Orðabókar- og rannsóknarverkefni

Beygingarlýsing íslensks nútímamáls

Markmið verkefnisins *Beygingarlýsing íslensks nútímamáls* er að koma upp beygingarlýsingu á tölvutæku formi til birtingar á vefsíðu Orðabókar Háskólans og til nota í ýmiss konar tungutækniverkefni.

Verkefnið á rætur að rekja til þess að Orðabókin og Edda hf. fengu styrk frá verkefnisstjórn menntamálaráðuneytisins í tungutækni árið 2002. Gengið var til samninga um verkið í ágúst 2002. Samkvæmt samningi þessum skyldu vera 100 þúsund beygingardæmi í Beygingarlýsingunni. Orðabókin sá að öllu leyti um vinnuna og lagði til húsnæði, alla aðstöðu og aðgang að gögnum. Edda hf. lagði til beygingarlýsingu þá sem unnin var fyrir tölvuútgáfu *Íslenskrar orðabókar* (2000). Þessum áfanga verksins lauk í febrúar 2004 þegar menntamálaráðuneytinu var afhentur geisladiskur með útgáfu 1.0 af Beygingarlýsingunni, alls 173.389 beygingardæmi á formi XML-skráa.

Beygingardæmunum var komið fyrir á vefsíðu Orðabókarinnar jafnhliða því að unnið var að útgáfunni sem ætluð var til tungutækninota. Slóðin er http://www.lexis.hi.is/beygingarlysing.

Á vefsíðunni eru nú tæplega 200 þúsund beygingardæmi með öllum beygingarmyndum hvers orðs. Í safninu eru beygingar ósamsettra og samsettra orða úr almennu nútímamáli, auk mannanafna. Takmarkið með birtingu beygingardæmanna er að einskorða efnið við raunverulegar myndir hvers orðs, þ.e. að sýna afbrigði þar sem það á við og eyður þar sem beygingarmyndir eru ekki til. Þannig eru t.d. gefin tvö afbrigði af þágufalli eintölu af orðinu *spölur*, þ.e. *speli/spöl* en fleirtölu af eintöluorðum er sleppt (t.d. af *sykur* og *kaffi*) og germynd er

ekki sýnd af miðmyndarsögnum (t.d. af *óttast*). 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, lýsingarhátta og viðskeyttra spurnarmynda (t.d. *ferðu* og *fórstu* af sögninni *fara*). Alls geta beygingarmyndir sagnar orðið 109, án afbrigða. Beygingarmyndir lýsingarorðs sem stigbreytist eru 120, án afbrigða, ef engar eyður eru í beygingardæminu.

Orðaforðinn í Beygingarlýsingunni var upprunalega fenginn úr 3. útgáfu *Íslenskrar orðabókar* og úr söfnum Orðabókarinnar en nú er verið að vinna við viðbótarefni úr öðrum heimildum. Helstu heimildir við rannsóknir á einstökum orðum og beygingarflokkum eru Ritmálsskrá og Textasafn Orðabókarinnar, auk handbóka, greina og ritgerða um íslenska málfræði.

Ítarleg beygingarlýsing er grundvöllur að vélrænni greiningu á íslenskum textum, nauðsynlegur undanfari orðflokkagreiningar og setningagreiningar. Beygingarlýsingin nýtist t.d. við mörkun texta og við gerð leitarvéla, auk þess að vera forsenda skilvirkrar orðabókargerðar og heimildasöfnunar. Þá er Beygingarlýsingin einnig nauðsynlegur efniviður við gerð leiðréttingar- og þýðingarforrita.

Beygingarlýsingin er nú notuð í ýmsum verkefnum Orðabókar Háskólans, t.d. við vinnu við ÍSLEX, við Markaða íslenska málheild og við leit í Textasafninu. Þá hefur tekist gott samstarf við hugbúnaðarfyrirtækið Spurl ehf. sem notar Beygingarlýsinguna í verk sín, t.d. í leitarvélina Emblu sem er á vefsíðu Morgunblaðsins og við leit í símaskránni, á ja.is.

Nú er unnið að nýjum gagnagrunni fyrir Beygingarlýsinguna í samvinnu við Spurl ehf., en styrkur til verkefnisins fékkst úr Tækniþróunarsjóði haustið 2005. Nýi gagnagrunnurinn mun auðvelda alla umsýslu við Beygingarlýsinguna og bæta leitaraðgang á vefsíðu Orðabókarinnar til muna.

Menntamálaráðuneytið hefur með samningi falið Orðabókinni umsjón með og ráðstöfun á Beygingarlýsingunni til tungutækninota. Gerður er samningur um afnot af verkinu við hvern leyfishafa þar sem kveðið er á um skilmála.

Verkefnisstjóri er Kristín Bjarnadóttir. Faglega verkefnisstjórn skipa Eiríkur Rögnvaldsson, Sigrún Helgadóttir og Þórdís Úlfarsdóttir.

