Orð og tunga

9

Orð og tunga
9

Ritstjóri
Guðrún Kvaran

Ritnefnd

Ásta Svavarsdóttir, Jón Hilmar Jónsson, Veturliði Óskarsson

Stofnun Árna Magnússonar í íslenskum fræðum
Reykjavík 2007

ISSN 1022-4610

Umbrot: Bessi Aðalsteinsson.
Hönnun kápu: Björg Vilhjálmsdóttir.
Prentun og bókband: Leturprent.

Efnisyfirlit

Formáli ritstjóra

Sú nýbreytni var tekin upp frá og með 7. hefti tímaritsins að binda fyrri hluta þess ákveðnu þema. Breytingarnar á tímaritinu hafa mælst vel fyrir og er því þessi árgangur með sama sniði. Samfara breytingunni var ákveðið að fyrri hluti ritsins helgaðist ákveðnu þema á sviði orðfræði eða orðabókarfræði. Þema þessa heftis var rætt á málstofu sem fram fór í safnaðarheimili Neskirkju 17. febrúar 2006 á vegum tímaritsins og Orðabókar Háskólans undir heitinu *Tungutækni og orðabækur*. Þau sem erindi fluttu voru Anna Björk Nikulásdóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson og Sigrún Helgadóttir. Greinar unnar upp úr fyrirlestrunum birtast í þessu hefti.

Auk þemagreinanna eru í ritinu tvær greinar á sviði orðfræði eftir þau Erlu Hallsteinsdóttur og Veturliða Óskarsson. Ítarlegur ritdómur eftir Lars S. Vikør er um *Stóru orðabókina* eftir Jón Hilmar Jónsson sem gefin var út 2005.

Í 8. hefti var tekin upp sú nýbeytni að kynna árlega helstu orðabókarverkefni og rannsóknarverkefni sem unnið er að og tengjast orðabókargerð eða styrkja hana á einhvern hátt. Að þessu sinni eru kynnt þrjú verkefni: Íslenskt orðanet, verkefni á sviði tungutækni og ný þýsk-íslensk orðabók. Umsagnir eru einnig um fjórar nýjar orðabækur.

Allur undirbúningur undir prentun *Orðs og tungu* fór fram á Stofnun Árna Magnússonar í íslenskum fræðum — orðfræðisviði og var það verk eins og áður í höndum Bessa Aðalsteinssonar. Björg Vilhjálmsdóttir grafískur hönnuður, sem hannaði nýtt útlit fyrir 7. árganginn, vann einnig kápuna á þetta hefti.

Guðrún Kvaran

Dr. Jakob Benediktsson

Aldarminning

Í júlímánuði í ár verða hundrað ár liðin frá fæðingu Jakobs Benediktssonar, fyrrum forstöðumanns Orðabókar Háskólans, en hann fæddist á Fjalli í Seyluhreppi í Skagafirði 20. júlí 1907. Jakob var settur til náms og lauk gagnfræðaprófi utanskóla á Akureyri 1922. Veturinn 1923–1924 sat hann í fjórða bekk Menntaskólans í Reykjavík og lauk þaðan stúdentsprófi utanskóla 1926. Eftir það lá leiðin til Kaupmannahafnar til náms í latínu og grísku en cand.mag. prófi í þeim greinum lauk hann við Hafnarháskóla 1932.

Að námi loknu kaus Jakob að dveljast áfram í Kaupmannahöfn þar sem honum bauðst árið 1933 að verða aðstoðarmaður hjá Christian Blinkenberg prófessor við útgáfu á grískum áletrunum frá Lindos. Að útgáfunni vann hann til 1942 en kenndi jafnhliða við ýmsa skóla í Kaupmannahöfn. Áður en útgáfunni með Blinkenberg var að fullu lokið varð Jakob styrkþegi Árnanefndar og vann á vegum hennar við orðabók yfir fornískensku á árunum 1939–1946. Síðustu þrjú árin var hann einnig bókavörður við Háskólabókasafnið í Kaupmannahöfn. Árið 1946 fluttist Jakob til Íslands ásamt konu sinni Grethe Kyhl, mag. art. í klassískri fornleifafræði. Jakobi hafði þá boðist að gerast forstjóri Máls og menningar. Hann tók því boði og gegndi starfinu í tvö ár að hann sneri sér að því sem átti eftir að verða hans ævistarf.

Nokkur umræða hafði á árunum á undan farið fram í heimspekideild Háskóla Íslands um sögulega íslenska orðabók og höfðu menn *Ordbog over det danske sprog* að fyrirmynd sem Verner Dahlerup hafði hafist handa við um aldamótin 1900. Árið 1943 samþykkti heimspekideildin áskorun til háskólaráðs um að láta hefja undirbúning sögu-

legrar íslenskrar orðabókar eftir siðskipti og var Alexander Jóhannesson prófessor aðalhvatamaður þess. Verkið fór hægt af stað. Árni Kristjánsson var lausráðinn til orðtöku 1943 og vann við hana ásamt kennslustörfum. 1947 var Ásgeir Blöndal Magnússon fenginn til verksins og voru þeir Árni nú fastráðnir til að safna efni til sögulegrar íslenskrar orðabókar. Yfirstjórn orðabókarverksins, sem í sátu þrír prófessorar við deildina, töldu mikilvægt að ráða forstöðumann til verksins og leituðu til Jakobs Benediktssonar sem féllst á að taka starfið að sér. Hann var ráðinn til þess 1. janúar 1948 og gegndi því til ársloka 1977 að hann lét af störfum fyrir aldurs sakir. Orðabók Háskólans hafði fest sig í sessi.

Í upphafi lagði Háskóli Íslands út fyrir öllum kostnaði við undirbúning orðabókarstarfsins og síðar fyrstu orðtökuna. Árið 1947 fékkst það samþykkt að háskólinn og ríkið skiptu kostnaði jafnt sín á milli og þannig var staðan þegar Jakob tók við forstöðumennskunni. Orðabókin varð síðar ríkisstofnun 1964 og við það voru fjárveitingar orðnar tryggar þótt ekki væru þær háar.

Mikið verk var fram undan þegar Jakob settist við skrifborðið sitt á efstu hæð í norðurenda aðalbyggingar háskólans. Eftir var að safna orðaforðanum úr heimildum fjögurra alda, skrifa orðabókargreinar og gefa út bók í mörgum bindum, eins og ætlunin var þá, starfsmenn aðeins þrír og ekkert fé til að greiða aðstoðarfólki fyrstu árin. Þeir félagarnir lásu því handrit og bækur, skrifuðu á seðla og röðuðu þeim í kassa. Þannig var unnið árum saman.

