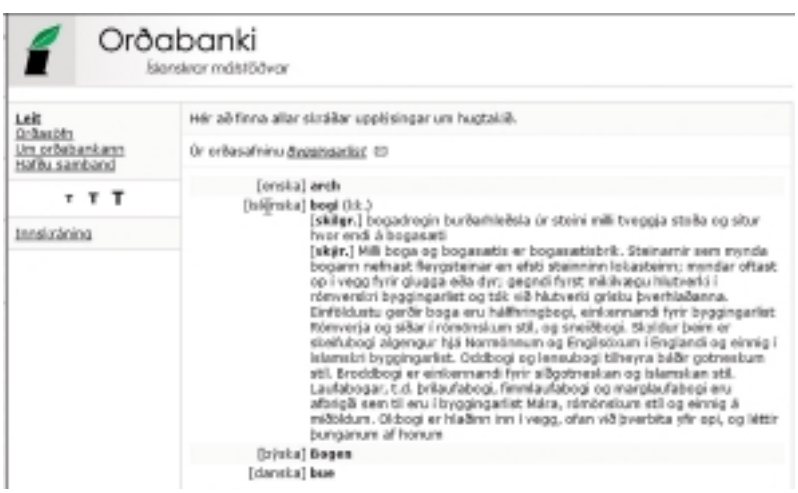
Orðabankinn skiptist í tvo hluta. Svokallaður birtingarhluti orðabankans er hinn sýnilegi orðabanki, þ.e. sá hluti bankans sem allir geta leitað í. Vinnsluhluti orðabankans er ætlaður til orðabókarsmíða og er skráningarkerfi fyrir orðasöfn. Hann skiptist í svæði sem hvert og eitt tilheyrir höfundi tiltekins orðasafns. Í vinnsluhlutanum geta bæði verið orðasöfn í frumvinnslu og í endurskoðun. Almennir notendur hafa engan aðgang að vinnsluhlutanum.

Viðmót orðabankans er ekki aðeins á íslensku heldur einnig á ensku og á öllum Norðurlandamálunum og mörg orðasafnanna eru með þýðingum á fleiri en einu tungumáli. Íslenska er algengasta tungumálið í orðabankanum og því næst enska. Vegna þess að orðabankinn er á Netinu getur fólk um allan heim bæði flett upp í orðasöfnunum og unnið að gerð nýrra safna á vinnusvæðum sínum í honum.

Upphafssíða birtingarhlutans lítur svona út að loknu verkefninu:

Leitum t.d. að íslenska orðinu þorskur og sjáum hvað birtist:

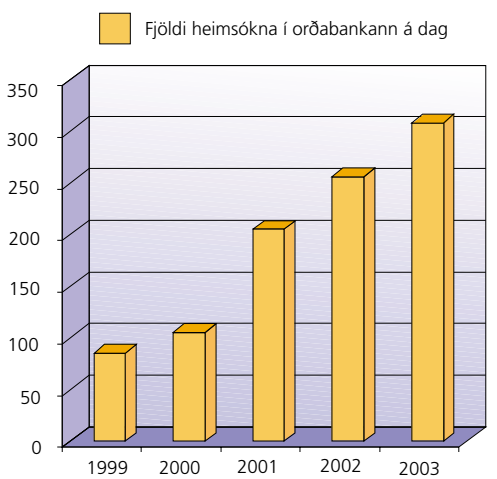
Með einni leit í orðabankanum er hægt að leita samtímis í öllum orðasöfnum hans. Í orðabankanum fæst þannig góð yfirsýn yfir mörg svið. Sá sem leitar í orðabankanum gæti einnig valið að takmarka leit sína strax í upphafi við ákveðið orðasafn. Eins er hægt að leita í öllum söfnum í upphafi og velja svo áfram eitt ákveðið orðasafn. Ef leitað er t.d. að orðinu *bogi* í öllum söfnum orðabankans í einu koma 12 niðurstöður. Ef svo *Orðasafn um byggingarlist* er valið sérstaklega koma aðeins upplýsingar um hugtakið úr því safni eins og sjá má á eftirfarandi mynd:



Þegar orðabankinn var opnaður 15. nóvember 1997 voru í honum 14 orðasöfn en nú, þegar orðabankinn fagnar 7 ára afmæli sínu, eru þau orðin 47 talsins í birtingarhlutanum. Þeim til viðbótar eru 24 önnur orðasöfn í vinnslu og væntanlega munu mörg þeirra birtast eitt af öðru á næstu árum. Sem dæmi um væntanlegar viðbætur má nefna eftirtaldar sérgreinar: fornleifafræði, hagrannsóknir, myndlist, tannlækningar og lyfjafræði.

Eftir því sem orðasöfnum fjölgar í orðabankanum fjölgar hugtökum hans. Í orðabankanum eru nú hátt í 170.000 hugtök. Heiti (íðorð) í orðabankanum eru raunar mun fleiri enda eru þau jafnan fleiri en hugtökin. (Til samanburðar má nefna að í prentaðri *Íslenskri orðabók* frá Eddu eru um 90.000 uppflettiorð.)

Aðsóknin að orðabankanum eykst hröðum skrefum ár frá ári eins og sést glöggt á þessari mynd sem sýnir fjölda heimsókna á dag að meðaltali.



Uppflettingar í orðabankanum eru að sjálfsögðu miklu fleiri en heimsóknirnar því að notendur fletta oft mörgum atriðum upp í hverri lotu. Að meðaltali er flett 6 sinnum í hverri heimsókn í orðabankann. Árið 2003 voru heimsóknir í orðabankann 110.610 eða 303 á dag að meðaltali. Uppflettingar í orðabankanum 2003 voru samtals 616.845 en það jafngildir 1.690 uppflettingum að jafnaði á dag allt árið.

Endurforritunarverkefnið

Reynslan af orðabankanum hefur verið mjög góð en eigi að síður var orðið nauðsynlegt að endurbæta orðabankann með tilliti til þeirra framfara sem orðið hafa í smíði vefforrita. Því var ráðist í það að vinna að endurbótum á orðabankanum. Myndirnar hér á undan sýna útlit orðabankans eftir breytingarnar.

Tungutækniverkefni á vegum menntamálaráðuneytisins fjármagnaði endurforritun orðabankans 2003-2004. Það hófst í maí 2003 og verður nánast lokið í árslok 2004. Það sem eftir stendur þá verður einkum gæðaeftirlit, eftirfylgni og öryggisprófanir sem áætlaðar eru 2005-2006. Auk endurforritunarinnar sjálfrar, sem skipt var í grunnverkþætti og aukaverkþætti, studdi tungutækniverkefnið tvo stoðþætti fyrir orðastarfið, annars vegar útgáfu leiðbeininga um íslenska orðmyndun og hins vegar íslenska þýðingu á leiðbeiningarritinu *Guide to terminology* eftir finnska íðorðafræðinginn Heidi Suonuuti. Um er að ræða hand-

hægar og hagnýtar leiðbeiningar um árangursríkt íðorðastarf og samræmdan frágang íðorðasafna sem styðjast við alþjóðlega staðla.

Þegar má sjá árangur verkefnisins í nýju útliti, nýjum leitarmöguleikum, prentmöguleikum og öðrum umbótum sem snúa að ritstjórum einstakra íðorðasafna. Ný útgáfa orðabankans byggist að miklu leyti á tækninýjungum, t.a.m. er gagnagrunnurinn mun sveigjanlegri en gagnageymsla eldri útgáfunnar. Endurbæturnar snúa fyrst og fremst að vinnsluhluta orðabankans og þær koma því þeim vel að notum sem vinna að gerð orðasafna í orðabankanum. Skráningar- og leitarmöguleikar hafa orðið betri en áður og auðveldara er að færa inn breytingar á skrám. Hægt er að færa inn leiðréttingar á einstökum orðum án þess að uppfæra orðasafn í heild eins og áður þurfti að gera. Það kemur sér t.d. mjög vel þegar unnið er að breytingum á orðasafni. Auðveldara er að fá fram allar tölfræðilegar upplýsingar nú en áður. Fyrir ritstjóra orðabankans er einnig hagræði að nýju útgáfunni. Hann á nú auðveldara en áður með aðgerðir og þarf minna að leita til kerfisstjóra. Í því felst umtalsverður sparnaður fyrir Íslenska málstöð sem á eftir að skila sér vel um ókomin ár.

Verktaki við endurforritunina var Guðmundur Ingi Gunnarsson. Af hálfu Íslenskrar málstöðvar komu að verkefninu Ari Páll Kristinsson, Ágústa Þorbergsdóttir og Dóra Hafsteinsdóttir.

(Í samantekt þessari er m.a. stuðst við grein Ágústu Þorbergsdóttur í 22. hefti tímaritsins *Málfregna*, 2003, um orðabanka Íslenskrar málstöðvar.)

Björn Jónsson

Hagnýt notkun tungutækni í símtölvunarlausnum

Björn Jónsson

Hagnýt notkun tungutækni í símtölvunarlausnum

Inngangur

Grunnur fékk 2,5 milljónir í styrk frá Tungutæknisjóði í framhaldi af umsókn, þar sem heiti verkefnisins var „Hagnýt notkun tungutækni í símtölvunarlausnum“. Markmið verkefnisins var að kanna hagnýtingu tungutæknieininga, sem til eru fyrir íslensku, m.a. í kerfum sem Grunnur hefur þróað og eru í almennri notkun.