Kristín Bjarnadóttir

ISLEX – Íslensk-norræn veforðabók

ISLEX er orðabókarverk sem er unnið á vegum Orðabókar Háskólans í samstarfi við stofnanir á Norðurlöndum. Áætlað er að verkið taki fimm ár í vinnslu og er stefnt að því að ljúka því árið 2011.

Um er að ræða rafræna orðabók með um 50.000 íslenskum uppflettiorðum ásamt þýðingum á sænsku, norsku og dönsku. Orðabókin á að endurspegla íslenska málnotkun samtímans og er markmiðið að gera hana öllum aðgengilega á vefnum. Við ritstjórnina er notaður veftengdur gagnagrunnur sem er hannaður sérstaklega fyrir þetta verkefni. Gagnagrunnurinn gerir mögulegt að ritstjórnarvinnan fari fram samhliða í mörgum löndum.

Verkið hefur margvíslegt gildi í norrænu samstarfi, jafnt opinberu sem almennu. Það á að nýtast öllum þeim sem þurfa á íslensk-norrænum orðabókum að halda en sérstaklega verður leitast við að sinna þörfum sænskra, norskra og danskra notenda, ekki síst þýðenda úr íslensku.

ISLEX er samstarfsverkefni Orðabókar Háskólans og þriggja stofnana á Norðurlöndum, *Institutionen för svenska språket* við Gautaborgarháskóla, *Nordisk Institutt* við háskólann í Bergen og *Det Danske Sprog- og Litteraturselskab* í Kaupmannahöfn.

ISLEX-orðabókin á að birtast á vefnum og er leitast við að láta margvíslega möguleika rafrænnar miðlunar njóta sín. Þar verður því að finna ýmis nýmæli í íslenskri orðabókargerð, til dæmis litmyndir, hreyfimyndir, framburð orða og önnur hljóð. ISLEX-gagnagrunnurinn er enn í mótun og eftir er að ljúka frekari forritunar- og tæknivinnu, m.a. við þá hluta verksins sem snúa að endanlegri birtingu efnisins.

Orðaforðinn í verkinu, um 50.000 flettur, er að mestu kominn inn í gagnagrunninn. Auk flettiorðanna eru í ISLEX-orðabókinni tilfærð um 150.000 samsett orð, svokallaðar virkar samsetningar, sem ekki eru flettiorð en er skipað undir viðkomandi grunnorð. Þessi samsettu orð sýna m.a. hversu virk grunnorðin eru í orðmyndun og gefa vísbendingar um merkingartilbrigði þeirra.

Ætlunin er að sýna fullt beygingardæmi allra beygjanlegra flettiorða ISLEX-orðabókarinnar og nýta í því skyni tungutækniverkefnið *Beygingarlýsing íslensks nútímamáls*. Er þetta eina íslenska orðabókin þar sem beygingar eru sýndar í heild sinni og án skammstafana.

Orðasamböndum er ætlað að skipa veigamikinn sess í orðabók-

arverkinu og er nú unnið að því að móta fyrirkomulagið á þeim. Framsetning á setningarlegum upplýsingum er einnig í mótun. Þetta varðar þætti eins og fallstjórn sagna og dæmigerð fylgiorð þeirra, svo dæmi séu nefnd.

Greining og vinnsla efnisins er komin nokkuð á veg og í fyrsta áfanga er lögð áhersla á frágang nafnorðanna.

Nú er hafin vinna við að bæta inn þýðingum á norsku, sænsku og dönsku. Verið er að leggja á ráðin um skipulag þýðingarvinnunnar og endanlega framsetningu markmálanna.

Þórdís Úlfarsdóttir

Mörkuð íslensk málheild

Á vegum Orðabókar Háskólans er nú unnið að gerð markaðrar íslenskrar málheildar. Verkið er styrkt af tungutækniverkefni menntamálaráðuneytisins. Verkið hófst um mitt ár 2004 og skal því lokið árið 2007.

Með **markaðri málheild** (e. *tagged corpus*) er átt við safn fjölbreyttra textabúta sem hafa verið greindir á málfræðilegan hátt. Málheildin er geymd í rafrænu formi, venjulega í stöðluðu sniði. Hverjum textabút fylgja upplýsingar um textann sem búturinn er úr og hverri orðmynd fylgir **nefnimynd** (e. *lemma*) og greiningarstrengur, sem kallast **mark** (e. *tag*) og sýnir orðflokk og málfræðilega greiningu orðsins. Nefnimynd nafnorða er nefnifall eintölu, nefnimynd fornafna er nefnifall eintölu í karlkyni og nafnháttur er nefnimynd sagna. Taka má sem dæmi setningarbrotið *ég sagði*. Nefnimynd persónufornafnsins *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.