Þegar heim kom á daginn settist Jakob við að ljúka doktorsritgerð sinni um Arngrím Jónsson lærða sem hann varði 1957 og kom út í Kaupmannahöfn sama ár undir heitinu *Arngrímur Jónsson and his works*. Á sama hátt vann hann að öllum þeim verkum sem eftir hann liggja, frumsömdum og þýddum, en í þá daga var ekki til það sem nú kallast „frjálsar rannsóknir“ sem hluti vinnutímans. Ég ætla ekki að gera hér að umræðuefni hinar fjölmörgu greinar Jakobs, útgáfur og þýðingar, sem flestir í íslenskum fræðum þekkja, aðeins orðabókarstarfið sem hann tók að sér og vann að í þrjátíu ár.

Ég kom fyrst á Orðabókina haustið 1965 sem stúdent á þriðja ári, nýbúin að ljúka fyrri hluta prófi. Mér hafði boðist þar vinna með námi og kveið ég því heil ósköp að berja að dyrum og spyrja eftir Jakobi. Sá kvíði reyndist ástæðulaus. Við mér tók brosandi maður í reykmettuðu herbergi sínu, sló úr pípunni og tók að spyrjast fyrir um námið.

Vinátta okkar var hafin og stóð hún þar til Jakob lést háaldraður. Eftir drykklanga stund vísaði hann mér í herbergi Ásgeirs Blöndals en þar átti ég að sitja og raða seðlum í bunka eftir upphafsstaf orðanna. Þessi fyrsti vetur minn á orðabókinni líður mér seint úr minni. Á hverjum degi mætti ég til vinnu eftir hádegi og sátum við Ásgeir bæði steinþegjandi við vinnu okkar fram undir vor að hann einn daginn yrti allt í einu á mig og þar eignaðist ég annan náinn vin á Orðabókinni til fjölmargra ára. Klukkan þrjú hringdi Jón Aðalsteinn Jónsson bjöllu og kominn var kaffitími. Hver maður var með sitt neskaffi, vatn var hitað í heldur laslegum katli og hver settist í sitt sæti í örlitla kaffikróknum. Gestir af ganginum komu nær daglega til að rabba, kennarar í íslensku, sagnfræði, lögfræði. Óláir erlendir gestir, sem áttu erindi til landsins, litu inn einkum til að hitta Jakob. Hann var ávallt hrókur alls fagnaðar, tók öllum jafnvel, gerði ekki mannamun.

Fjárhagur stofnunarinnar hafði aðeins rýmkast, unnt var að ráða fólk til þess að skrifa upp úr orðteknum bókum og létti það talsvert á starfsmönnum. Jakob sinnti þessu fólki öllu allt til starfsloka, taldi seðlana, reiknaði út launin, spjallaði og lét það fá nýjar bækur til að vinna með.

Sjálfur valdi Jakob sér mörg tímafrek og erfið verkefni. Hann las og orðtók t.d. Nýja testamenti Odds Gottskálkssonar og bar saman við orðalista í bók Jóns Helgasonar um Nýja testamentisþýðingu Odds. Guðbrandsbiblíu las hann og bar biblíuútgáfu Þorláks Skúlasonar að henni og skráði lesbrigði.

Erfiðasta og tímafrekasta verk Jakobs var þó án efa uppskriftir hans úr orðabókarhandriti Jóns Ólafssonar úr Grunnavík (AM 433 fol.). Það var þá varðveitt í níu bindum í Árnasafni í Kaupmannahöfn en Orðabókin hafði eignast ljósmyndir af handritinu og eftir þeim vann Jakob. Um er að ræða íslenskt orðabókarhandrit með skýringum að mestu á latínu en Jón greip til dönsku og jafnvel íslensku ef latínan ein nægði honum ekki. Handrit Jóns er ekki árennilegt. Hann hafði valið þá leið að skrifa á blöð en ekki seðla og komst fljótt í þrot með rúm innan stafrófsraðarinnar. Þá tók hann að skrifa á spássíur og jafnvel milli lína, hvar sem auðan blett var að finna. Jakob vann það þrekvirki að fara yfir allt handritið, skrá orð og orðskýringar á seðla og raða þeim í stafrófsröð í fjórtán kassa. Alls munu seðlarnir vera um 50.000 en einstök orð eru rúmlega 40.300. Að auki skráði Jakob kvæði, vísur, gátur, tilvísanir í þjóðtrú og tilvitnanir í nafngreinda menn og

er það safn geymt í sérstökum kassa. Þessu verki lauk hann skömmu áður en hann lét af störfum.

Árið 1956 hófst samvinna milli Ríkisútvarpsins og Orðabókar Háskólans um útvarpsþætti. Jón Aðalsteinn Jónsson hafði þá um skeið flutt útvarpsþætti um íslenskt mál en ákveðið var nú að tengja þættina orðabókarstarfinu að hluta. Jakob reið á vaðið og flutti fyrsta þáttinn 6. nóvember og lýsti hugmyndinni á eftirfarandi hátt:

> Þó að þessum þætti sé framar öllu ætlað að vera fræðsluþáttur fyrir almenning, þá skal hinu ekki leynt að við orðabókarmenn væntum okkur af honum nokkurs fróðleiks frá þeim sem á hann hlusta og senda honum bréf. Hvorttveggja er, að spurningar um íslenska tungu geta veitt bæði beina og óbeina fræðslu um ýmsa hluti sem okkur fýsir að vita, og eins er ekki loku fyrir skotið að við munum stundum beina spurningum til hlustenda um atriði sem okkur skortir tilfinnanlega vitneskju um. Mættu svo báðir aðiljar, hlustendur og við, hafa af þættinum nokkurt gagn, og samstarf takast sem báðum gæti komið að notum.

Ekki stóð á liðsinni almennings. Fyrst í stað var efni þáttanna orðfræði og málfarslegar leiðbeiningar auk orðasöfnunar en eins og þeir vita sem fylgdust með þættinum *Íslenskt mál* þróaðist hann smám saman í að verða samvinnuvettvangur Orðabókarinnar og hlustenda ríkisútvarpsins. Ég á ekki von á að Jakob og samstarfsmenn hans hafi árið 1956 grunað að þátturinn yrði á dagskrá í tæpa fimm áratugi.