Notast skyldi við þá talgreina og talgervla sem aðgangur væri að. Gert var ráð fyrir að notast við talgervil (Snorri) sem til er fyrir íslensku. Til samanburðar var einnig nýttur enskur talgervill frá Microsoft. Við upphaf verkefnisins var ekki til íslenskur talgreinir en stefnt var að því að nýta *open source*-talgreini frá Carnigie Mellon og kenna honum einföld íslensk orð (já, nei, tölur o.s.frv.) sem nauðsynleg eru í símtölvunarlausnum. Með tilkomu Hjal-verkefnisins og þeirrar ákvörðunar Landssíma Íslands að kaupa Scansoft-talgreini með íslensku sem byggist á niðurstöðum Hjal-verkefnisins var eðlilega tekin ákvörðun um að byggja á Scansoft-talgreininum í staðinn. Notast var við íslensku útgáfuna, en sú enska var einnig notuð til hliðsjónar.

Ákveðið var að útbúa einfalt talsetur (e. Voice Portal) sem innihéldi dæmi um virkni eins og upplestur frétta, upplýsingar um flug, uppflettingu í símaskrá, kaup á vöru og ýmislegt fleira. Því til viðbótar skyldi nýta þessa tækni í tengslum við Tímon og Snældu sem eru símtölvunarlausnir sem Grunnur hafði þróað. Á verkefnatímanum sameinaðist hugbúnaðarsvið Grunns Trackwell undir nafni Trackwell og var verkefninu haldið áfram í nafni Trackwell.

Forritun var að mestu unnin í Visual C++ 6.0. Þetta gerði okkur kleift að nota talgreinana á lágu plani, þ.e. án millilaga. Að sumu leyti er auðveldara að forrita tungutækniforrit í málum, s.s. Vxml (voice-xml), og útbjuggum við einnig ákveðin tungutæknidæmi með VoiceXML til samanburðar. VoiceXML hentar ágætlega þegar útbúa skal einföld kerfi, en þegar útbúa þarf flóknari kerfi teljum við að alvöru forritunarmál eins og C++ henti betur. Meðal annarra þægilegra möguleika við gerð svona kerfa má nefna SALT (speech-application-langu-

age-tags). Notkun C++ gerði okkur tæknilega mögulegt að rannsaka betur hina ýmsu möguleika tungutækninnar, þar sem við höfum betri stjórn á því sem verið er að vinna með.

Helstu niðurstöður

Gerð þessara lausna tókst vel, en reyndist töluvert meiri vinna en áætlað var, auk þess sem yfir ákveðinn þröskuld var að fara til að byrja með, sérstaklega hvað varðar notkun talgreinanna. Annar tæknilegur þröskuldur var samvirkni talgreinisins og svarþjónskjarna Grunns. Umbreyta þurfti svarþjónskjarnanum til að hann gæti haft talviðmót til viðbótar við hefðbundið DTMF-viðmót (veldu 1 fyrir X, veldu 2 o.s.frv.). Þetta var leyst almennt og er því svarkerfiskjarninn tilbúinn til notkunar fyrir þróun talgreinalausna. Slík samþætting er hins vegar einfaldari ef notast er við VoiceXML. Í öðru lagi kom í ljós, eins og raunar vitað var, að talgreinir er í eðli sínu þung keyrsla í samanburði við t.d. svarþjónskerfið sem byggist á DTMF-viðmóti.

Það kom líka skýrt í ljós að notkun tungutæknieininga með símtölvunarlausnum krefst ekki eingöngu forritunarhæfni heldur er þekking á málfræðiþættinum nauðsynleg.

Talgervlar

Notkun talgervla er tæknilega einföld, sér í lagi þeir ef þeir uppfylla SAPI 5-tungutæknistaðalinn. Vandamálið er að íslenski talgervilinn Snorri er lélegur, reyndar það vondur að deila má um hvort hann sé nothæfur í almennar lausnir. Snorri venst reyndar og smám saman læra notendur að skilja hann. Undantekningalítið sögðust þó nýir notendur eiga erfitt með að skilja hann. Í það minnsta er hann fráhrindandi og ráðleggjum við þeim sem útbúa vilja símtölvunarlausnir að notast við samsettar og forupptaknar hljóðskrár þar sem slíkt er mögulegt.

Talgreinar

Langmesta vinnan fór í notkun talgreina og fékkst mjög góð reynsla af notkun þessara lausna sem á eftir að nýtast vel við frekari gerð tungutæknilausna. Talgreinirinn frá Scansoft er svokallaður stakorðagreinir og einskorðaðist notkunin við það. Talgreinirinn virkaði vel í flestum tilfellum og sá íslenski betur en sá enski, sem ekki er óeðlilegt þar sem notendur voru allir íslenskir auk þess sem íslenskan er einsleitara tungumál en enska. Það er samt vísbending um að vel hafi tekist til með gerð íslenska talgreinisins og að hann eigi eftir að nýtast vel við gerð mismunandi kerfa.

Eftirfarandi eru nokkur atriði sem rétt er að hafa í huga við gerð slíkra lausna sem snúa að viðmóti við notendur, sem er ákaflega mikilvægur þáttur við gerð tungutæknilausna:

Skilgreining orðasafna. Mikilvægt er að skilgreina hvaða orð eigi að skiljast á mis-

munandi stöðum í svartrénu. Eðlilegt er að hafa jafnvel mörg leyfileg orð fyrir sömu skipun.

Ágiskunarstilling. Hægt er að stilla svokallaðan ágiskunarstuðul sem segir til um hvort reynt sé að giska á um hvaða orð er að ræða ef talgreinirinn er ekki viss um hvað sagt var. Það fer eftir eðli lausna hvernig ber að stilla þennan stuðul.

Staðfestingartextar. Í sumum tilfellum er nauðsynlegt að spyrja notandann hvort hann hafi sagt ákveðið orð til staðfestingar, þetta á ekki síst við þegar um kaup á vöru eða þjónustu er að ræða. Slíkt ferli hægir á vinnslu og því mikilvægt að nota rétt.

Staðfestingartónar. Mikilvægt er að viðmótið virki þægilega á notandann, t.d. að ekki komi langar þagnir meðan kerfið er að sækja gögn, svo dæmi sé tekið. Með því að nota staðfestingartóna og önnur hljóðtákn, t.d. þegar verið er að vinna eða sækja gögn, má gera flæðið mun eðlilegra.

Hjálp á öllum stigum. Æskilegt er að notendur geti fengið aðstoð hvar sem er í kerfinu, einfaldlega með því að segja „hjálp". Helst þarf hjálpin að vera þannig uppbyggð að hún byggist á því hvar viðkomandi er staddur í valtrénu.

Geta hætt hvar sem er. Á sama hátt er æskilegt að notendur geti hvar sem er sagt „hætta" til að hætta vinnslu eða fara efst í valtréð.

Þótt notkun talgreina gefi marga nýja möguleika fram yfir hefðbundin svarkerfi eru samt ýmsir vankantar við notkun þessarar tækni. Þar má t.d. nefna eftirfarandi atriði.

Seinlegt að segja talnarunur. Með talgreini þarf vanalega að segja eina tölu í einu. Í flestum tilfellum er fljótlegra að slá inn tölur á takkaborði símtækis.

Staðfestingar. Staðfestingar á svörum eru þreytandi en samt oft nauðsynlegar, t.d. þegar tölur eða upphæðir eru lesnar inn. Innsláttur frá takkaborði Símans er öruggari.

Hægvirkari vinnsla. Í mörgum tilfellum er fljótlegra að nota hefðbundin svarkerfi en með talgreini. Þetta á ekki síst við þegar framkvæma þarf sömu aðgerðina oft.

Skilningur. Framleiðendur talgreina lofa að þeir skilji það sem sagt er við þá í 97% tilfella. Dregið er í efa að slíkur skilningur sé fyrir hendi, a.m.k. kemur í ljós að ef um bakgrunnshávaða er að ræða minnkar þessi skilningur. Á sama hátt virðist sem skilningur úr GSM-símum sé lakari en úr borðsímum. Hefðbundin svarkerfi, þar sem tölur eru slegnar inn á takkaborði, hafa 100% skilning.

Notkunarmöguleikar þessara tungutæknilausna eru miklir og má telja upp fjölda möguleika í því sambandi. Á einfaldan hátt má þó skipta þessum þjónustum í nokkra flokka:

- Viðbót við núverandi svarkerfislausnir (upplýsinga- og fréttaþjónustur)

- Þjónustur sem spara mannskap (símaskrá, þjónustuver)

- Tekjuskapandi lausnir (kaup og sala á ýmsum vörum og þjónustu)

Lokaorð

Í upphafi var gerð ákveðin verk- og tímaáætlun sem lögð var fram með umsókn. Verkefnið stóð yfir í u.þ.b. eitt og hálft ár með hléum. Þetta er ákveðin seinkun sem var vísvitandi gerð þar sem beðið var eftir að íslenskur talgreinir yrði tilbúinn til prófana. Jafnframt var haldið utan um þá vinnu sem fór í verkefnið en hún fór allnokkuð fram úr áætlun, aðallega þar sem forsendur breyttust á verktímanum.