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. Stefnt er að því að í málheildinni verði í fyrstu um 25.000.000 lesmálsorð sem skiptast í um 900 textabúta. Hámarksstærð hvers textabúts verður 40.000 orð. Aldrei er tekinn heill texti. Ef texti er styttri en 40.000 orð er 10% af textanum sleppt.

Þegar vinna við málheildina var skipulögð var ekki gert ráð fyrir að safnað yrði talmáli og var það aðallega vegna þess hversu tímafrekt og dýrt það er. Nú hefur hins vegar komið í ljós að málheildin getur fengið talmálstexta úr öðrum verkefnum. Í fyrsta lagi er þar um að ræða texta sem hefur verið safnað á vegum verkefnisins *ÍSTAL – Íslenskur talmálsbanki* sem unnið var fyrir styrk frá Tæknisjóði á árunum 1999-2001. Í öðru lagi má nefna umræður á Alþingi sem var safnað á vegum verkefnisins *Tilbrigði í setningagerð* sem hlaut öndvegisstyrk frá RANNÍS 2005. Í þriðja lagi má nefna hópviðtöl um tökuorð og erlend áhrif sem voru hljóðrituð vegna norrænnar rannsóknar um viðhorf til tökuorða en verða fullskráð og frágengin á vegum verkefnisins *Tilbrigði í setningagerð*.

Stofn málheildarinnar er textasafn sem var útbúið vegna vinnu við *Íslenska orðtíðnibók* sem kom út á vegum Orðabókar Háskólans 1991. Í því safni eru um 500.000 lesmálsorð 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. Árið 2002 veitti menntamálaráðuneytið styrk til verkefnis sem fólst í því að gera tilraunir 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. Einnig hafa verið gerðar tilraunir við að finna nefnimyndir orða á vélrænan hátt. Stefnt er að því að lesmálsorð verði greind á vélrænan hátt með um 90% nákvæmni.

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 upphaflega gerð fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins en hefur síðan verið aukin verulega á vegum Orðabókar Háskólans. Einnig hefur verið aflað skráa yfir mannanöfn, örnefni, heiti fyrirtækja og skammstafanir.

Málheildir eru venjulega skráðar með stöðluðu sniði til þess að

tryggja að sem flestir geti nýtt efnið þrátt fyrir að menn noti ólíkar tölvur og hugbúnað. 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.

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. 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.

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.

Gerður hefur verið samningur við menntamálaráðuneytið um að Orðabók Háskólans visti málheildina og veiti aðgang að henni. Ráðgert er að málheildin verð til ráðstöfunar til rannsókna í tungutækni og til þróunar tungutæknitóla. Einnig er stefnt að því að veita aðgang að málheildinni á vefsetri Orðabókar Háskólans með sérstökum leitarhugbúnaði.

Til þess að unnt sé að hafa opinn aðgang að málheildinni er nauðsynlegt að semja við rétthafa texta um hvernig birtingu skuli háttað. Í því sambandi skiptir höfuðmáli að engir textar verða birtir í heild í málheildinni þannig að útilokað er að endurgera verk með textum sem þar eru geymdir.

Ráðgert er að textabútar sem mynda málheildina verði sóttir í textasafn Orðabókar Háskólans. Í textasafninu eru textar af ýmsu tagi og frá ýmsum tímum en tækifærið verður notað til þess að auka það.

Verkefnisstjóri í verkinu er Sigrún Helgadóttir. Verkefnisstjórn,

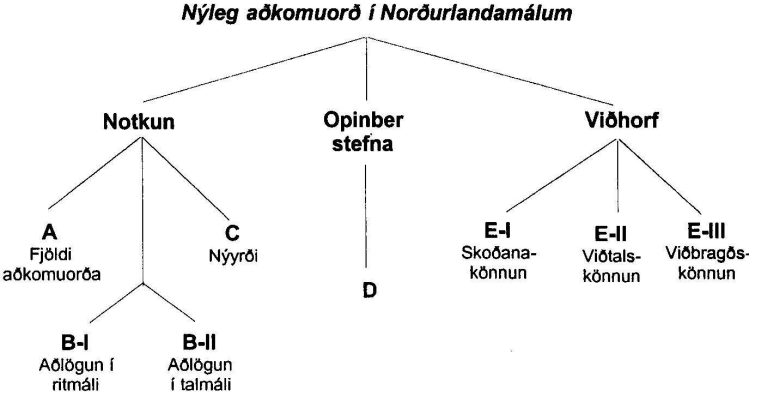
skipuð Ástu Svavarsdóttur, Eiríki Rögnvaldssyni og Kristínu Bjarnadóttur starfar með verkefnisstjóra.