Jakob var ákaflega ljúfur yfirmaður. Hann var vinnusamur en gaf sér þó alltaf tóm til að spjalla við samstarfsmenn og þá sem til hans leituðu. Þeir voru ófáir stúdentarnir sem leituðu ráða hjá honum og fóru sjaldan burt án þess að fá einhverja úrlausn. Sem sumarstarfsmaður á efstu hæð Árnagarðs minnist ég þess hvernig dagur hans hófst. Hann kom með strætisvagni ofan úr Hlíðum um hálftíu, settist stundarkorn í sófann inni hjá Ásgeiri Blöndal, þar sem mér hafði verið holað niður, sló úr pípunni, tróð í aftur og spjallaði, sagði sögur og hló á meðan hann reykti pípuna til enda. Þá stóð hann upp og hóf dagleg störf. Þessar fáu mínútur eru ljúf minning um góðan yfirmann.

Guðrún Kvaran

Anna Björk Nikulásdóttir

Sjálfvirk greining merkingarvensla í *Íslenskri orðabók*[1]

1 Inngangur

Með tilkomu orðabóka á tölvutæku formi opnast nýir möguleikar í orðabókahönnun. Þróunin í þá átt að hanna rafrænar orðabækur og önnur rafræn uppflettirit á sjálfstæðan hátt, óháð fyrirrennurum sínum á prenti, er þó enn á byrjunarstigi. Þar ræður mestu um að flestar tölvuorðabækur byggjast að einhverju eða öllu leyti á prentaðri útgáfu (Storrer 2001:54). Útgáfa á prenti lýtur öðrum lögmálum og því hefur ekki enn verið hægt að nýta möguleika rafræna formsins nema að hluta. Í bókaútgáfu er pláss veigamikill kostnaðarliður og er því kappkostað að hafa skýringar og aðrar upplýsingar í orðabókum sem knappastar. Skipulagi upplýsinganna eru einnig skorður settar, til að mynda skipa flestar orðabækur flettum í stafrófsröð. Fyrir orðabækur á tölvutæku formi hafa þessi atriði ekki eins mikið vægi. Ekki þarf að hugsa um að spara pláss, flettum og öðrum upplýsingum má raða á ýmsan hátt og gefa notanda kost á mismunandi upplýsingum allt eftir hans þörfum. Hér skiptir mestu að notendaviðmót og leitarmöguleikar séu sem best úr garði gerð.

[1] Þessi rannsókn er hluti af lokaverkefni mínu til MA-prófs í tungutækni frá Ruprecht Karls-Universität Heidelberg, Þýskalandi.

Orð og tunga 9 (2007), 5–24.

Íslensk orðabók (ÍO) var gefin út af Eddu í tölvuútgáfu árið 2000 og í endurbættri útgáfu árið 2003. Um nokkurt skeið hefur hún verið aðgengileg á vefnum (sjá *Vefbækur Eddu*) þar sem hún er í stöðugri þróun. Hvati rannsóknarinnar, sem hér verður lýst, var sá að það væri eftirsóknarvert að taka meira tillit til merkingarlegra þátta við uppsetningu orðabókarinnar en nú er gerlegt. Fyrsta skrefið í þessa átt hefur reyndar þegar verið tekið þar sem við leit að ákveðnu orði birtist listi þeirra flettna sem hafa leitarorðið í skýringu sinni og tengjast því leitarorðinu mjög líklega merkingarlega. Þessi breyting á framsetningu upplýsinga ætti að koma notendum til góða því ástæða þess að fólk flettir upp í orðabók er yfirleitt sú, að leitað er eftir merkingu ákveðins orðs (Herbst og Klotz 2003:33).

Markmið rannsóknarinnar var að kanna möguleikana á því að greina sjálfvirkt merkingarvensl milli nafnorðaflettna og orða úr skýringartextum þeirra. Í þeim tilgangi var forritið MerkOr þróað sem greinir 12 mismunandi merkingarvensl. Niðurstöðurnar sýna á ótvíræðan hátt að sjálfvirk greining merkingarvensla er raunhæfur möguleiki. Stór hluti nafnorðaflettna ÍO hefur verið tengdur við merkingarlega skyld orð, alls 96,45% merkingarliða.

Greining merkingarvenslanna byggist á orðflokkamynstrum skýringanna og því skilar rannsóknin einnig úttekt á því hvaða form skýringartexta eru best fallin til vélrænnar greiningar og hvaða form síður. Þessar upplýsingar gætu komið að notum við endurbætur á ÍO.

2 Aðferð

Lagt var upp með þá tilgátu að unnt sé að greina merkingarvensl milli flettna og orða í skýringum þeirra út frá orðflokkamynstri. Nokkrar rannsóknir af þessu tagi hafa verið gerðar fyrir önnur tungumál og má þar nefna greiningu ensku orðabókarinnar *Longman Dictionary of Contemporary English* (Alshawi 1987), greiningu basknesku orðabókarinnar *Euskal Hiztegia* (Agirre o. fl. 2000) og greiningu yfirheitastigveldis úr þýsku orðabókinni *Wörterbuch der deutschen Gegenwartssprache* (Geyken og Ludwig 2003). Skýringartextar orðabóka henta sérstaklega vel til greiningar eftir föstum mynstrum þar sem form þeirra er í nokkuð föstum skorðum. Algengast er að skýringar nafnorða tilheyri öðrum af eftirfarandi flokkum:

a) samheiti / jafnheiti[2]eða upptalning samheita / jafnheita.

(1) **fagnaður 1** ánægja, gleði[3]

b) yfirheiti (*genus proximum*) ásamt nánari lýsingu, oft lýsingarorði eða -orðum (*differentia specifica*) (Svensén 1993:122).

(2) **breiðband** breitt tíðnisvið notað til fjarskipta, [...]

Einnig er samsetning beggja flokka algeng:

(3) **fagnaður** [...] **2** höfðinglegar móttökur, gleðskapur

Með því að greina orðflokkamynstur skýringanna og samhengi þeirra við merkingarvensl sést til dæmis að í (2) og (3) er höfuð (fyrsta) nafnliðarins yfirheiti flettunnar en hvert orð í upptalningu eins og í (1) er samheiti / jafnheiti hennar. Ýmis önnur mynstur koma fyrir í skýringatextum ÍO þó ekki séu þau eins algeng og þau í (1) – (3) og verður nánar fjallað um nokkur þeirra hér fyrir neðan.