Að verkinu komu eftirfarandi starfsmenn Grunns:

Ívar Ragnarsson
Orri Eiríksson
Björn Jónsson

Það er mat okkar að þetta verkefni hafi tekist vel og muni nýtast okkur og vonandi öðrum sem stefna að gerð símtölvunarlausna sem nýta eiga sér tungutæknieiningar. Tungutækni í símtölvunarlausnum er mjög öflug viðbót en ýmsa pytti ber að varast til að árangurinn verði góður.

Helga Waage

Hjal – gerð íslensks stakorðagreinis

Helga Waage

Hjal – gerð íslensks stakorðagreinis

Hjal-verkefnið nefnist verkefni sem sneri að gerð talgreinis fyrir íslensku. Tilurð verkefnisins var sú að menntamálaráðuneytið ákvað að efna til tungutækniátaks til að styrkja íslenskustuðning í ýmsum tölvukerfum. Um svipað leyti var stofnuð tungutækniskor við Háskóla Íslands, þverfaglegt nám er miðast að því að mennta fólk til rannsókna og starfa á mörkum málvísinda og upplýsingatækni. Nokkur fyrirtæki á hugbúnaðar- og fjarskiptasviðum töldu að skilningur á íslensku mæltu máli væri mikilvægur, bæði fyrir þá þjónustu sem þau vildu geta boðið upp á en einnig væri mikilvægt fyrir íslenska tungu að hún væri jafnsjálfsagt viðmót við tæki og önnur tungumál. Upp úr þessum farvegi varð til samstarfshópur um gerð íslensks stakorðagreinis – Hjalhópurinn.

Að Hjali stóðu Háskóli Íslands, Hex hugbúnaður, Síminn, Nýherji og Grunnur-Gagnalausnir (nú Trackwell). Ákveðið var að leita samstarfs við öflugan framleiðanda á sviði talgreina og varð þýska fyrirtækið Philips (nú í eigu Scansoft) fyrir valinu. Það val helgaðist fyrst og fremst af því að tækni þess var þróuð með það í huga að styðja við mörg tungumál og því var gerð tungumálapakkans vel skilgreind eining innan kerfisins. Íslenska var 48. tungumál talgreinisins og nutum við góðs af yfirgripsmikilli þekkingu Scansoft-manna af gerð talgreina fyrir önnur tungumál.

Hvað er stakorðagreinir?

Stakorðagreinir er talgreinir sem skilur öll orð í einhverju tungumáli – bara ekki öll í einu. Stakorðagreinum er ætlað að skilja alla málhafa en einungis takmarkað orðamengi hverju sinni. Hversu takmarkað orðamengið er fer eftir ýmsu, til dæmis hversu lík orðin eru, hversu einsleitur framburður málhafa er, hversu vel þjálfað orðalíkanið er og hversu öflug tölvan er sem talgreinirinn keyrir á. Stærð orðamengisins getur þannig verið frá örfáum orðum upp í um 2 milljónir. Algeng stærð á orðamengi fyrir þjónustu er af stærðargráðunni 100–1000 orð.

Einnig eru til talgreinar sem skilja samfellt mál – dictational-kerfi. Það eru kerfi sem eru þjálfuð til að skilja samfellt tal hjá einum eða mjög fáum einstaklingum, eða þá að skilja mjög takmarkaðan orðaforða. Einfalt

dæmi um slíkan talgreini er talgreinirinn sem fylgir með Windows-stýrikerfinu og áhugasamir geta dundað sér við að þjálfa upp til að taka niður skjöl sem lesin eru fyrir. Einnig eru til svona talgreinar fyrir mjög takmörkuð orðamengi, til dæmis fyrir skurðlækna. Enn er ekki til talgreinir af þessu tagi fyrir íslensku.

En aftur að stakorðagreininum. Stakorðagreinir vinnur yfirleitt sem viðmót á annað kerfi, til dæmis upplýsingasíma um færð á þjóðvegum. Hann er gjarna notaður sem viðmót í gegnum símkerfi þar sem hann hlustar á viðmælanda sinn og greinir út úr hljóðaflaumnum orð og orð á stangli sem hann sendir síðan til samstarfskerfisins ásamt upplýsingum um það hversu öruggur hann sé um orðagreininguna. Ef um er að ræða upplýsingar um færð á íslenskum hálendisvegum væru orðin sem greinirinn hlustar eftir heiti á borð við *Steingrímsfjarðarheiði, Hellisheiði og Holtavörðuheiði* sem öll eru ólík og greinirinn greinir með miklu öryggi á milli þeirra. Upplýsingaveitur á borð við 118 þurfa hins vegar að greina á milli þúsunda nafna sem hafa þann ágalla að vera ekki einkvæm (margir heita *Guðmundur Jónsson*), nöfn sem hljóma svipað (eins og *Ebba Dóra og Edda Þóra*) svo ekki sé talað um þá sem hringja í 118 ef þá vantar uppskrift að sósu eða eru að leita að góðri hársnyrtistofu. Það er því mun auðveldara að búa til þjónustu sem miðlar sjálfvirkt upplýsingum um færð á hálendisvegum til símnotenda heldur en þjónustu sem veitir upplýsingar um símanúmer þótt bæði kerfin noti sömu grunntæknina.

Framkvæmd Hjalverkefnisins

Sótt var um styrk til verkefnisins í árslok 2002 og er ljóst að ekki hefði orðið af verkefninu ef það hefði ekki verið styrkt á þann máta sem gert var. Undirbúningur verkefnisins hófst síðan í ársbyrjun 2003. Verkefnisstjóri var Sæmundur Þorsteinsson hjá Landssíma Íslands en rekstur verkefnisins var í höndum Helgu Waage hjá Hex. Scansoft lagði línurnar með hvaða verkþætti þyrfti að vinna og hvernig væri best að manna þá. Var ákveðið að ráða þrjá mastersnema í tungutækni til verksins og skyldu þeir vinna meginþorra vinnunnar sumarið 2003. Það þurfti því að tryggja að allur undirbúningur væri með þeim hætti að sá tími sem nemarnir hefðu til umráða myndi nýtast sem best.

Verkáætlun verkefnisins var í stórum drátt-um þessi:

- Málfræðileg forvinna – safna saman þeim gögnum sem þarf til að gera talgreininn
- Safna talsýnum – fá um það bil 2000 Íslendinga til að taka þátt í söfnun á taldæmum
- Skrá talsýni – hlusta á talsýni og skrá hvað hver einstaklingur segir og hvernig það er sagt
- Þjálfa talgreini – gögn send til Þýskaland til að þjálfa tungumálaeininguna

Málfræðileg forvinna

Forsenda þess að vel tækist til með talgreininn var sú að málfræðileg forvinna væri vel úr garði gerð – að öll málhljóð og málhljóðasambönd í íslensku væru þekkt og vitað hvar og hvernig þau kæmu fyrir. Einnig þurfti að útvega lista af staðarheitum, mannanöfnum, fyrirtækjaheitum, algengum fyrirskipunum og töluorðum. Að lokum þurftum við að útbúa eðlilegar setningar samkvæmt niðurstöðum málhljóðagreiningarinnar. Úr þessum gögnum voru útbúin blöð sem send yrðu til þátttakenda í talsýnatöku og fylgir eitt slíkt blað með þessum pistli. Orðalisti með rúmlega 30.000 algengum orðum var undirbúinn til hljóðritunar. Hljóðritaði listinn yrði síðan hafður til hliðsjónar við þjálfun talgreinisins. Að auki voru önnur orð af setningablöðunum sett á listann.

Eiríkur Rögnvaldsson, prófessor við Háskóla Íslands, tók saman megnið af þessum gögnum, sér í lagi var mikilvægt að fá yfirlit yfir öll málhljóð í íslensku.

Við undirbúning gagna kom í ljós að leiðbeiningar Scansoft áttu ekki fyllilega við íslensku. Til dæmis eru töluorð bæði fleiri og flóknari í íslensku en í flestum öðrum tungumálum en mállýskur fáar og ekki mikill munur á þeim frá sjónarhóli þarfa talgreinisins.

Afurðir þessa verkþáttar var Sampa – hljóðritunarstaðall fyrir íslensku og 1000 mismunandi bréf til að nota við talsýnatöku.

Söfnun upptakna

Markmið verkefnisins var að safna 1500-2000 upptökum af fólki 14 ára og eldra af öllu landinu. Sérstaklega var lögð áhersla á að ná málhöfum með norðlenskan framburð svo og konum eldri en 40 ára, sem okkur var sagt að skiluðu sér almennt fremur illa í verkefni af þessu tagi. Söfnunin var tvíþætt – í upphafi var leitað til fjölmiðla um að kynna verkefnið til að fá inn þátttakendur sem hefðu áhuga á verkefninu sem slíku og vildu leggja því lið. Síðan var leitað til Gallup og beðið um aðstoð við að safna því úrtaki sem upp á vantaði. Þátttakandinn fékk sent til sín bréf til að lesa og síðan hringdi hann inn og fylgdi fyrirmælum á blaðinu.