Sigrún Helgadóttir

Rannsókn á aðkomuorðum í Norðurlandamálum

Um nokkurra ára skeið hefur verið í gangi samanburðarrannsókn á nýlegum aðkomuorðum í Norðurlandamálum (*Moderne importord i språka i Norden)*, sem nær til íslensku, dönsku, norsku, sænsku, finnsku, finnlandssænsku og færeysku. Meginmarkmið hennar er að skoða tiltekna þætti sem varða aðkomuorð og bera þá saman í málunum sjö. Með *aðkomuorðum* er átt við öll orð sem eiga uppruna í öðrum málum og koma fyrir í norrænu textasamhengi, allt frá fullaðlöguðum tökuorðum til erlendra orða sem bregður fyrir í textunum; *nýleg* merkir í rannsókninni orð sem hafa borist í Norðurlandamálin eftir seinni heimsstyrjöld (ca. 1945 eða síðar). Á þessu tímabili er fyrst og fremst um ensk áhrif að ræða og gildir það jafnt um öll málin.

Rannsóknin skiptist í þrjá meginhluta: rannsókn á *notkun* aðkomuorða, umfjöllun um *opinbera málstefnu* gagnvart aðkomuorðum og rannsókn á *viðhorfi almennings* til aðkomuorða. Þeir skiptast síðan flestir í rannsóknarþætti eins og sýnt er á eftirfarandi skýringarmynd:



Notkunarhlutinn skiptist í þrennt. Í fyrsta lagi (A) var kannaður hlutfallslegur fjöldi aðkomuorða í dagblaðatextum á tungumálunum sjö (blöð frá 1975 og 2000). Í öðru lagi (B) var rannsökuð aðlögun aðkomu-

orða í töluðu og rituðu máli, m.t.t. framburðar, stafsetningar, beyginga og orðmyndunar. Rannsóknin fólst annars vegar í spurningakönnun (talmál) og hins vegar í frekari athugun og úrvinnslu á dagblaðaefninu úr A-hluta verkefnisins (ritmál). Í þriðja lagi (C) voru athuguð nýyrði sem ætlað er að leysa aðkomuorð af hólmi og útbreiðslu þeirra. Rannsóknin beinist að ritmáli og styðst m.a. við dagblaðatextana sem notaðir voru í A. Annar hluti verkefnisins (D) felst í sögulegri lýsingu á opinberri málstefnu hvers lands á tímabilinu 1850–2000 og sáu málnefndir þátttökulandanna um þann þátt. Viðhorfsrannsóknirnar skiptast í þrjá rannsóknarþætti. Sá fyrsti fólst í skoðanakönnun þar sem fjöldi fólks (500-1000 eftir löndum) var spurður um ákveðin atriði sem snerta aðkomuorð í móðurmáli þátttakenda (E-I). Viðtalskönnun (E-II) gegnir því hlutverki að varpa skýrara ljósi á niðurstöður skoðanakönnunarinnar og gefa færi á því að kafa dýpra í ákveðin atriði með viðtölum við tiltölulega lítinn hóp fólks. Markmið viðbragðskönnunarinnar (E-III), sem nær til nokkurra stórra hópa, er að kalla fram viðbrögð fólks við notkun aðkomuorða. Tilgangurinn með því að skipta rannsókninni niður í marga smærri þætti er ekki síst sá að skoða viðfangsefnið frá ýmsum hliðum og með mismunandi rannsóknaraðferðum þannig að bæði sé hægt að bera saman niðurstöðurnar innbyrðis og á milli mála.

Að rannsókninni stendur hópur norrænna fræðimanna og háskólastúdenta undir stjórn Helge Sandøy, prófessors við háskólann í Bergen. Íslensku þátttakendurnir eru Guðrún Kvaran (C-hluti) og Ásta Svavarsdóttir (B-I og B-II) á Orðabók Háskólans, Kristján Árnason (E-I), Hanna Óladóttir (E-II) og Halldóra Björt Ewen (E-III og a.n.l. E-II) við Háskóla Íslands. Auk þeirra hafa aðrir komið að einstökum rannsóknarþáttum, einkum stúdentar við HÍ, og Ari Páll Kristinsson skrifaði skýrslu um opinbera stefnu gagnvart aðkomuorðum (D-hluti) fyrir Íslenska málnefnd. Rannsóknirnar hafa verið styrktar af Norrænu málráði (Nordisk sprogråd), NOS-H, Norræna menningarsjóðnum (Nordisk kulturfond), Rannsóknasjóði Háskóla Íslands og Nordplus sprog.

Niðurstöður úr öllum þáttum rannsóknarinnar eru birtar í *ritröðinni Moderne importord i språka i Norden* hjá Novus í Noregi. Þrjú rit eru þegar komin út. Í því fyrsta eru fyrirlestrar frá ráðstefnu sem haldin var í Bergen haustið 2002 *(Med 'bil' i Norden i 100 år. Ordlaging og tilpassing av utalandske ord.* Oslo 2003; ritstjóri Helge Sandøy), annað

er greinasafn um opinbera málstefnu á Norðurlöndunum *(Normering av importord i Norden. Historikk*. Oslo 2004; ritstjórar Helge Sandøy og Jan-Ola Östman) og það þriðja er skýrsla um sænsku viðtalsrannsóknina á viðhorfum málnotenda til aðkomuorða og erlendra máláhrifa (*Teamwork? Man kan lika gärna samarbeta. Svenska åsikter om importord*. Oslo 2005; höfundur Catharina Nyström Höög). Rit með niðurstöðum úr öðrum rannsóknarþáttum eru væntanleg á árinu 2006. Auk þess voru niðurstöður rannsóknarinnar í heild kynntar á ráðstefnu í Kaupmannahöfn í árslok 2005 og einnig hefur verið fjallað um verkefnið eða hluta þess í greinum og fyrirlestrum einstakra fræðimanna, bæði hérlendis og erlendis.