Gögnin, sem rannsóknin byggist á, eru markaðir skýringartextar nafnorða úr gagnagrunni ÍO. Skýringartextarnir voru markaðir með TnT-málfræðimarkara Brants sem þjálfaður hefur verið á íslensku.[4] Stór kostur ÍO er að skýringum er skipt niður í svokallaðar flettutegundir í gagnagrunninum. Flettutegundir eru helsta viðfangsefni rannsóknarinnar og kýs ég að kalla þær skýringarhluta til einföldunar. Eftirfarandi skýring er til að mynda þrískipt:

(4) **dílaburkni** [1] íslensk burknategund [2] *(Dryopteris assimilis)* [3] af þrílaufungsætt, með fjaðurskiptum blöðum, vex í gjám og kjarri

Þetta skipulag auðveldar til muna vélrænan aðgang að einstökum hlutum skýringanna og skilar þar af leiðandi betri niðurstöðum en ef öll skýringin er vistuð sem ein heild. Í greiningu MerkOr er hver skýringarhluti meðhöndlaður sem sjálfstæð eining. Við hann er flettan tengd ásamt kenninúmeri sem og það orðflokkamynstur sem markarinn skilar. Skýringin í (4) hefur eftirfarandi greiningu:

[3]Sjá nánari umfjöllun um samheiti og jafnheiti í köflum 3.2 og 3.8.

[3]Öll orðabókardæmi eru úr vefútgáfu *Íslenskrar orðabókar* (sbr. *Vefbækur Eddu*). Málfræðiupplýsingum (kyni og beygingu) er sleppt í dæmunum.

[4]Gagnagrunnurinn var fenginn til rannsóknarinnar hjá Eddu útgáfu og kann ég þeim bestu þakkir fyrir og þá sérstaklega starfsmönnum orðabókadeildar, þeim Laufeyju Leifsdóttur og Marinó Njálssyni. Kærar þakkir einnig til Sigrúnar Helgadóttur hjá Orðabók Háskólans sem markaði skýringatextana.

(5) íslensk burknategund l_nn[5]
af þrílaufungsætt, með fjaðurskiptum blöðum, vex í gjám og kjarri[6]
a_nþ_,_a_nþ_nþ_,_s_a_nþ_c_nþ

Markarinn skilar mun nákvæmari upplýsingum en hér sjást en einungis er notast við orðflokkamerkinguna, auk fallupplýsinga ef um nafnorð er að ræða. Upphafsorðflokkur hvers mynsturs segir mikið til um það hvaða merkingarvensl er von til þess að greina í viðkomandi skýringu. Mynstrunum er því skipt upp í flokka eftir því. MerkOr er skrifað á hlutbundna forritunarmálinu *Smalltalk* og greiningarklasinn hefur undirklasa fyrir hverja tegund af mynstrum. Greiningin byggist á reglustigveldum, einu fyrir hvern flokk af mynstrum. Í flestum tilvikum nægir að greina orðflokkamynstrið til þess að fá niðurstöðu, sú er raunin í fyrri hluta (5) með mynstrið l_nn. Stundum er þó nauðsynlegt að kanna skýringarstrenginn sjálfan og er seinni hluti (5) dæmi um það. Hér gefur orðið *af* sem upphaf texta með orðflokkamynstrið a_nþ_* vísbendingu um að merkingarvenslin ÆTT sé að finna í skýringunni. Flettan *dílaburkni* hlýtur eftirfarandi greiningu í MerkOr:

(6) YFIRHEITI (dílaburkni, burknategund)[7]
EIGINLEIKI (dílaburkni, íslensk)
ÆTT (dílaburkni, þrílaufungsætt)

Í næsta kafla verður fjallað nánar um merkingarvenslin og greiningu þeirra.

3 Merkingarvensl

Alls innihalda niðurstöðurnar 10 mismunandi merkingarvensl: yfirheiti, undirheiti, samheiti, eiginleiki, ætt, heildheiti, hlutheiti, hluti hóps, tengt lýsingarorð og tengt sagnorð. Auk þess voru jafnheiti og vísanir tekin með en þessi sambönd eru merkt sérstaklega innan orðabókargagnagrunnsins. Ákveðið var að einskorða rannsóknina ekki

[6]Markarinn skilar niðurstöðum í samræmi við greiningarstrengi *Íslenskrar orðtíðnibókar*: a = atviksorð, l = lýsingarorð, nn = nafnorð í nefnifalli, nþ = nafnorð í þágufalli, s = sögn, c = samtenging.

[6]*Dryopteris assimilis* er sleppt. Skýringarhluta á erlendu tungumáli er hægt að tengja við flettuna eftir á án þess að til komi sérstök greining.

[7]Lesist: *burknategund* er yfirheiti *dílaburkna.*

við gildandi merkingarvensl merkingarfræði orða heldur að leggja áherslu á að finna þau sambönd sem eru fyrir hendi í skýringatextum ÍO og gætu komið að notum við uppsetningu hennar.

3.1 Yfirheiti og undirheiti

Yfirheiti er algengustu merkingarvenslin sem er að finna í skýringatextum nafnorðaflettna ef jafnheitin eru ekki meðtalin, alls greindust 43.066 yfirheiti. Y er yfirheiti X ef *A er X* felur í sér *A er Y* en ekki öfugt:

(7) Þetta er fólksbíll og þar af leiðandi er þetta farartæki.
*Þetta er farartæki og þar af leiðandi er þetta fólksbíll.

Fyrir yfirheitasambönd gildir, að ef Y er yfirheiti X þá er X undirheiti Y:

YFIRHEITI (X, Y) → UNDIRHEITI (Y, X)

Þar af leiðir að ef yfirheiti flettu finnst í skýringartexta þá er flettan jafnframt undirheiti þess. Þannig geta orðið til yfirheitastigveldi innan flettna orðabókarinnar (undirstrikuðu orðin eru yfirheiti):

(8) a. **glímukappi** frábær glímumaður
b. **glímumaður 1** keppandi í glímu **2** maður sem iðkar glímu
c. **keppandi** sá sem keppir, tekur þátt í keppni (t.d. í íþróttum)
d. **maður 1** tvífætt og tvíhent spendýr sem talar, [...]
[...]
7 rún sem samsvarar *m*

Í (8c) kemur orðið *maður* ekki fyrir í skýringunni en er engu að síður greint sem yfirheiti. Þetta skýrist af því að skýringar, sem hefjast á orðunum *sá sem...*, vísa til þess að um manneskju sé að ræða og er reglan því sú að greina orðið *maður* sem yfirheiti flettna með þess konar skýringar (jafnvel þó að í einhverjum tilvikum gæti einnig verið vísað til dýra með viðkomandi flettu).

