

Orðabókar- og rannsóknarverkefni

Tungutækni-verkefni sem Orðabók Háskólans tekur þátt í

Orðabók Háskólans (nú orðfræðisvið Stofnunar Árna Magnússonar í íslenskum fræðum) er einn þriggja aðila að *Tungutækni*setri, ásamt Málvísindastofnun Háskóla Íslands og verkfræðideild Háskólans í Reykjavík. Setrið sér um vefsetrið www.tungutaekni.is, sem er upplýsingavefur um íslenska tungutækni og var áður rekinn af verkefnisstjórn í tungutækni. Vefurinn var frá upphafi hýstur hjá Skýrr en haustið 2006 var hann settur upp á tölvukerfi Orðabókarinnar. Eiríkur Rögnvaldsson hefur haft umsjón með vefnum.

Tungutækni-setur gekkst fyrir ráðstefnunni *Íslensk tungutækni 2006* hinn 26. maí og voru þar fluttir sjö fyrirlestrar um fjölbreytt tungutækni-verkefni. Ráðstefnan var haldin í Háskólanum í Reykjavík og sóttu hana um 50 manns. Í desember hófst fyrirlestraröð Tungutækni-seturs en stjórn setursins áformar að gangast fyrir mánaðarlegum fyrirlestrum fram á vor. Stefán Briem reið á vaðið með fyrirlestur um vélrænar þýðingar, sem rúmlega 60 manns hlýddu á. Á vegum setursins vann Valdís Ólafsdóttir, M.A. í tungutækni, um mánaðarskeið við uppfærslu og endurbætur á íðorðasafni í tungutækni sem opnað var í Orðabanka Íslenskrar málstöðvar haustið 2005.

Orðabókin og Háskóli Íslands héldu áfram þátttöku 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 verkefnum (verkefnisstjóri Bente Møgegaard). Vinnan í verkefnum á árinu fólst einkum í því að vinna úr ýmsum orðalistum sem aflað hafði verið til að unnt væri að nýta þá í verkinu. Frumgerð af margmála leitarvél fyrir norræn tungumál (og ensku) er nú að finna á vefsíðu Norrænu ráðherranefndarinnar, www.norden.org. Eiríkur Rögnvaldsson tók þátt í verkefnum af Íslands hálfu.

Norræn upplýsingasetur um tungutækni og samtök iðnaðarins á Norðurlöndum sóttu sameiginlega til Norrænu nýsköpunarmiðstöðvarinnar um styrk til tveggja ára verkefnis, *Nordic ICT and Language*, þar sem ætlunin var að fylgja eftir þeirri stefnu sem mörkuð var í forverkefnum *NLTNet* sem sagt var frá í skýrslu síðasta árs. Styrkurinn fékkst ekki. Auk þess var umsókn um verkefnið *EuroDocNet* send inn að nýju í 6. rammaáætlun Evrópusambandsins. Styrkur fékkst ekki heldur til þess verkefnis.

Eiríkur Rögnvaldsson

Íslenskt orðanet

Frá árinu 2004 hafa Jón Hilmar Jónsson og Þórdís Úlfarsdóttir unnið að verkefnum *Íslenskt orðanet*. Verkefnið sækir að stofni til efnivið sinn til orðasambandalýsingarinnar í orðabókunum *Orðastað* og *Orðaheimi* auk orðasambandaskrár Orðabókar Háskólans. Grunnhugmyndin er sú að rekja megi merkingarvensl innan orðaforðans út frá setningarlegum og orðmyndunarlegum venslum á milli orða. Með rækilegri lyklun orðasambanda er byggð upp orðaskrá þar sem einstök orð geta sameinað mikinn fjölda orðasambanda með merkingarlega samstæðu orðafari. Sjálf orðaskráin leiðir auk þess í ljós orðmyndunarleg samkennti sem jafnframt fela í sér merkingarlegan skyldleika. Auk þess er byggt á orðmyndunarlegum venslum í samsettum orðum og merkingarflokkun samsetninga sem rakin er í *Orðastað*. Megináhersla er lögð á samheitavensl en einnig verður gerð grein fyrir lausari merkingarvenslum milli „skyldheita“ auk andheitavensla. Greining efnisins og samtenging orðanna fer fram í skipulegum áföngum þannig

að smátt og smátt hlaðast upp stærri heildir merkingartengdra orða innan einstakra orðflokka.

Tengslin við orðasambandalýsinguna birtast m.a. í því að hin fastari orðasambönd (þ.á m. orðtök) fá sjálfstæða stöðu til jafns við stök orð og verða á þann hátt virk í samheitasamböndum og öðrum merkingarvenslum sem rakin eru innan orðaforðans. Hér gegnir hugtakaflokkun orðasambandanna í *Orðaheimi* víða mikilvægu hlutverki við samtengingu merkingarskyldra orða og orðasambanda.

Meðferð sagna og sagnasambanda er óvenjuleg að því leyti að sagnmyndin birtist hverju sinni með rökliðum sínum svo að af einni sögn geta sprottið fjölmargar sagnafluttur, hver með sínum orðtengslum, setningargerð og merkingareinkennum. Föstum orðasamböndum, svo sem orðtökum, sem innihalda sögn, eru og gerð skil sem flettum. Þannig skilar sagnagreiningin jafnframt mikilvægum áfanga í flokkun og greiningu nafnorða og er í raun forsenda fyrir skipulegri flokkun þeirra. Með staðlaðri framsetningu á sagnafluttunum fæst víða glöggð yfirlit um setningarleg einkenni sagna og sagnarsambanda. Þeirri greiningu er fylgt eftir með málfræðilegri mörkun flettnanna, bæði til yfirlits um einkenni einstakra sagna og til að afmarka og virða fyrir sér tiltekna formgerðir (án þess að hugsað sé til sérstakrar sagnar).