Sjá nánar um verkefnið: http://www.hf.uib.no/moderne/.

Ásta Svavarsdóttir

Spænsk-íslensk, íslensk-spænsk orðabók

Fyrsti vísir að orðabók á Íslandi var baskneskt-íslenskt orðasafn sem fæddist af samskiptum Íslendinga og baskneskra sjómanna á Vestfjörðum síðla 16. aldar og fram á þá 17.[2] Síðan þá hafa samskipti við Spán farið sívaxandi og nú er Spánn það land sem flestir Íslendingar sækja heim. Að sama skapi hefur Rómanska Ameríka laðað unnendur spænskrar tungu til sín og hefur vægi spænsku sem viðskiptamáls aldrei verið meira. Hefur það leitt til þess að spænskunemum hefur fjölgað þannig að flestir framhaldsskólar landsins bjóða upp á spænsku sem þriðja mál. Þar af leiðandi hefur skortur á nútímalegri spænsk-íslenskri orðabók aldrei verið jafn aðkallandi.

Fram til þessa hafa verið gefnar út tvær spænsk-íslenskar orðabækur. Sú fyrri er *Spænsk-íslensk vasaorðabók* eftir Elisabeth Hangartner Ásbjörnsson og Elvira Herrera Ólafsson (1978). Hún hefur nýst framhaldsskólanemum í fyrsta og öðrum áfanga spænskunáms en er barn síns tíma og gefur ekki ítarlegri upplýsingar um orð en kyn þeirra. Málnotkunardæmi eru engin. Síðari bókin, *Spænsk-íslensk orðabók,* er

[2]Basknesk íslensku orðasöfnin er að finna í þremur óprentuðum handritum. Eitt þeirra er á Árnastofnun (AM 987 4to) og hin er að finna á handritadeild Landsbókasafns (JS 284 8vo og JS 401 4to).

stærri.[3] Hún er eftir Sigurð Sigurmundsson frá Hvítárholti og kom hún út árið 1973 og aftur árið 1995. Bókin ber þess ýmis merki að höfundur hafði ekki menntun í spænsku en eftir stendur engu að síður aðdáun á frumkvöðlaverki sjálfmenntaðs manns.

Þrátt fyrir augljósa þörf á spænsk-íslenskri orðabók tók þó nokkurn tíma að afla fjár til að hægt væri að hefjast handa við gerð orðabókarinnar. Verkið er kostnaðarsamt og enginn fjárhagslegur ábati fyrir bókaforlög af því að leggja í slíkt verk. Því var nauðsynlegt að fá öfluga liðsmenn til að hleypa verkinu af stað. Höfðinglegur styrkur frá Menntamálaráðuneyti og Minningarsjóði Margrétar Björgólfsdóttur gerði Háskólanum í Reykjavík og Eddu útgáfu kleift að setjast að samningaborði og hefja samvinnu um verkið.

Til verksins voru ráðnir 3 sérfræðingar í spænsku í rúmlega tvö stöðugildi, þær Guðrún H. Tulinius, Ragnheiður Kristinsdóttir og Sigrún Á. Eiríksdóttir. Ritstjóri verksins er Margrét Jónsdóttir og Laufey Leifsdóttir kemur að verkinu fyrir hönd Eddu útgáfu.

Til að spara tíma og fé var ákveðið að fá aðkeyptan orðabókargrunn frá hinni virtu útgáfu Harper-Collins. Um er að ræða rúmlega 20.000 flettur í hvora átt og er ætlunin að auka við þann fjölda sérvöldum orðaforða, þannig að spænsk-íslenski hlutinn innihaldi um 25.000 flettur. Íslensk-spænski hlutinn verður svipaður að umfangi.

Við val á gagnagrunni var haft í huga að bókin henti byrjendum í spænsku og nemendum sem enn eru ekki komnir það langt í að tileinka sér málið að þeir geti notast við spænsk-spænskar orðabækur. Jafnframt er miðað við að bókin nýtist öðrum, s.s. ferðamönnum, húseigendum á Spáni og þeim sem stunda viðskipti við hinn spænskumælandi heim. Þó íslensk-spænski hlutinn miðist aðallega við Íslendinga í spænskunámi, er einnig tekið mið af spænskumælandi fólki á Íslandi sem ekki hefur íslensku að móðurmáli.