Þegar flettur eru tengdar saman í yfirheitastigveldi eins og í (8) verður að hafa í huga að yfirheitin, sem greind eru fyrir hverja flettu, geta haft margar mismunandi merkingar og í versta falli er um að ræða samhljóma flettur. Að svo stöddu er ekki mögulegt að segja til um hvaða merkingarliður flettu á við það orð sem greint er sem yfirheiti (eða önnur merkingarvensl) og því verður að taka stigveldunum

með fyrirvara. Í (8d) má til dæmis sjá að YFIRHEITI (maður, rún) er gilt en orðið *rún* er ekkert tengt orðunum *glímukappi* eða *keppandi* sem eru neðar í stigveldinu. Þrátt fyrir þetta gæti listi allra orða í slíku stigveldi með tengingu við viðkomandi flettu(r) komið að gagni við leit í orðabókinni. Greining yfirheita gefur einnig möguleika á því að finna öll undirheiti ákveðins orðs:

(9) UNDIRHEITI (stelpa, (dugga, flekon, fruska, gaflhlað, gandála, glofra, hveðra, lausastelpa, skoffín, sleggja, stelpugopi, stelpuskjáta, stelputrippi, strákaflenna, strækni, trilla, trýta, vammaskjóða)

Greining undirheita úr skýringum er ekki algeng. Eins og er skila einungis skýringar af forminu *X(-) eða (-)Y* undirheitum:

(10) **fyrirtak 1** ágæti, úrvalsmaður eða -hlutur
UNDIRHEITI (fyrirtak, úrvalsmaður)
UNDIRHEITI (fyrirtak, úrvalshlutur)

3.2 Samheiti

Samheiti teljast þau orð sem í ákveðnu samhengi er hægt að skipta út hvort fyrir annað án þess að merking eða sannleiksgildi setningarinnar breytist:

(11) a. *Merin* hans Jóns er inni í hesthúsi
b. *Hryssan* hans Jóns er inni í hesthúsi

Algjör samheiti, það er að segja tvö orð sem geta í öllum hugsanlegum tilvikum komið hvort í stað annars, eru sjaldgæf eða ekki til (sjá t.d. Cruse 1986:265ff). Við greiningu samheita úr ÍO er heldur ekki leitað eftir algjörum samheitum í þessum skilningi heldur er gengið út frá því að flettur, sem eru skýrðar með einu orði eða upptalningu orða, hafi í ákveðnu samhengi sömu merkingu og orðið (orðin) í skýringunni, sbr. (1) hér að ofan. Samheiti eru reyndar ekki hátt hlutfall heildargreiningarinnar þar sem skýringar eins og í (1) eru yfirleitt merktar sem jafnheiti í gagnagrunni ÍO og hljóta greiningu samkvæmt því. Eftirfarandi eru dæmi um skýringar þar sem MerkOr greinir samheiti:

(12) a. **rúmfatnaður 1** sængurfatnaður
SAMHEITI (sængurfatnaður, rúmfatnaður)
b. **hormotta** yfirvaraskegg
SAMHEITI (yfirvaraskegg, hormotta)
c. **garg** [...] fuglahljóð
SAMHEITI (fuglahljóð, garg)

Hér sést að niðurstöður samheitagreiningarinnar eru ekki einsleitar. Dæmi (12a) sýnir samheiti sem geta yfirleitt, ef ekki alltaf, komið í stað hvort annars. Í (12b) er merking orðanna á vissan hátt sú sama en á þeim er stór stílmunur þannig að þau er ekki hægt að nota á víxl í hvaða samhengi sem er. Í (12c) er um að ræða yfirheiti og undirheiti: YFIRHEITI (garg, fuglahljóð). Upphaflega tilgátan um að skýringar sem eru eitt nafnorð (eða upptalning nafnorða) væru alltaf samheiti flettunnar stenst því ekki að öllu leyti.

3.3 Eiginleiki, tengt lýsingarorð

Merkingarvenslin eiginleiki og tengt lýsingarorð eru vensl milli nafnorðaflettu og lýsingarorðs úr skýringartextanum. Eiginleiki lýsir á einhvern hátt því sem flettan vísar til og oftar en ekki greinir eiginleiki flettuna frá systrum hennar í yfirheitastigveldi:

(13) a. **busl** ólögulegt sund
b. **sprettsund** stutt sund

Yfirheiti beggja flettna í (13) er *sund* en þær hafa mismunandi eiginleika:

c. EIGINLEIKI (busl, ólögulegt)
EIGINLEIKI (sprettsund, stutt)

Eins og er fylgja lýsingarorðin í niðurstöðunum eftirfarandi nafnorði í kyni og eru því ekki flettur í ÍO nema nafnorðið sé í karlkyni. Í útgáfu ÍO á netinu (sjá *Vefbækur Eddu*) er þó hægt að slá inn leitarorð í hvaða formi sem er þar sem flett er upp í *Beygingarlýsingu íslensks nútímamáls* (BÍN) ef leitarorðið finnst ekki sem fletta. Einnig stendur til að fara yfir niðurstöður MerkOr með hjálp BÍN, bæði til þess að setja orð merkingarvensla í kenniform og til þess að sigta út villur í greiningunni (sjá kafla 4). Hér má einnig benda á að vitanlega eru ekki öll orð, sem fram koma í greiningunni, flettur í ÍO (sjá einnig Kristínu Bjarnadóttur 1998:38–39).

Tengt lýsingarorð er lýsingarorð af sama stofni og flettan eða lýsingarorð með sömu eða svipaða merkingu og slíkt orð. Skýringar, sem innihalda tengt lýsingarorð, eru af forminu *það að vera X* eða *eitthvað X*:

(14) a. **bæklun** það að vera bæklaður, lemstrun
TENGT_LO (bæklun, bæklaður)
b. **dásemd** ágæti, e-ð aðdáanlegt
TENGT_LO (dásemd, aðdáanlegt)

3.4 Ætt

Skýringar orða úr plöntu- eða dýraríkinu hafa sérstöðu að því leyti að þær innihalda oft fræðilegar upplýsingar úr líffræði. Þetta eru upplýsingar sem notandi gæti verið að leita eftir þó að merking flettunnar sé honum ljós (Herbst og Klotz 2003:35). Sem dæmi má nefna að þó notandi viti að *grænlilja* sé plöntutegund áður en hann flettir upp í orðabók gæti vitneskjan um að hún sé af vetrarliljuætt nýst honum:

(15) **grænlilja** íslensk plöntutegund *(Orthilia secunda)* af vetrarliljuætt, [...]
ÆTT (grænlilja, vetrarliljuætt)
YFIRHEITI (grænlilja, plöntutegund)
EIGINLEIKI (grænlilja, íslensk)

Þó segja megi að yfirheitið *plöntutegund* eigi einnig uppruna sinn í líffræði var ákveðið að halda þeirri greiningu í stað þess að setja upp fleiri sérhæfð sambönd fyrir þennan flokk flettna.