Samheitatenging nafnorða er komin nokkuð á veg með flokkun orðasambanda sem tengjast lýsingarorða- og sagnafluttum. Samræðum nafnorðum sem fram koma í orðastæðum undir þessum tveimur flettutegundum er skipað í merkingarhópa. Með sérstakri forritun eru merkingarhópar með samstæðu innihaldi felldir saman, jafnframt því sem fram kemur hversu oft einstök samheiti koma fyrir í samstæðunni. Þar með er lagður grunnur að því að gera grein fyrir innbyrðis vægi einstakra orða innan samheitaklasa. Þessi aðgerð hefur skilað álitlegum stofni í samheitasafn íslenskra nafnorða sem síðar verður aukið með frekari greiningu.

Við þessa yfirferð hefur komið skýrt fram að markviss samheitatenging og merkingarflokkun nafnorða er háð því að nafnorðafluttunarnar séu merkingarlega einræðar, svo að ekki slái saman óskyldu orðafari innan samheitaklasanna. Í þessu efni þarf víða að ganga nokkru lengra en hefð er fyrir í almennri orðabókarlýsingu.

Íslenskt orðanet hefur að ýmsu leyti sérstöðu meðal sambærilegra merkingarneta. Þar má í fyrsta lagi nefna samhengið við víðtæka og

heildstæða orðasambandalýsingu. Í öðru lagi gegna orðmyndunarleg vensl mikilvægu hlutverki sem tengiliður. Í þriðja lagi spannar orðanetið feikilega víðtækan orðaforða, án tíðni- eða aldursbundinnar takmörkunar. Loks eiga orðasambönd sjálfstæðari aðild að sjálfu netinu en venja er. Jafnframt því að skila fræðilegri greiningu og flokkun á íslenskum orðaforða er orðanetinu ætlað að vera undirstaða nýrra orðabókarverka, einkum samheita- og hugtakaorðabóka.

Forritun og tæknivinna við gerð og uppbyggingu orðanetsins er í höndum Ragnars Hafstað.

Jón Hilmar Jónsson

Ný þýsk-íslensk orðabók

Liðin eru meira en sjötíu ár frá því að Jón Ófeigsson gaf út þýsk-íslensku orðabókina en hún kom út árið 1935 hjá Bókaverslun Sigfúsar Eymundssonar. Þrátt fyrir að ýmsar nýrri orðabækur hafi verið gefnar út frá þeim tíma, hefur bókin verið notuð fram til dagsins í dag vegna þess hversu yfirgripsmikil hún er. Það er hins vegar ljóst að þessi góða bók er fyrir löngu orðin úrelt og þörfin fyrir nýja þýsk-íslenska orðabók orðin knýjandi.

Nú stendur hins vegar til að gefa út nýja þýsk-íslenska orðabók og er það PONS forlagið í Þýskalandi sem vinnur hana í samvinnu við Stofnun Vigdísar Finnbogadóttur í erlendum tungumálum og Stofnun Árna Magnússonar í íslenskum fræðum. Gerð nýju orðabókarinnar er fjármögnuð af tveimur sjóðum í Þýskalandi, Würth-Stiftung í Künzelsau og Robert-Bosch-Stiftung í Stuttgart auk þess sem vonast er eftir framlögum frá íslenskum aðilum. Háskóli Íslands leggur verkefninu til húsnæði og tölvur, en Guðrún Kvaran verður tengiliður Íslands í verkefninu.

Bókin er unnin hjá PONS-forlaginu í Stuttgart og er notaður við verkið þýsk-enskur orðabókagrunnur sem forlagið leggur til. Um er að ræða rúmlega 40.000 flettið auk 25.000 dæmasetninga og orðasambanda. Þá var ákveðið að bæta við 3000 flettið sem tengdust Íslandi með einhverjum hætti, en mikil áhersla er lögð á að bókin nýtist bæði íslenskum lesendum og þýskumælandi fólki sem vill tjá sig á íslensku.

Bókin verður unnin í tölvuforritinu XML-SPY frá austurríska fyrirtækinu Altova, en forritið hefur reynst vel í orðabókavinnslu. Einn helsti styrkur þess er að það leyfir mismunandi meðferð á frum- og markmálinu, en slíkt hjálpar mikið til þegar að umbroti og lokafrágangi kemur.

Að ýmsu verður að gæta þegar menn vinna svona verkefni samkvæmt erlendri fyrirmynd. Þannig verða höfundar að taka mið af fyrirframingefinni ritstjórnarstefnu forlagsins en gæta þess jafnframt að viðhalda þeim stöðlum sem íslenskir notendur orðabóka eru vanir.

Ráðinn hefur verið verkefnastjóri, Heimir Steinarsson, B.A. í þýsku, sem mun ásamt hópi af höfundum og þýðendum annast vinnslu bókarinnar. Tíminn sem hópurinn hefur til að ljúka verkinu eru tvö ár og stefnt er að því að allri höfundar- og ritstjórnarvinnu sé lokið í september 2008 þannig að bókin verði kominn í verslanir um jólin það ár.

Heimir Steinarsson