Vinnan við orðabókina hófst í október 2005. Notast er við orðabókarforritið *Lexa* sem skrifað var af sérfræðingum á orðabókardeild Eddu. Ætlunin var að flytja Harper-Collins grunninn inn í Lexuna en þar sem samningar um kaup á orðagrunninum drógust á langinn hefur danskur grunnur frá GADE útgáfunni verið lagður til grundvallar og viðbætur og breytingar færðar inn í hann eftir prentuðu orðabókinni frá Harper-Collins. Vinnan hefur því aðallega verið fólgin í því að þýða ýmist úr dönsku eða ensku. Verkið hefur gengið vel og

[3] Sigurður Sigurmundsson, *Spænsk – íslensk orðabók.* Reykjavík 1995.

er innslætti á fyrri hluta bókarinnar lokið og yfirlestur hafinn. Vonir standa til að sá hluti verði aðgengilegur á veraldarvefnum á haustmánuðum 2006.

Í orðabókinni verður algengasti orðaforði spænskrar tungu undanfarinna ára, þ.á m. orð frá spænskumælandi löndum Ameríku. Ýmis orð eru sérmerkt annaðhvort Spáni (ESP) eða Ameríku (AM) og fyrir kemur að orð sé sérmerkt einstöku landi þar. Öll algengustu orð og hugtök eru uppflettiorð í bókinni. Gerð er ítarleg grein fyrir merkingu þeirra og notkun og fjölmörg notkunardæmi eru birt í bókinni. Eitt sérkenni bókarinnar eru upplýsingar af menningarlegum toga, sem ætlunin er að bæta verulega við. Einnig eru upplýsingar um fjölda gagnlegra skammstafana og málfræðilegar upplýsingar, þ.á m. um beygingu sagna. Starfsmenn orðabókarinnar hafa bætt við talsverðum sérhæfðum orðaforða t.d. varðandi atvinnugreinar hér á landi, dýralíf, lögfræði, hagfræði, viðskipti, listir, heilbrigðismál, frístundir, ferðalög, jarðfræði og margt fleira. Er ætlunin að hafa í íslenska-spænska hlutanum samskonar menningarlegar upplýsingar og finna má í þeim spænsk-íslensk, en eitt af markmiðum ritstjórnar er að auka menningarlæsi þeirra sem nýta sér bókina, hvort sem um spænskumælandi lönd er að ræða eða Ísland.

Þegar fyrri hluta er lokið verður orðunum varpað yfir í hinn hlutann þannig að öll þau orð sem hafa einfalda skýringu fara sjálfvirkt inn í íslenska-spænska hlutann. Dæmi: *orgullo*: nm, stolt, mont. Bæði þýðingarorðin, *stolt* og *mont*, fara inn sem íslensk uppflettiorð og munu starfsmenn greina orðin og finna dæmi um notkun þeirra. Einnig verður grunnurinn borinn saman við aðrar tvímála orðabækur sem Edda útgáfa gefur út sem og íslenskan orðagrunn útgáfunnar.

Ef áætlanir standast kemur spænsk-íslensk, íslensk-spænsk orðabók út árið 2008. Þar sem verkinu hefur miðað framar vonum er líklegt að svo verði. Er það von undirritaðra að hér sé á ferðinni orðabók sem stenst tíma- og kostnaðaráætlanir og verði lifandi vitnisburður um gott samstarf háskóla og atvinnulífs.

Guðrún H. Tulinius og Margrét Jónsdóttir

Tilbrigði í setningagerð

Rannsóknarverkefnið *Tilbrigði í setningagerð* hefur það meginmarkmið að gera grein fyrir ýmiss konar tilbrigðum í íslenskri setningagerð, athuga útbreiðslu þeirra og leitast við að skýra eðli þeirra og einkenni. Tilbrigðin eru skoðuð út frá landfræðilegri dreifingu, félagslegum breytum (t.d. kyni og aldri málnotenda) og mismunandi málaðstæðum (talmál/ritmál, formlegt/óformlegt mál o.s.frv.). Enn fremur felur verkefnið í sér rannsóknir á tilteknum atriðum í færeyskri setningagerð og samanburði þeirra við íslensku. Íslenska verkefnið er hluti af norræna verkefninu *Scandinavian Dialect Syntax* (ScanDiaSyn), sem miðar að því að rannsaka og bera saman mállýskubundin setningatilbrigði á öllu norræna málsvæðinu, þvert á mörk þjóðmálanna, og að koma upp aðgengilegum gagnagrunni með efniviði til setningafræðilegra rannsókna. Einnig hefur verið stofnað til tengsla og samvinnu við sambærileg verkefni annars staðar í Evrópu, einkum í Hollandi og á Norður-Ítalíu.