3.5 Tengt sagnorð

Nafnorð, sem leidd eru af sögnum (eða öfugt), eru oft skýrð með viðkomandi sögn í ÍO. Einnig nafnorð sem lýsa í raun verknaði án þess að til sé sögn af sama stofni. Þessar skýringar eru af forminu *það að X*:

(16) a. **íhugun** það að íhuga
b. **eftirför** það að elta

Í þessum tilvikum greinir MerkOr merkingarvenslin tengt sagnorð: TENGT_SO (eftirför, elta), TENGT_SO (íhugun, íhuga). Skýringar af þessu tagi innihalda oft orðasambönd:

(17) **útungun** það að unga út eggjum, klak

MerkOr greinir einungis sambönd milli einstakra orða og greinir því ekki tengt sagnorð úr skýringum eins og þeirri í (17) enda væri TENGT_SO (útungun, unga) röng greining.

3.6 Heildheiti

„Heildheiti/holonym er heiti heildar e-s þegar hlutheiti/meronym X er heiti hluta þess. Y er heildheiti X ef X er hluti Y.“ (*Vefbækur Eddu*: Ensk orðabók). Greining heildheita úr ÍO fer eftir mynstrinu *hluti X / hluti af X*. Ef yfirheiti, sem finnst innan skýringartexta, er orðið *hluti* er næsta nafnorð á eftir að öllum líkindum heildheiti flettunnar. Í þessum tilvikum er hætt við að greina yfirheiti en heildheiti skilað í staðinn:

(18) **drag** innsti hluti dals þar sem fjöllin eru lág og aflíðandi
HEILDHEITI (drag, dals)

Skýringar af þessu tagi eru ekki hátt hlutfall skýringartextanna en þó greindust 382 heildheiti. Í einstaka tilvikum var sama heildheitið greint oftar en einu sinni og því hægt að fá lista yfir hlutheiti þess: HLUTHEITI (dals, (drag, framdalur, dalbotn, dalsmynni, uppdalur)). Heildheiti

er meðal þeirra merkingarvensla sem er nauðsynlegt að greina með hjálp BÍN til þess að hægt sé að hafa þau í nefnifalli í niðurstöðunum.

3.7 Hlutheiti

„Hlutheiti er heiti sem nær yfir samsetningarlið orðs, efni eða kjarna e-s, eða að eitthvað tilheyri ákv. hópi. X er hlutheiti Y ef X er hluti Y.“ (*Vefbækur Eddu*: Ensk orðabók). Hlutheiti eiga venjulega við áþreifanlega hluti og hluta þeirra, sígilt dæmi er *fingur* sem hlutheiti orðsins *hönd* (Cruse 1986:160–161):

(19) Fingur er hluti handar
Hönd er með fingur

MerkOr greinir hlutheiti í skýringum sem hafa mynstur sem samsvara reglulegu segðinni n_(,_n_)*c_n.* þar sem samtengingin (c) er *og*:

(20) **þvagfæri** nýru og þvaggangur
HLUTHEITI (þvagfæri, (nýru, þvaggangur))

Hlutheitin í niðurstöðunum eru þó af mismunandi gerð og í mörgum tilfellum er ekki um að ræða eiginleg hlutheiti heldur tengd sambönd (sjá Cruse 1986:172ff). Prófunarsetningin *X er hluti Y* gildir til að mynda ekki fyrir öll samböndin. Það sem þau eiga sameiginlegt er að upptalning þeirra gefur nokkuð góða mynd af merkingu flettunnar. Eftirfarandi eru nokkur dæmi um hlutheiti og hvernig þau skiptast í flokka eftir því hvaða prófunarsetning gildir:

(21) a. X er hluti Y: HLUTHEITI (Eyjaálfa, (Ástralía, Kyrrahafseyjar, Nýja-Sjáland))
b. X og Z mynda Y: HLUTHEITI (brúðhjón, (brúðgumi, brúður))
c. Það að vera X og Z felst í því að vera Y: HLUTHEITI (trúbador, (skáld, tónlistarmaður))

Sérstakur flokkur hlutheita er sá sem lýsir meðlimum hóps. Fyrir þessi vensl gildir að myndi hópur af X Y, þá er X hluti hóps Y. Þessi gerð greindist einungis 87 sinnum í ÍO en mynstrið, sem gildir fyrir þessi vensl, er *hópur (af) X*:

(22) a. **holl** hópur manna, hluti heildar
HLUTI_HÓPS (holl, manna)
b. **grind** [...] **5** hópur af hvölum
HLUTI_HÓPS (grind, hvölum)
c. **gæslusveit** hópur eftirlitsmanna við tiltekið verkefni, einkum friðargæslu
HLUTI_HÓPS (gæslusveit, eftirlitsmanna)

Síðan kemur til kasta BÍN til þess að setja orðin *manna, hvölum, eftirlitsmanna* í nefnifall eintölu. Þá lítur (22a) svona út: HLUTI_HÓPS (holl, maður).

3.8 Jafnheiti og vísanir

Jafnheiti og vísanir eru sérstaklega merkt innan gagnagrunns ÍO og MerkOr framkvæmir enga frekari greiningu á þeim skýringum nema skipta þurfi upptalningum upp í einstök orð.

Hugtakið *jafnheiti* er reyndar ekki viðeigandi í þessu samhengi. Þetta hugtak er notað um „orð með samsvarandi merkingu“ (Jón Hilmar Jónsson 2005:26) í tvímála orðabókum. Hafa ber því í huga að jafnheiti í niðurstöðum MerkOr eru merkt sem slík í gagnagrunni

ÍO en lýsa ekki eiginlegum merkingarvenslum. Jafnheitaflokkun ÍO er notandanum ósýnileg og ekki er alltaf hægt að útskýra mun á flokkun skýringa:

(23) **leppur 1** tuska [merkt í gagnagrunni sem skýring]
[...]
4 handbendi [merkt í gagnagrunni sem jafnheiti]

Við yfirferð niðurstaðna MerkOr komu í ljós nokkrir flokkar jafnheita. Þetta er ekki tæmandi listi þar sem dæmi voru valin af handahófi:

(24) a. Jafnheitið er nánast sama orð og flettan:
afrak afhrak; **snjár** snjór; **ábreiðsla** ábreiða; **sauðarlús** sauðalús
b. Jafnheitið er samheiti flettunnar:
hettuselur blöðruselur; **aflgjafi** orkugjafi; **vefdagbók** bloggsíða
c. Jafnheiti hefur samsvarandi merkingu og flettan en háð samhengi. Jafnheitin í skýringunni eru ekki endilega samheiti innbyrðis:
krakki barn, barnungi, krógi
blástur kaldi, stinningsgola, vindur, þurrkur
d. Jafnheiti er undirheiti flettunnar:
sjómaður farmaður, fiskimaður
feiti viðbit

Einnig koma fyrir skýringar þar sem jafnheitin eru af mismunandi gerð. Í eftirfarandi dæmi eru samheiti, yfirheiti og undirheiti í sömu upptalningunni:

(25) **dalur** dollari, peningur, ríkisdalur

Í öllum tilfellum ætti þó tenging við viðkomandi jafnheiti sem flettur (ef þær eru fyrir hendi í ÍO) að koma notandanum að gagni en hann verður að meta að hvaða leyti merking þeirra er samsvarandi því sem hann er að leita að.