Safnað var talsýnum 2005 einstaklinga, en ríflega 3000 manns skráðu sig til þátttöku. Í 89% tilvika kláraði þátttakandinn símtalið. Yngsti þátttakandinn var 8 ára stúlka (sem því miður varð að sía frá, þar sem hún er of ung til að gagnast þjálfun talgreinisins) en sá elsti var 83 ára karl. Alls hringdu 14 einstaklingar 70 ára og eldri. Flestir sem hringdu voru á aldrinum 18 til 40 ára og voru konur heldur duglegri að hringja inn, eða 55%.

Þátttakendur voru um 1% þjóðarinnar (14 ára og eldri) og þykir það einstakt að svo stór hluti þjóðar hafi tekið þátt í verkefni af þessu tagi.

Úrvinnsla

Úrvinnsla gagna var tvíþætt. Annars vegar hljóðritun orðalista en hins vegar skráning á tali þess sem hringdi inn.

Hljóðrita þurfti öll orð sem komu fyrir í talsýnunum og auk þess um 30.000 algengar orðmyndir. Við gerð þessa orðalista (sem endaði í rúmlega 50.000 orðum) var þá valin sú mynd eða þær myndir sem taldar voru algengastar.

Við skráningu á talsýnum var hlustað á hverja einustu upptöku og skráð nákvæmlega niður hvað viðmælandinn sagði, hljóð sem heyrðust í bakgrunni, hóstar, hik og önnur málhljóð. Einnig voru mállýskur, mismæli og ýmis afbrigði skráð sérstaklega.

Upptökurnar og orðalistarnir voru síðan sendir til Scansoft sem fór yfir þá og notaði síðan gögnin til að þjálfa mállíkan talgreinisins.

Niðurstaða

Fullbúnum talgreini var skilað í byrjun nóvember 2003. Talgreinirinn virkar vel, enda kom í ljós að íslenska er vel fallin til að greina á þennan máta. Hrynjandin í málinu er nokkuð regluleg, það er fremur langt á milli hljóða og framburður er nokkuð einsleitur.

Talgreinirinn hefur nú verið í notkun í rúmt ár og er fyllilega samanburðarhæfur við stakorðagreina fyrir önnur tungumál. Hann hefur verið notaður til að veita ýmiss konar þjónustu í gegnum síma, bæði viðskiptalegs eðlis, til að veita almannaþjónustu og til skemmtunar.

Aðrar afurðir verkefnisins eru hljóðritaður orðalisti sem ekki var áður til fyrir íslensku, upptökur af talmáli 2000 kyn-, aldurs- og búsetugreindra einstaklinga og hljóðritunarstaðall fyrir íslensku.

Dæmi um innhringiblað

Velkomin(n) í hljóðsöfnun Hjals. Hljóðsöfnunin fer þannig fram að fyrst verður þú beðin(n) að segja auðkennistöluna þína. Hana er að finna í bréfinu sem þú fékkst sent. Síðan koma fáeinar spurningar sem við viljum biðja þig um að svara. Að því loknu verður þú beðin(n) að segja setningar og setningabrot úr bréfinu. Kerfið mun leiðbeina þér. Hafðu engar áhyggjur þótt eitthvað komi upp á í upptöku. Haltu áfram eins og ekkert hafi í skorist.

Það væri gott ef þú læsir bréfið einu sinni alveg í gegn áður en við byrjum hljóðsöfnunina.

Í hverri umferð gefur kerfið fyrirmæli og síðan heyrist hljóðmerki. Þú segir svarið eftir að hljóðmerkið heyrist.

Segðu auðkennistölu þína.

Hver er fæðingardagur þinn og ár?
Ertu karl eða kona?

Hvaðan ertu?

Segðu einhverja tölu milli 0 og 10 milljóna.

Hringir þú úr farsíma?

Segðu eftirfarandi staðanöfn:

1. Ólafsvík
2. Víðinesi
3. Héðinsgötu
4. Mjóafirði
5. Hegranesi
6. Holtagerði

Segðu eftirfarandi mannanöfn:

1. Pál Böðvarsson
2. Sólveig Logadóttir
3. Pétur Hreggviðsson
4. Guðrún Sigríður Erlendsdóttir
5. Anna María Ingimundardóttir

Segðu eftirfarandi nöfn stofnana og fyrirtækja:

1. Orkustofnun
2. Forsætisráðuneytið
3. Eimskipafélag Íslands
4. Mjólkursamsalan
5. Flugstöð Leifs Eiríkssonar

Segðu eftirfarandi tölur:

1. annan annar annarri annars ein eina
2. einar fern fernra fjórar fjórði fjórir
3. fjögur fyrsta fyrsti fyrstu níu níundi
4. sjötti tvenn tvennar tvennir tvo tvö
5. þrem þremur þriðji þrjár þrjú öðrum

Segðu eftirfarandi skipanir:

1. ég vildi
2. starta
3. taka frá miða

Segðu eftirfarandi peningaupphæðir:

1. 281 kr.
2. 88.123 kr.
3. 290.284 kr.
4. 1.800.000 kr.
5. 4.000.000 kr.

Segðu eftirfarandi setningar:

1. Jamm, hér er víst hægt að hnjóta um hnullung, sagði hann.
2. Ég gleymdi mér andartak og gerðist montinn.
3. Raunsæismanneskjan hafði rétt fyrir sér.
4. Við Stefán talaði hann lágt og blíðlega og Stefán hváði.
5. Þau horfðu á mig vantrúuð en höfðu hvorki önnur ráð né betri.

Segðu eftirfarandi tímasetningar:

1. Miðvikudagur 5. febrúar
2. Á eftir

Segðu eftirfarandi setningabrot:

1. Mætti ég
2. Svo er nú það

Stafaðu eftirfarandi stafarunur:

1. B A K A R Í I Ð
2. H V E R G I
3. H L M X M Ú X

Kærar þakkir fyrir þátttökuna!

Sigrún Helgadóttir

Markari fyrir íslenskan texta

Sigrún Helgadóttir

Markari fyrir íslenskan texta

Inngangur

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998-1999 lagði m.a. til að „unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði".

Það að greina orð eftir orðflokkum og beygingu er kallað *tagging* í ensku og greiningarstrengurinn kallast *tag*. Forrit eða kerfi sem framkvæmir þetta verk kallast á ensku *tagger*. Lagt er til að greiningarstrengurinn kallist **mark** á íslensku, aðgerðin verði köllluð **mörkun** og forritið eða kerfið kallist **markari**.

Í anda tillögunnar var gerð málfræðilegs markara fyrir íslensku eitt af þeim verkefnum sem voru styrkt af tungutækniverkefni menntamálaráðuneytisins í apríl 2002. Markmið verkefnisins var að finna aðferðir til þess að greina íslenskan texta vélrænt í orðflokka og eftir beygingu. Afurð verkefnisins átti að vera annaðhvort reglusafn til að nota með svokölluðum Brill-markara eða sérstakt forrit. Verkefnið þróaðist þó á þann veg að prófaðar voru fjórar aðferðir við mörkun íslensks texta. Einnig var prófað að setja þrjár af þessum aðferðum saman eftir tilteknum reglum til þess að ná sem bestum árangri við mörkun.

Í ýmsum tungutækniverkefnum þar sem unnið er úr texta er ávinningur að því að orð í textanum séu greind í orðflokka og beygingarmyndir. Má þar nefna greiningu texta í setningahluta (e. *partial parsing*), orðtöku úr texta fyrir gerð orðasafns, upplýsingaheimt, talkennsl, talgervingu, vélrænar þýðingar, orðabókargerð, fyrirspurnarkerfi og leiðréttingarforrit. Einnig er nauðsynlegt að orð í texta séu greind eftir orðflokkum og beygingu ef gera á tíðnikönnun á texta eins og þá sem birt er í Íslenskri orðtíðnibók (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991).

Handvirk greining texta eftir orðflokkum og beygingu er mjög tímafrek og heldur leiðinleg iðja. Þess vegna hefur lengi verið fengist við að þróa vélrænar aðferðir við það verkefni. Þetta svið hefur því fengið mikla umfjöllun hjá þeim sem vinna við máltækni.

Vélrænar aðferðir við mörkun eru venjulega flokkaðar í tvo flokka, regluaðferðir og tölfræðilegar aðferðir. Fyrstu vélrænu aðferðirnar sem var beitt voru regluaðferðir. Orðasafn var notað til að merkja sérhvert orð í texta með öllum hugsanlegum greiningarstrengjum. Síðan voru notaðar reglur til þess að skera úr um hvaða greiningarstrengur væri réttur. Þessar reglur voru byggðar á málfræði hvers tungumáls og venjulega samdar af málfræðingum.

Tölfræðilegar aðferðir byggjast allar á því að orðum í textasafni hefur verið úthlutað mörkum og þau leiðrétt handvirkt (*data-driven methods*). Forrit er síðan látið búa til líkan á grundvelli þessara gagna sem þegar hafa verið greind. Aðrar aðferðir sem ekki eru beinlínis flokkaðar sem tölfræðilegar byggjast einnig á sama vinnulagi. Má þar nefna aðferð sem mætti kalla leiðréttingaaðferð og byggist á því að skipta um greiningarstreng þegar ákveðnum skilyrðum í umhverfi orðsins er fullnægt (e. *transformation-based learning*). Forrit eða kerfi sem nota fyrir fram greint textasafn til þess að læra af mætti kalla námfúsa markara.

Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni.

Efniviður og aðferðir

Prófaðar voru nokkrar aðferðir við mörkun sem allar eiga það sameiginlegt að reynt er að læra af fyrir fram greindum gögnum. Námfús markari lærir fyrst af textasafni sem hefur verið greint í orðflokka og eftir beygingu. Markarinn nýtir síðan þessa kunnáttu til þess að marka orð í texta sem hann hefur ekki lesið áður. Til þess að ná sem bestum árangri er æskilegt að textinn sem á að marka sé sem líkastur textanum sem markarinn lærði af.

Til þess að prófa námfúsan markara á nýju tungumáli eða nýrri gerð af texta er nauðsynlegt að hafa aðgang að stóru textasafni þar sem hvert orð hefur verið greint eftir orðflokki og beygingu. Textasafninu er venjulega skipt í tvo hluta. Annar hlutinn, sem gæti verið um 90% af safninu, er notaður til þess að þjálfa markarann og kallast þjálfunarsafn. Hinn hlutinn (10% af safninu) er notaður til þess að prófa það sem markarinn hefur lært og kallast prófunarsafn. Orðum er úthlutað marki og útkoman síðan borin saman við rétt mark.

Í þeirri vinnu sem hér er greint frá var notað textasafn sem varð til við undirbúning *Íslenskrar orðtíðnibókar*. Í textasafninu eru 590.297 lesmálsorð sem birtast í 59.358 mismunandi orðmyndum að meðtöldum greinarmerkjum. Lesmálsorðunum fylgja 639 mismunandi greiningarstrengir að meðtöldum greinarmerkjum.

Ákveðið var að prófa fjórar aðferðir við mörkun. Tvær þeirra teljast til tölfræðilegra aðferða, eina mætti kalla leiðréttingaaðferð (e. *error-driven transformation-based learning*) og ein byggist á minnistækni (e. *memory-based technique*). Alls voru próf-

aðir fimm markarar sem unnt er að þjálfa á íslenskum texta og eru fáanlegir endurgjaldslaust. Tölfræðimarkararnir sem voru prófaðir voru **TnT**, sem byggist á Markovslíkani, og **MXPOST** sem byggist á svo kölluðu hámarksóreiðulíkani (e. *Maximum Entropy Model*). Tveir markarar, **µ-TBL** og **fnTBL**, sem byggjast á leiðréttingaaðferðinni voru prófaðir og einn markari, **MBT**, sem byggist á minnistækni.

Prófanir

Tölvuskrár Orðtíðnibókarinnar eru skipulagðar þannig að í hverri skrá er textabútur úr einni heimild. Hverri skrá var skipt í 10 nokkurn veginn jafna búta. Úr þessum 10 bútum voru búin til 10 pör af skrám þannig að skrárnar í hverju pari skarast ekki. Í hverju pari er ein skrá með um 90% af lesmálsorðum úr textasafninu og önnur með um 10% af lesmálsorðum úr textasafninu. Stærri skráin er notuð sem þjálfunarsafn og sú minni sem prófunarsafn. Í hverju pari eru því textar sem eiga að vera dæmigerðir fyrir alla textaflokka í textasafninu. Prófunarsöfnin 10 eru óháð hvert öðru en þjálfunarsöfnin hafa um 80% sameiginlega texta. Allir markarar voru prófaðir á öllum 10 pörum og fundin meðalnákvæmni (*ten-fold cross-validation*).

Tveir markaranna, µ-TBL og MBT, virtust ekki henta fyrir íslenska textasafnið en markararnir TnT, MXPOST og fnTBL voru prófaðir á öllum 10 pörunum. Eins og sést af töflu 1 gaf TnT-markarinn besta niðurstöðu, þá MXPOST-markarinn og sístur var fnTBL-markarinn.

Nákvæmni er sýnd fyrir öll orð, þekkt orð og óþekkt orð. Óþekkt orð eru orð sem eru í prófunarsafni en ekki viðkomandi þjálfunarsafni. Tölur í töflu 1 eru fengnar með því að leggja saman prófunarsöfnin 10 og telja rétt greind orð fyrir hverja mörkunaraðferð. Tölur um nákvæmni gefa því meðalnákvæmni fyrir pörin 10. Meðalhlutfall óþekktra orða í prófunarsöfnunum var 6,84%.

Eins og sést á töflunni eru markararnir þrír misjafnlega duglegir við að greina óþekkt orð, þ.e. orð sem þeir hafa ekki séð áður. Markararnir nota mismunandi aðferðir við greiningu óþekktra orða. TnT-markarinn virðist hafa yfir að ráða betri aðferð en hinir markararnir við að greina óþekkt orð og fær því besta heildarniðurstöðu eða **90,36**%.

Vert er að benda á að mark er talið rangt þó að aðeins eitt af sex atriðum í greiningarstreng sé rangt.

Niðurstöður mörkunar voru skoðaðar nákvæmlega og greindar til þess að finna hvers konar villur markararnir gera og hvernig mætti bæta árangurinn. Algengustu villurnar sem allir markarar gera er að rugla saman fallstjórn forsetninga. Næst kemur ruglingur á milli beygingarmynda nafnorða sem hafa sömu mynd. Má þar nefna þolfall og þágufall kvenkynsorða í eintölu (`þf`. *konu;* `þgf`. *konu*) og nefnifall og þolfall hvorugkynsorða í eintölu (`nf`.

barn; þf. *barn*). Ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum er líka algengur þar sem þessar beygingarmyndir líta eins út (*ég fer; hann fer*). Einnig má nefna nafnhátt og þriðju persónu fleirtölu í nútíð en þessar beygingarmyndir líta eins út (*að fara; þeir fara*).

Markararnir gera að nokkru leyti ólíkar villur og það má nota á ýmsa vegu til þess að bæta árangur mörkunar. Prófað var að kjósa á milli marka sem markarar úthluta. Í íslenska verkefninu voru prófaðir þrír markarar. Sú aðferð við kosningu á milli þeirra sem gaf besta niðurstöðu fólst í því að velja það mark sem tveir eða fleiri voru sammála um. Ef allir þrír eru ósammála var valið mark þess markara sem stóð sig best, í þessu tilviki mark TnT. Í töflu 1 sést að með því að kjósa á milli markara á þennan hátt fæst **91,54**% nákvæmni.

Greining textans í textasafni orðtíðnibókarinnar er mun ítarlegri en tíðkast í tungumálum sem þessi kerfi höfðu verið prófuð á. Skrá yfir alla greiningarstrengi eða mörk sem koma fyrir í tilteknu mörkuðu textasafni er oft kölluð **markaskrá** (e. *tagset*). Markaskrá orðtíðnibókarinnar er mjög stór og ítarleg. Sú greining sem þar er notuð er ekki endilega sú eina rétta og verið getur að sumar tungutæknilausnir geti nýtt sér greiningu sem er ekki jafn ítarleg. Sum tungutækniverkefni gætu þurft mikla nákvæmni í mörkun en ekki mjög ítarlega greiningu. Í viðauka sést hvernig greiningarstrengir Orðtíðnibókarinnar eru settir saman.

Tafla 1. Nákvæmni þriggja markara og nákvæmni sem fæst með því að kjósa á milli niðurstöðu markaranna og nákvæmni þegar greiningarstrengir eru einfaldaðir. Að lokum er beitt reglum.
Einnig niðurstöður miðað við að nota orðasafn.
Einnig niðurstöður miðað við að aðeins orðflokkar séu greindir

Markaskrá	Orða-safn[1]	Aðferð	Óþekkt orð		Þekkt orð		Öll orð	
			Tíðni	%	Tíðni	%	Tíðni	%
Alls			40.392	6,84	549.905	93,16	590.297	100,00
Óbreytt	Nei	MXPOST	25.246	62,50	500.617	91,04	525.863	89,08
Óbreytt	Nei	fnTBL	21.823	54,03	502.378	91,36	524.201	88,80
Óbreytt	Nei	TnT	28.919	71,60	504.484	91,74	533.403	**90,36**
Óbreytt		Kosið milli m	29.003	71,80	511.348	92,99	540.351	**91,54**
Einföldud[2]		MXPOST	25.255	62,52	508.753	92,52	534.008	90,46
Einföldud[2]		fnTBL	21.831	54,05	509.309	92,62	531.140	89,98
Einföldud[2]		TnT	28.925	71,61	513.173	93,32	542.098	**91,83**
Einf. f. kosn.		Kosið milli m	29.010	71,82	517.342	94,08	546.352	**92,56**
		Mark MXPOS	29.141	72,15	518.775	94,34	547.916	**92,82**
Óbreytt	Nei	MXPOST	25.252	62,50	500.611	91,04	525.863	89,08
Óbreytt	Já	fnTBL	28.461	70,44	503.142	91,50	531.603	90,06
Óbreytt	Já	TnT	34.859	86,28	505.511	91,93	540.370	**91,54**
Óbreytt		Kosið milli m	34.331	84,97	512.044	93,12	546.375	92,56
Einföldud[2]		MXPOST	25.261	62,52	508.747	92,52	534.008	90,46
Einföldud[2]		fnTBL	28.467	70,46	509.788	92,71	538.255	91,18
Einföldud[2]		TnT	34.863	86,29	513.797	93,44	548.660	**92,95**
Einf. f. kosn		Kosið milli m	34.336	84,98	517.773	94,16	552.109	**93,53**
		Mark MXPOS	34.013	84,18	518.818	94,35	552.831	**93,65**
Orðfl greindir		MXPOST	35.621	88,19	538.697	97,96	574.318	97,29
		fnTBL	33.181	82,15	540.859	98,35	574.040	97,25
		TnT	37.349	92,47	541.946	98,55	579.295	98,14

[1] Orðasafn hefur u.þ.b. helming óþekktra orða
[2] Einföldun felst í að greina ekki atviksorð og ekki heldur samtengingar
Fornöfn eru sett í einn flokk en að öðru leyti er greining þeirra eftir kyni, tölu og falli látin haldast.