Í rannsóknunum er annars vegar byggt á spurningalistum sem lagðir verða fyrir fjölda málnotenda víða um land, bæði munnlega og skriflega. Með þeim verða rannsökuð valin atriði sem talin eru áhugaverð frá setningarlegu sjónarmiði. Áhersla verður lögð á atriði þar sem einhvers konar tilbrigði koma fram í málnotkun og getur munurinn þá ýmist verið á milli einstaklinga (t.d. eftir uppruna þeirra eða aldri) eða í máli sama einstaklings (t.d. eftir aðstæðum). Hins vegar byggjast rannsóknirnar á samfelldum textum úr rituðu og þó einkum töluðu máli og verkefnið felur m.a. í sér umfangsmikla efnissöfnun úr talmáli. Þar er einkum um það að ræða að safna saman efni sem þegar er til, ljúka nauðsynlegri skráningu þess og ganga frá því í aðgengilegu formi. Þannig hefur verið lokið frágangi á efni sem safnað var í verkefninu ÍSTAL — *Íslenskur talmálsbanki* á árunum 1999–2000 (sjálfsprottin, persónuleg samtöl) og skráningu lokið á efni sem safnað var fyrir rannsókn á aðkomuorðum í íslensku árið 2002 (MIN-*verkefnið*; viðtöl) og á umræðum frá Alþingi, sem teknar voru upp og frumskráðar á vegum þingsins. Alls eru þetta nálægt 50 klst. af hljóðrituðu og umrituðu efni. Auk þess hefur verið unnið að skráningu og undirbúningi að frekari úrvinnslu hljóðritana í vörslu Árnastofnunar sem geyma viðtöl við Vestur-Íslendinga og einnig hefur verið rætt við aðstandendur fleiri rannsóknarverkefna um afnot af efni frá þeim.

Tekist hefur samvinna við verkefnið *Mörkuð íslensk málheild*, sem unnið er að við Orðabók Háskólans (sjá verkefnislýsingu Sigrúnar Helgadóttur í þessu hefti), um úrvinnslu og greiningu talmálsefnisins. Munu bæði verkefni njóta góðs af samstarfinu. Lokið verður við umritun efnisins innan setningafræðiverkefnisins og gengið frá því með samræmdu sniði. Efnið verður síðan lagt inn í fyrirhugaða málheild, þar sem ekki hefði verið kostur á að hafa talmálsefni ef slík samvinna hefði ekki hefði komið til. Á móti kemur að málheildarverkefnið skilar vélrænni mörkun og greiningu efniviðarins, sem gerir hann miklum mun aðgengilegri til leitar og rannsókna.

Verkefnið *Tilbrigði í setningagerð* fékk veglegan öndvegisstyrk frá RANNÍS 2005 til þriggja ára. Verkefnisstjóri er Höskuldur Þráinsson, prófessor við HÍ, og aðrir í stjórn verkefnisins eru Eiríkur Rögnvaldsson, Jóhannes Gísli Jónsson og Sigríður Sigurjónsdóttir frá HÍ, Þórunn Blöndal frá KHÍ og Ásta Svavarsdóttir frá OH. Auk þeirra koma fleiri fræðimenn að verkefninu auk fjölmargra háskólastúdenta í íslensku.

Frekari upplýsingar um verkefni sem vísað er til má finna á eftirtöldum vefsíðum:

Tilbrigði í setningagerð (verkefnislýsing á ensku):
http://uit.no/scandiasyn/island/

ScanDiaSyn: http://uit.no/scandiasyn/scandiasyn/

ÍSTAL — *Íslenskur talmálsbanki*: http://www.hi.is/~eirikur/istal/

MIN (Morderne importord i språka i Norden):
http://www.hf.uib.no/moderne/

Mörkuð íslensk málheild: http://www.lexis.hi.is/malheild.htm

Ásta Svavarsdóttir

Tungutækniverkefni sem Orðabók Háskólans tekur þátt í

Frá árinu 2001 hefur Orðabók Háskólans rekið *Íslenskt upplýsingasetur um tungutækni*, sem kostað var af norrænu tungutækniáætluninni fram á mitt ár 2005. Setrið hefur tekið þátt í samstarfsneti norrænna upplýsingasetra, *NorDokNet*, www.nordoknet.org, en starfstíma þess lauk einnig á árinu. Upplýsingasetrið beitti sér fyrir samningu íslensks íðorðasafns í tungutækni, en fé fékkst til þess úr norrænu tungutækni-

áætluninni. Valdís Ólafsdóttir, meistaranemi í tungutækni, var ráðin til að leggja drög að safninu og vann hún við það um þriggja mánaða skeið (í 75% starfi). Verkið var unnið í samvinnu við Orðabanka Íslenskrar málstöðvar, og var íðorðasafnið opnað sem hluti bankans 1. september.