Töluvert algengt er að lengri skýringartextar séu merktir sem jafnheiti. Þessi tilfelli eru ekki greind í þessari fyrstu útgáfu MerkOr en þau þyrfti að taka til athugunar, sem og hvort ekki væri ástæða til að merkja þessa skýringarhluta öðruvísi í gagnagrunninum. Eftirfarandi er dæmi um lengri skýringu sem er merkt sem jafnheiti:

(26) **keppni** [...] **3** skipuleg samkoma eða mót þar sem menn reyna með sér í íþróttum eða öðru kappi

Vísanir eru merktar fyrir notanda og benda honum á að gagnlegt gæti verið að fletta upp á viðkomandi orði. Hér eru eftirfarandi merkingarvensl algeng:

(27) a. undirheiti
gafl fótagafl, höfðagafl
b. yfirheiti
bandprjónn prjónn
c. andheiti
vörn sókn
d. hlutheiti
brot nefnari, teljari

Einnig sambönd eins og lýst er í (24a):

(28) a. **pepilrás** pöpulrás
b. **fögnuður** fagnaður

Vísanir eru því tengdar flettunni á ólíkan hátt og ekki er alltaf ljóst af hverju eitt orð er skráð sem vísun en annað ekki. Af hverju er til dæmis *pöpulrás* í (27a) vísun en *snjár* í (23a) jafnheiti? Það að ákveðið ósamræmi sé til staðar skýrist væntanlega að einhverju leyti af því að efni ÍO er frá mismunandi tímum. Hluti skýringanna er óbreyttur síðan í upphaflegri útgáfu ÍO frá árinu 1963, aðrar hafa verið endurskoðaðar fyrir síðari útgáfur verksins og enn öðrum bætt við. Spurning er hvort vísanir í því formi sem þær birtast í prentaðri útgáfu eigi alltaf erindi í tölvuútgáfu orðabókar. Ef núverandi uppsetning ÍO á vefnum er tekin sem dæmi, má sjá að til vinstri við skýringartextana birtist listi af tenglum sem allir vísa í flettu innan orðabókarinnar og þetta er því eins konar listi með vísunum.

4 Niðurstöður

Reynslan af þessari fyrstu útgáfu af MerkOr sýnir að hægt er að ná góðum árangri í sjálfvirkri greiningu merkingarvensla. Með frekari fínstillingu reglnanna má reikna með að árangurinn geti batnað enn frekar. Þær niðurstöður sem þegar hafa fengist mætti byrja að nýta til þess að prófa áfram notendaviðmót og leitarmöguleika ÍO á netinu þó

ljóst sé að nokkur fjöldi orða sé tengdur merkingarvenslum við röng orð.

	alls	heimt
merkingarliðir	77.348	96,45%
skýringarhlutar	106.972	92,61%

Tafla 1: Heimt MerkOr úr skýringartextum nafnorðafletta *Íslenskrar orðabókar*

Þar sem hver merkingarliður flettu er greindur sjálfstætt er fjöldi merkingarliða notaður sem viðmiðun en ekki fjöldi flettna. Hver merkingarliður getur svo samanstaðið af tveimur eða fleiri skýringarhlutum en hver skýringarhluti myndar eigið orðflokkamynstur sem er grundvöllurinn fyrir greiningu merkingarvensla. Eins og sjá má í töflu 1 hafa fundist merkingarvensl við 96,45% allra merkingarliða en það þýðir að við stærsta hluta merkingarliða nafnorðaflettna er hægt að setja vísun í orð sem stendur í merkingarvenslum við flettuna. Þessi tala er mikilvæg fyrir hönnuði og höfunda orðabókarinnar til þess að sjá hve algengt það er að fletta hafi vísun í merkingarlega tengt orð.

Hin talan, hlutfall greindra mynstra, sýnir hve hátt hlutfall orðflokkamynstra hlaut greiningu. Með því að skoða þau mynstur sem MerkOr hafnar mætti annars vegar bæta hönnun tólsins og hins vegar benda á aðra möguleika til þess að orða skýringar. Skýringarnar verða þó alltaf skrifaðar með notendur efst í huga og því er ekki við því að búast að hægt verði að greina allt vélrænt.

	fjöldi merkingarliða	hlutfall
rétt greining	786	82,13%
ófullnægjandi greining	121	12,64%
röng greining	50	5,22%

Tafla 2: Niðurstöður prófunar (prófunarsett: 1034 merkingarliðir sem greindir voru handvirkt)

Prófunarsettið er tilviljunarúrtak heildargagnanna, um 1,34%. Þessir 1034 merkingarliðir voru greindir handvirkt og síðan keyrðir saman við niðurstöður úr MerkOr. Ef merkingarliður telst rétt greindur þýðir það að MerkOr fann öll merkingarvensl sem greind voru handvirkt

og ekki fleiri. Þegar MerkOr finnur eingungis hluta þeirra merkingarvensla sem greind voru handvirkt en greinir þó ekkert rangt er um ófullnægjandi greiningu að ræða. Samanlagt voru því 94,77% allra greindra merkingarliða villulausir. Rangt greindur er hver sá merkingarliður þar sem MerkOr greinir að minnsta kosti eitt rangt orð eða röng merkingarvensl, jafnvel þó hluti greiningarinnar sé réttur. Þessi aðferð prófunar var valin þar sem mikilvægt er fyrir höfunda ÍO að vita hve hátt hlutfall merkingarliða gætu haft rangt greind merkingarvensl til þess að meta hvort bæta eigi niðurstöðunum hráum inn í ÍO. Mögulegt væri einnig að gera villutölfræðina eftir flettum eða mynstrum en þær tölur gæfu ekki nógu nákvæma mynd af því hve margir merkingarliðir væru rangt greindir.