Prófað var að einfalda greiningarstrengi á þrennan hátt. Einföldunin felst í því að líta aðeins á fyrsta staf í greiningarstreng fyrir atviksorð og samtengingar, þ.e. greina þessa orðflokka ekki í undirflokka, og slá saman fornafnaflokkum en láta greiningu fornafna halda sér að öðru leyti. Í töflu 1 sést að TnT-markarinn nær 91,83% nákvæmni eftir að markaskrá hefur verið einfölduð.

Eins og þegar er getið skiptir miklu máli að hafa góðar aðferðir til þess að greina óþekkt orð. Markararnir búa til orðasafn úr þjálfunarsafninu og nota það við mörkun texta sem þeir hafa ekki séð. Tveir af mörkurunum, TnT og fnTBL, gefa kost á að nota viðbótarorðasafn við mörkunina. Notað var orðasafn sem hefur um helming þeirra orða sem eru óþekkt í hverju prófunarsafni miðað við samsvarandi þjálfunarsafn og bætti það árangur nokkuð. Í töflu 1 sést að TnT-markarinn nær 91,54% nákvæmni með því orðasafni.

Flestir markararnir eiga í erfiðleikum með að greina á milli orðmynda sem líta eins út. Má þar nefna þolfall og þágufall eintölu kvenkynsorða og nefnifall og þolfall eintölu hvorugkynsorða. MXPOST-markarinn virtist gera færri slíkar villur en hinir markararnir tveir. Samdar voru reglur til þess að velja mark MXPOST-markarans frekar en niðurstöðu kosningar ef tilteknum skilyrðum var fullnægt. Með því móti mátti bæta niðurstöðu nokkuð. Tafla 1 sýnir helstu niðurstöður og að besta niðurstaðan fyrir þann efnivið sem var notaður varð **93,65**% nákvæmni.

Neðst í töflunni sést nákvæmni ef aðeins er litið á greiningu eftir orðflokkum. Þá nær TnT 98,14% nákvæmni, MXPOST 97,27% nákvæmni og fnTBL 97,25% nákvæmni. Í sumum tungutækniverkefnum nægir greining eftir orðflokkum.

Aðferðirnar prófaðar á nýjum textum

Aðferðirnar við mörkun sem hér hefur verið lýst voru prófaðar á textum sem ekki voru hluti af textasafni Orðtíðnibókarinnar. Fjögur aðskilin lítil textasöfn voru notuð. Í fyrsta safninu eru brot úr 13 skáldritum frá 19. öld og fyrri hluta 20. aldar, samtals 6.022 lesmálsorð að meðtöldum greinarmerkjum. Í öðru safninu eru brot úr níu skáldverkum frá því eftir 1980, samtals 3.601 lesmálsorð að meðtöldum greinarmerkjum. Í þriðja safninu eru textar um tölvur og tækni sem eru fengnir úr gagnasafni Morgunblaðsins, úr Fréttabréfi RHÍ og af vefsíðum ýmissa tölvufyrirtækja, samtals 2.926 lesmálorð að meðtöldum greinarmerkjum. Í fjórða safninu eru textar um lögfræði og viðskipti sem eru teknir úr Lagasafni, fréttabréfi fjármálaráðuneytis og Morgunblaðinu (viðskipti), alls 2.776 lesmálsorð að meðtöldum greinarmerkjum. Mörkun var síðan leiðrétt til þess að unnt væri að reikna út nákvæmni mörkunar með hinum ýmsu aðferðum.

Tafla 2. Nákvæmni við mörkun texta sem eru ekki í textasafni Orðtíðnibókar

Gamall bókmenntatexti						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	524	8,70	5.498	91,30	6.022	100,00
MXP	334	63,74	4.935	89,76	5.269	87,50
fnTBL	279	53,24	4.985	90,67	5.264	87,41
TnT	393	75,00	5.209	94,74	5.602	93,03
TnT, einf.	393	75,00	5.218	94,91	5.611	93,18
MXP, orðfl.	458	87,40	5.326	96,87	5.784	96,05
fnTBL, orðfl	409	78,05	5.374	97,74	5.783	96,03
TnT, orðfl.	472	90,08	5.430	98,76	5.902	98,01
Bókmenntatextar frá því eftir 1980						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	280	7,21	3.601	92,79	3.881	100,00
MXP	182	0,00	3.217	89,34	3.399	87,58
fnTBL	157	56,07	3.262	90,59	3.419	88,10
TnT	221	78,93	3.385	94,00	3.606	92,91
TnT, einf.	221	0,00	3.385	94,00	3.606	92,91
MXP, orðfl.	236	84,29	3.512	97,53	3.748	96,57
fnTBL, orðfl	221	78,93	3.537	98,22	3.758	96,83
TnT, orðfl.	257	91,79	3.561	98,89	3.818	98,38
Textar um tölvur og tækni						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	442	15,11	2484	84,89	2.926	100,00
MXP	186	42,08	2191	88,20	2.377	81,24
fnTBL	169	38,24	2190	88,16	2.359	80,62
TnT	222	50,23	2317	93,28	2.539	86,77
TnT einf.	222	50,23	2317	93,28	2.539	86,77
MXP, orðfl	364	82,35	2410	97,02	2.774	94,81
fnTBL, orðfl	356	80,54	2437	98,11	2.793	95,45
TnT, orðfl.	395	89,37	2453	98,75	2.848	97,33
Textar um lögfræði og viðskipti						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	390	14,05	2.386	85,95	2.776	100,00
MXP	236	60,51	2.042	85,58	2.278	82,06
fnTBL	213	54,62	2.041	85,54	2.254	81,20
TnT	284	72,82	2.174	91,11	2.458	88,54
TnT einf.	284	72,82	2.176	91,20	2.460	88,62
MXP, orðfl.	348	89,23	2.301	96,44	2.649	95,43
fnTBL, orðfl	336	86,15	2.309	96,77	2.645	95,28
TnT, orðfl.	366	93,85	2.337	97,95	2.703	97,37

Í töflu 2 sjást helstu niðurstöður mörkunar lesmálsorða í þessum textum. Hér kemur í ljós að TnT-markarinn nær bestum árangri. Markararnir MXPOST og fnTBL ná svo lélegum árangri að ekki reyndist unnt að bæta niðurstöðu TnT-markarans með því að nýta niðurstöður frá hinum mörkurunum tveimur. TnT-markarinn nær betri árangri við mörkun bókmenntatextanna heldur en við mörkun texta orðtíðnibókarinnar sjálfrar en verri árangri við mörkun textanna um tölvur og tækni og viðskipti og lögfræði.

Ekki var notað viðbótarorðasafn þannig að óþekkt orð eru þau orð sem ekki koma fyrir í textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hátt í öllum textunum og hærra en meðalhlutfall í prófunarsöfnum sem gerð voru úr textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hæst í textanum um tölvur og tækni og þar er árangur mörkunar slakastur. TnT-markarinn nær samt alls staðar viðunandi árangri ef aðeins er gerð krafa um réttan orðflokk.

Þessar niðurstöður benda til þess að nauðsynlegt sé að bæta árangur mörkunar óþekktra orða til þess ná viðunandi árangri í mörkun texta. Ein leið til þess að gera það er að hafa til umráða umfangsmiklar orðaskrár þar sem fram koma beygingarmyndir sem flestra orða og mörk þeirra. Nota má beygingarlýsingu íslensks nútímamáls, sem einnig var gerð fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins, sem efnivið í slíka orðaskrá. Einnig er nauðsynlegt að hafa tiltækar skrár með ýmiss konar sérnöfnum svo sem mannanöfnum, nöfnum fyrirtækja og stofnana og örnefnum.

Umsjón með verkinu

Verktakar við verkið voru Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir) og Orðabók Háskólans. Gerður var samningur milli ráðuneytisins og verktakanna 18. október 2002. Verkið hófst haustið 2002 og lokaskýrslu var skilað til menntamálaráðuneytisins í febrúar 2004. Verkefnisstjóri var Eiríkur Rögnvaldsson en Sigrún Helgadóttir mótaði vinnulag við prófun markaranna og vann meginhluta vinnunnar ásamt félögum í Málgreiningarhópnum. Orðabók Háskólans lagði til verksins aðstöðu og markað textasafn *Íslenskrar orðtíðnibókar*.