Þegar tungutækniáætlun menntamálaráðuneytisins lauk og verkefnisstjórn í tungutækni var lögð niður um áramótin 2004-5 yfirtók upplýsingasetrið vef hennar, www.tungutaekni.is. Til að halda utan um og ýta undir starfsemi á sviði íslenskrar tungutækni var ákveðið að stofna *Tungutæknisetur*, sem er samstarfsverkefni Orðabókarinnar, Málvísindastofnunar Háskóla Íslands og tækni- og verkfræðideildar Háskólans í Reykjavík. Í stjórninni situr einn fulltrúi frá hverjum samstarfsaðila, og er Sigrún Helgadóttir fulltrúi Orðabókarinnar. Meðal markmiða setursins er að halda árlega ráðstefnu um íslenska tungutækni, og verður sú fyrsta haldin nú í maí.

Norræna nýsköpunarmiðstöðin (Nordisk Innovationscenter) og Norræna rannsóknamiðstöðin (NordForsk) veittu norrænum upplýsingasetrum um tungutækni og samtökum iðnaðarins á Norðurlöndum styrk til forverkefnis sem nefnt var *NLTNet*. Styrknum var varið til að halda sameiginlegan fund í Kaupmannahöfn þar sem ræddar voru forsendur fyrir samvinnu um kynningu og eflingu tungutækni á Norðurlöndum. Þar var samþykkt að stefna að áframhaldandi samstarfi, og var gefin út skýrsla um niðurstöður fundarins – sjá www.nltnet.org. Snemma árs 2006 sendi sami hópur svo umsókn til Norrænu nýsköpunarmiðstöðvarinnar um nýtt tveggja ára verkefni, *Nordic ICT and Language*, þar sem ætlunin er að fylgja eftir þeirri stefnu sem mörkuð var í forverkefninu. Þegar þetta er ritað er ekki komið í ljós hvort styrkur fæst til þessa verkefnis.

Orðabókin og Háskóli Íslands tóku þátt í *Nordisk Netordbog*, samstarfsneti um rannsóknir og þróun í tungutækni, einkum margmála leit á netinu og í gagnabönkum. Aðrir þátttakendur í verkefninu eru Háskólinn í Bergen, Kungliga tekniska högskolan (KTH) í Stokkhólmi, Háskólinn í Helsinki og Center for sprogteknologi (CST) í Kaupmannahöfn, sem stýrir verkefninu (verkefnisstjóri Bente Maegaard). Norræna ráðherranefndin átti frumkvæði að þessu verki og hefur veitt fé til þess, ásamt Málráði Norðurlanda. Vinnan í verkefninu á árinu fólst einkum í því að afla ýmissa tví- eða margmála orðalista til að nýta í margmála leit. Þrír fundir hafa verið haldnir í verkefninu og hefur

Eiríkur Rögnvaldsson sótt þá. Framhald þessa verkefnis, *Tværsproglig søgning på tekster og ordbøger*, fékk svo styrk frá Nordplus Sprog í lok ársins. Þessi verkefni halda því áfram á árinu 2006.

Auk þessa átti Orðabókin aðild að þremur fjölþjóðlegum umsóknum um styrki til tungutækniverkefna á árinu 2005. Þar var í fyrsta lagi um að ræða verkefnið *Nordic Multilingual Technologies*, umsókn til NOS-HS um norrænt öndvegissetur í margmála upplýsingatækni. Forsvarsmaður umsóknarinnar var Koenraad de Smedt prófessor í Bergen. Í öðru lagi var verkefnið *NorPar*, umsókn til NORA, Norrænu Atlantshafsnefndarinnar, um gerð samskipaðrar málheildar (parallel korpus) fyrir dönsku, norsku, sænsku, færeysku og íslensku, með áherslu á tvö síðastnefndu málin. Forsvarsmaður umsóknarinnar var Janne Bondi Johannessen prófessor í Osló. Í þriðja lagi var verkefnið *EuroDocNet*, umsókn í 6. rammaáætlun Evrópusambandsins. Hans Uzkoreit prófessor í Saarbrücken var forsvarsmaður umsóknarinnar, en markmið verkefnisins var að koma á evrópsku samstarfsneti upplýsingasetra í tungutækni, koma upp upplýsingasetrum í Eystrasaltslöndunum, koma upp sameiginlegum gagnagrunni með upplýsingum um tungutækni, þróa aðferðir við margmála leit, þróa og samhæfa íðorðaforða greinarinnar á mörgum tungumálum o.fl.

Engin þessara þriggja umsókna hlaut brautargengi, en þátttaka í þeim var samt mjög gagnleg fyrir Orðabókina. Með henni fékkst dýrmæt reynsla af samningu umsókna af þessu tagi, auk þess sem Orðabókin komst í tengsl við ýmsar erlendar stofnanir og einstaklinga sem starfa á sviði tungutækni. Síðastnefnda umsóknin hefur nú verið endurskoðuð og send aftur til Evrópusambandsins en niðurstöðu er ekki að vænta fyrr en í haust.

Eiríkur Rögnvaldsson