Heimildir

Jörgen Pind (ritstj.), Friðrik Magnússon, Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

Viðauki

Skýring skammstafana í greiningarstrengjum Orðtíðnibókar

Dálkur	Formdeild	Greiningartákn-greiningaratriði
1	Orðflokkur	N-nafnorð
2	kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn, X-ókyngreint
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
5	Greinir	G-með viðskeyttum greini
6	Sérnöfn	M-mannsnafn, Ö-örnefni, S-önnur sérnöfn
1	Orðflokkur	L-lýsingarorð
2	Stig	F-frumstig, M-miðstig, E-efstastig
3	Beyging	S-sterk beyging, V-veik beyging, O-óbeygt
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	Tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	F-fornafn
2	Flokkur	A-ábendingarfornafn, B-óákveðið ábendingarfornafn, E-eignarfornafn O-óákveðið fornafn, P-persónufornafn, S-spurnarfornafn, T-tilvísunarfornafn
3	Kyn/Persóna	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	G-greinir
2	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	T-töluorð
2	Flokkur	F-frumtala
3	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	S-sögn (þó ekki lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	N-nafnh., B-boðh., F-framsöguh., V-viðtengingarh., S-sagnbót, L-lýsingarh. nútíðar
4	Tíð	N-nútíð, Þ-þátíð
5	Tala	E-eintala, F-fleirtala
6	Persóna	1-1. persóna, 2-2. persóna, 3-3. persóna
1	Orðflokkur	S-sögn (lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	Þ-lýsingarháttur þátíðar
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall
1	Orðflokkur	A-atviksorð
2	Stig	M-miðstig, E-efsta stig
3	Flokkur/Fallstjórn	A-stýrir ekki falli, U-upphrópun/ O-stýrir þolfalli, Þ-stýrir þágufalli E-stýrir eignarfalli
1	Orðflokkur	C-samtenging
2	Flokkur	N-nafnháttarmerki, T-tilvísunartenging
1	Flokkur	E-erlent orð
1		X-ógreint orð

Sigrún Helgadóttir

Mörkuð íslensk málheild

Sigrún Helgadóttir

Mörkuð íslensk málheild

Inngangur

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998-1999 benti á að til þess að unnt sé að útbúa forrit sem nýta og nota tungumál þurfi að vera fyrir hendi miklar og nákvæmar upplýsingar um tungumálið og notkun þess. Þar má m.a. nefna upplýsingar um tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerð og merkingu.

Þessar upplýsingar má fá úr textaheild eða málheild (e. *corpus*), þ.e. safni tölvutækra texta af ýmsu tagi svo sem blaðatexta, fræðitexta af ýmsum sviðum, bókmenntatexta og talmáls. Starfshópurinn lagði því m.a. til að komið yrði á fót slíkri málheild, sem gæti nýst „fyrirtækjum sem hráefni í afurðir".

Víða í grannlöndum okkar eru stórar málheildir þegar til eða verið er að koma á fót slíkum söfnum sem eru aðgengileg fyrir margvíslegar rannsóknir á málinu og til afnota fyrir þá sem búa til ýmiss konar tungutæknitól. Þar má nefna *British National Corpus* (BNC)[1] í Bretlandi, *Korpus 2000*[2] í Danmörku og *American National Corpus* (ANC)[3] í Bandaríkjunum. Það fer eftir aðstæðum í hverju landi hvernig notkun og aðgangi er háttað.

Breska málheildin BNC er stærst þessara málheilda og komin lengst á veg. Í henni eru 100 milljón orð af breskri ensku. BNC-málheildin var búin til á fyrri hluta 10. áratugar tuttugustu aldar. Að því stóðu útgefendur og háskólastofnanir og fékk verkefnið umtalsverða opinbera styrki. Í málheildinni eru fjölbreyttir textar í tilteknum hlutföllum auk talaðs máls sem er um 10% af safninu.

Hvað er mörkuð málheild?

Með **markaðri málheild** (e. *tagged corpus*) er átt við safn fjölbreyttra textabúta sem hafa verið greindir á málfræðilegan hátt. Málheildin er í rafrænu formi og venjulega geymd í stöðluðu sniði. Hverjum textabút fylgja upplýsingar um textann sem búturinn er úr og hverri orðmynd fylgir **nefnimynd** (*lemma*) og greiningarstrengur, sem kallast **mark** (e. *tag*) og sýnir orðflokk og beyging-

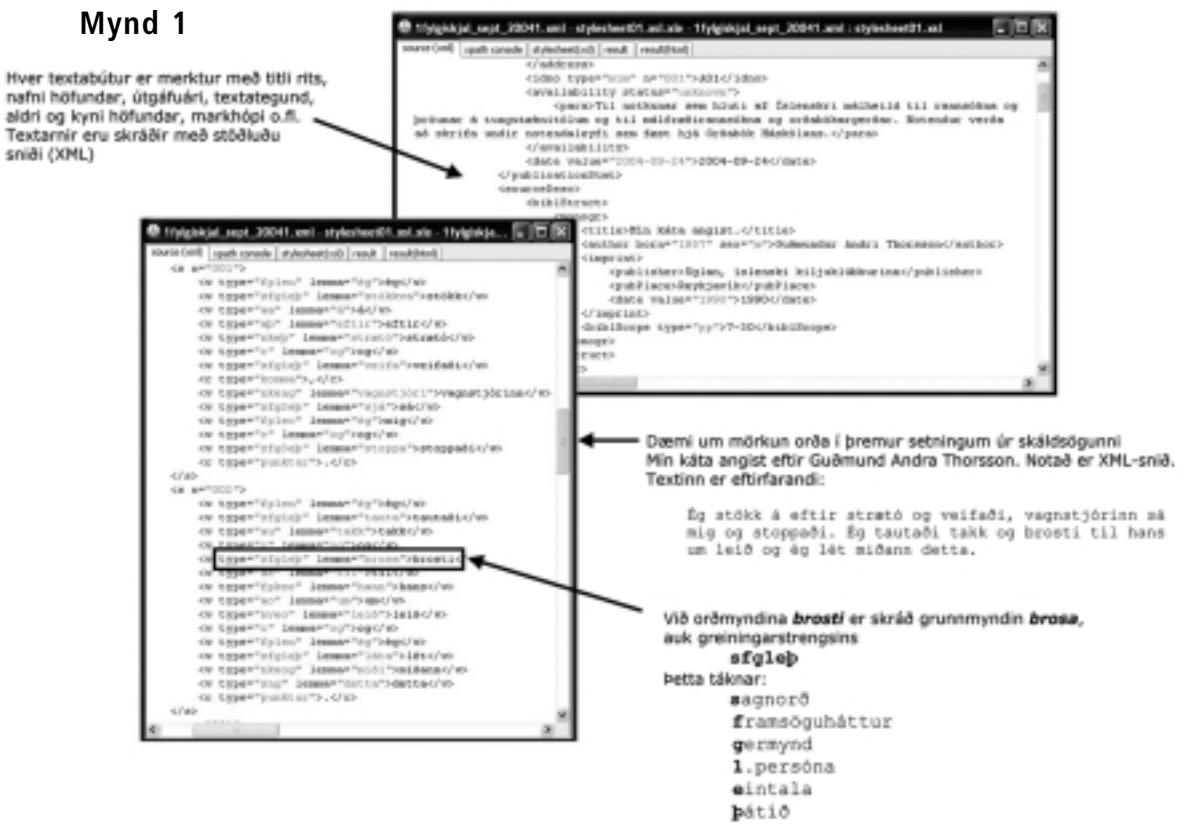
[1] http://www.natcorp.ox.ac.uk/
[2] http://korpus.dsl.dk/korpus2000/
[3] http://americannationalcorpus.org/

armynd orðsins. Nefnimynd nafnorða og fornafna er nefnifall og nafnháttur er nefnimynd sagna. Taka má sem dæmi setningarbrotið *ég sagði*. Nefnimynd fornafnsins *ég* er *ég* og markið verður `fp1en`, þar sem `f` táknar fornafn, `p` táknar persónufornafn, `1` táknar fyrstu persónu, `e` táknar eintölu og `n` táknar nefnifall. Nefnimynd sagnarinnar *sagði* er *segja* og markið verður `sfg1eþ` þar sem `s` táknar sagnorð, `f` táknar framsöguhátt, `g` táknar germynd, `1` táknar fyrstu persónu, `e` táknar eintölu og `þ` táknar þátíð.

Hvernig eru málheildir notaðar?

Notendur málheildarinnar eru einstaklingar, fyrirtæki og stofnanir sem vinna að orðabókargerð, margvíslegum tungutækniverkefnum og rannsóknum á íslensku nútímamáli. Úr málheildinni má lesa ýmiss konar gagnlegan fróðleik, t.d. upplýsingar um tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerð og merkingu eins og þegar er getið. Málheildir gefa einnig upplýsingar um hvernig tiltekið tungumál er notað á tilteknum tíma. Þær gefa vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti.

Mynd 1



Mörkuð málheild er því undirstaða fyrir þróun þýðingarforrita og mikilvæg fyrir nútíma orðabókargerð. Margir útgefendur orðabóka byggja nú gerð orðabóka á stórum mörkuðum málheildum. Upplýsingar sem fást úr markaðri málheild má einnig nota við gerð ýmissa tungutæknitóla, t.d. fyrir talgreiningu og talgervingu. Einnig eru slíkar upplýsingar nauðsynlegar við þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði. Mörg tungutæknitól af þessu tagi nýtast sérstaklega fyrir blinda, heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika.

Mörkuð íslensk málheild

Stefnt er að því að setja saman á næstu þremur árum málheild með íslenskum textum sem hafa að geyma um 25.000.000 orð. Árið 2002 veitti menntamálaráðuneytið styrk til verkefnis sem fólst í því að gera tilraunir með búnað til að marka íslenskan texta á vélrænan hátt.[4] Í tilraununum var notað textasafn sem var gert vegna *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991). Í því safni eru um 500.000 orð og fylgir hverri orðmynd nefnimynd og mark og hefur greining orða í textasafninu verið leiðrétt handvirkt. Textasafn orðtíðnibókarinnar verður því notað sem fyrsti vísir að málheildinni. Stefnt er að því að safna auk þess efni úr 900-1.000 textabútum sem skiptast á tiltekinn hátt eftir uppruna og efni. Hámarksstærð hvers textabúts verður 40.000 orð en aldrei er tekinn heill texti. Ef texti er styttri en 40.000 orð er 10% af textanum sleppt.

Valdir verða textar úr ritum sem gefin hafa verið út frá árinu 2000. Stefnt er að því að um 60% textanna komi úr bókum, 25% úr blöðum og tímaritum, 5-10% verði úr öðru útgefnu efni, 5-10% verði óútgefið efni og minna en 5% verði efni sem er skrifað til upplestrar. Enn fremur er stefnt að því að um 25% af textunum séu skáldverk og um 75% verði nytjatexti sem skiptist milli texta um hagnýtt vísindi, náttúrufræði, þjóðfélagsfræði, heimsmál, viðskipti, listir, trúarbrögð, heimspeki og tómstundir.

Orð í textunum verða greind á vélrænan hátt og er stefnt að um 90% nákvæmni. Hverri orðmynd í málheildinni á að fylgja nefnimynd orðsins og mark. Stefnt er að því að mörkun um einnar milljónar lesmálsorða í málheildinni verði leiðrétt handvirkt, þ.e. um 500.000 orð til viðbótar þeim 500.000 orðum sem þegar hafa verið greind.

Málheildir eru venjulega skráðar með stöðluðu sniði til þess að tryggja að sem flestir sem nota ólíkar tölvur og hugbúnað geti nýtt efnið. Notuð verður XML-útgáfa af sniði fyrir málheildir sem TEI-samtökin (TEI: *Text Encoding Initiative*) hafa skilgreint. Í þessu sniði er gert ráð fyrir að hverjum textabút fylgi haus þar sem skráðar eru margvíslegar upplýsingar um textann, höfund hans o.fl.

[4] Sjá greinargerð um gerð markara fyrir íslenskan texta annars staðar í þessu hefti

Í mynd 1 er sýnt er dæmi um skráningu textabrots með þremur setningum úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson. Fremst er haus þar sem eru upplýsingar um textann, síðan koma orðin í textanum ásamt nefnimynd þeirra og marki. Ekki er víst að þetta dæmi sýni endanlega gerð þess sniðs sem notað verður fyrir málheildina.

Söfnun texta

Til þess að raunhæft sé að koma þessu í kring er nauðsynlegt að semja við rétthafa texta um að fá efni án þess að greitt sé fyrir það. Rétthöfum verður því gerð grein fyrir hvernig aðgangur verður veittur að málheildinni þegar hún verður komin í notkun.

Ráðgert er að nýta það tækifæri sem nú gefst til textaöflunar til þess að bæta einnig textasöfn Orðabókar Háskólans. Textum verður fyrst komið fyrir í textasafni Orðabókarinnar og síðan verða textabrot sótt þangað til nota í málheildinni.

Til þess að geta valið texta er ráðgert að fá lista úr bókasafnskerfinu Gegni yfir efni sem gefið var út árið 2000 og síðar. Beita þarf öðrum aðferðum til þess að finna óútgefið efni og efni sem er ætlað til upplestrar. Einungis verður valinn tölvutækur texti.

Talmáli verður ekki safnað sérstaklega en reynt verður að fá afnot af tölvutækum skrám sem þegar eru til og geyma umritað talað mál.

Mörkun og hjálparskrár

Eins og þegar hefur verið getið veitti menntamálaráðuneytið styrk árið 2002 til verkefnis sem fólst í tilraunum til að marka íslenskan texta á vélrænan hátt. Vinna við verkið hófst síðla árs 2002 og var lokið í upphafi árs 2004. Niðurstöður verkefnisins verða nýttar við mörkun texta í málheildinni.

Við mörkunina þarf einnig að nota ýmsar hjálparskrár og orðasöfn. Stærst þessara hjálparskráa er orðasafn sem gert hefur verið úr beygingarlýsingu íslensks nútímamáls. Beygingarlýsingin var einnig gerð fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins. Í beygingarlýsingunni eru allar beygingarmyndir um 170.000 íslenskra orða. Einnig hefur verið aflað skráa yfir mannanöfn, örnefni, heiti fyrirtækja og skammstafanir.

Hvernig verður verkið unnið?

Gert er ráð fyrir að í upphafi verksins verði lögð áhersla á að skilgreina verkþætti og afla forrita og annarra verkfæra eða búa þau til. Einnig verður unnið við að undirbúa öflun texta. Nokkur vinna mun felast í því að fá leyfi til þess að fá að nota textana í málheildinni. Á fyrsta ári verkefnisins verður lögð áhersla á að safna textum sem frjáls aðgangur er að.

Þegar búið er að ná í textana þarf að koma þeim í vinnsluhæft form, m.a. með því að

hreinsa úr þeim prentskipanir og ganga frá neðanmálsgreinum og myndatextum.

Einnig þarf að efnisflokka textana og skrá ýmiss konar upplýsingar um þá, t.d. uppruna, höfund, birtingu og tímasetningu. Þá tekur við ýmiss konar forvinnsla sem felst m.a. í því að merkja málsgreinaskil á ótvíræðan hátt, leysa úr skammstöfunum, taka ákvörðun um meðferð nafna, talna og dagsetninga.

Síðan verður beitt þeim aðferðum við mörkun sem skilgreindar voru í verkefninu sem greint var frá hér að ofan. Aðferðirnar voru þróaðar með því að láta tiltekin forrit læra greiningu af textasafni *Íslenskrar orðtíðnibókar*. Val texta í textasafn Orðtíðnibókarinnar takmarkaðist af því hvaða textar voru tiltækir í tölvutæku formi þegar textunum var safnað. Þess vegna vega bókmenntatextar þyngra en æskilegt væri. Stefnt er að því að leiðrétta handvirkt til viðbótar greiningu um 500.000 orða til þess að úthlutun marka verði nákvæmari. Stefnt er að því að hlutföll milli textategunda í því milljón orða safni þar sem mörkun hefur verið leiðrétt verði lík hlutföllum í málheildinni allri. Fyrst verður tekinn fyrir textabútur með um 100.000 orðum og orðin mörkuð með þeim aðferðum og gögnum sem þegar eru til. Mörkunin verður síðan leiðrétt handvirkt. Leiðrétti búturinn verður þá lagður við textasafn orðtíðnibókarinnar og forritin látin læra af stækkuðu textasafni. Þannig verður haldið áfram þangað til fyrir liggur safn með um 1.000.000 orðum þar sem mörkun hefur verið leiðrétt handvirkt. Sá lærdómur sem dreginn verður af þessu safni verður síðan notaður til að marka texta sem síðar verður bætt við málheildina.

Rekstur málheildar

Ekki hefur verið ákveðið hvar málheildin verður vistuð né hvernig veittur verður að henni aðgangur. Nauðsynlegt er að ákveða þetta svo gera megi rétthöfum texta grein fyrir þessum atriðum áður en þeir gefa leyfi til þess að textar þeirra verði notaðir.

Umsjón með verkinu

Verkefnið „Mörkuð íslensk málheild" er unnið af Orðabók Háskólans samkvæmt samningi við menntamálaráðuneytið frá 14. júní 2004. Verkinu á að skila í júní 2007. Orðabók Háskólans leggur m.a. til verksins aðstöðu og markað textasafn *Íslenskrar orðtíðnibókar*. Verkefnisstjóri er Sigrún Helgadóttir (sigrunh@lexis.hi.is). Skipuð hefur verið verkefnisstjórn sem vinnur með verkefnisstjóra. Í verkefnisstjórninni eiga sæti Ásta Svavarsdóttir, Eiríkur Rögnvaldsson og Kristín Bjarnadóttir.

Heimildir

Jörgen Pind (ritstj.), Friðrik Magnússon, Stefán Briem. 1991. *Íslensk orðtíðnibók.* Orðabók Háskólans, Reykjavík.


TUNGUTÆKNI