4.1 Helstu vandamál

Algengasta ástæða rangt greindra merkingarvensla eru villur í greiningu markarans. Nákvæmni markarans var ekki prófuð enda leiða ekki allar villur markarans til rangrar greiningar merkingarvensla og þar af leiðandi er ekki beint samband þar á milli. Í sumum tilvikum voru reglur MerkOr einnig lagaðar að villum markarans til þess að ná betri árangri. Eins og Þórdís Úlfarsdóttir bendir á (2006:141) er í „þessum stuttu textum [orðabókarskýringanna] lítið setningarlegt samhengi fyrir markarann að styðjast við“[8]. Árangur markarans er engu að síður ágætur eins og sjá má af því að villur, sem leiða til rangt greindra merkingarvensla, eru aðeins um 2% af prófunarsettinu.

Lýsingarorð eru oft greind sem nafnorð og MerkOr greinir þau þar af leiðandi sem yfirheiti:

(29) **bylgusa** skammvinn hríð, snjógangur [...]

Hér er greining markarans sú að *skammvinn* sé nafnorð og niðurstaða MerkOr var YFIRHEITI (bylgusa, skammvinn) í staðinn fyrir YFIRHEITI (bylgusa, hríð), EIGINLEIKI (bylgusa, skammvinn).

Önnur tegund af villum er þegar vensl eru út af fyrir sig rétt greind en gefa samt sem áður engar nothæfar upplýsingar. MerkOr útilokar nokkur slík orð við greininguna en lista yfir þessi orð þarf að lengja. Dæmi um slík orð úr niðurstöðum prófunarinnar:

[8] Þórdís Úlfarsdóttir fjallar í grein sinni um mörkun orðasambanda og tekur einnig dæmi um mörkun orðabókartexta.

(30) a. **bulla** svipuð stöng í vél eða dælu
EIGINLEIKI (bulla, svipuð)
b. **spaði** þannig lagað kvenhöfuðfat
EIGINLEIKI (spaði, lagað)

5 Form skýringartexta

Í sambandi við form skýringatexta eru tveir meginþættir sem þarf að athuga: Hvaða form eru vel eða illa til þess fallin að greina þau vélrænt og hvaða form eru algengust í nafnorðaskýringum ÍO. Nákvæmri greiningu er ekki lokið en hér verður bent á nokkur dæmi. Hafa ber í huga að tölurnar eru byggðar á niðurstöðum markarans og verður því að lesa þær með þeim fyrirvara að greining hans er ekki alltaf rétt. Eins og sést í töflu 1 hér að ofan er heildartala skýringarhluta 106.977. Flestir þeirra hefjast á nafnorði, alls 73.290 og þar af eru 35.158 einungis eitt nafnorð. Upptalning á nafnorðum eru 6.913 skýringarhlutar og 26.085 skýringarhlutar passa við mynstrið nafnorð_komma_frekari skýring. Önnur mynstur eru til dæmis nafnorð_atviksorð_nafnorð og nafnorð_eignarfallseinkunn. Í skýringarhlutum, sem hafa mynstur sem hefst á nafnorði, er yfirleitt að finna samheiti eða yfirheiti og eru þær því vel fallnar til greiningar merkingarvensla. Sem stendur er litið fram hjá því sem stendur í svigum og því eru skýringar eins og eftirfarandi ekki greindar:

(31) **handaburður** handa- (og handleggja)hreyfing(ar)

Önnur form skýringarhluta, sem skila áreiðanlegum niðurstöðum, eru form sem hefjast á lýsingarorði. Alls eru slík mynstur í nafnorðaskýringum ÍO 14.684 talsins, þar af kemur mynstrið lýsingarorð_nafnorð 5.672 sinnum fyrir. Þessi mynstur skila yfirleitt merkingarvenslunum eiginleika og yfirheiti. Að síðustu má nefna mynstur sem hefjast á fornafni en þau eru 7.362 talsins. Úr þessum mynstrum eru það fyrst og fremst merkingarvenslin tengd sögn (*það að X*) og orðið *maður* sem yfirheiti (*sá / sú sem...*) sem koma til greina.

Auk skýringarhluta eins og í (17) *það að unga út eggjum*, eru ýmiss konar form skýringa sem ekki er hægt að greina með tilliti til merkingarvensla:

(32) a. **horgemlingur** einnig um horað fólk
b. **umritun** um að venjulegum orðum er skipt út fyrir ný orð [...]

Þetta þýðir þó ekki endilega að engin merkingarvensl finnist fyrir viðkomandi merkingarlið flettu. Til dæmis skila aðrir skýringarhlutar viðkomandi merkingarliðs fyrir flettuna *umritun* (*það að umrita* og *umorðun*) merkingarvenslum TENGT_SO (umritun, umrita) og JAFNHEITI (umritun, umorðun).

Notkun sviga í skýringatextum skapar oft vandamál í vélrænni greiningu. Svigar eru oft notaðir til þess að draga fram atriði sem skýra merkingu orðsins betur eða til þess að benda á atriði sem yfirleitt (en ekki endilega alltaf) tilheyra merkingu orðsins. Herbst og Klotz benda á að innihald sviga skýri betur viðkomandi „Prototyp“ (2003:36). Eins og er er það sem stendur í svigum ekki tekið með í greiningu MerkOr og þær skýringar sem hefjast á sviga eru í heild sinni látnar eiga sig. Þetta verður þó skoðað betur í áframhaldandi þróun tólsins.

Í skýringum ÍO er að finna „margvíslegt misræmi“ (Mörður Árnason 1998:4) sem skýrist, eins og áður sagði, af sögu bókarinnar og af því hvernig hún hefur verið unnin. Notkun sviga er eitt af því sem þyrfti að vera skýrt skilgreint við endurskoðun skýringatexta ef auðvelda á vélræna greiningu. Í inngangi þriðju útgáfu *Íslenskrar orðabókar* eru skýringar um notkun tákna og greinimerkja. Þar segir: „Svigar eru notaðir í dæmum um uppflettiorð þannig að það sem er innan þeirra getur komið í stað þess sem stendur á undan (einnig notaðir á hefðbundinn hátt í skýringartextum, þ.e. um upplýsingar sem eru til nánari útskýringar á einhvern hátt).“ (*Íslensk orðabók* 2002: xiii). Dæmi um sviganotkun, sem hefur áhrif á núverandi greiningu MerkOr, eru lýsingarorð í sviga í upphafi skýringarhluta. Í eftirfarandi dæmum eru skýringarhlutar með lýsingarorðinu *lítill* skoðaðir. Það liggur í eðli lýsingarorða að þau eru til nánari skýringar og því fellur það hlutverk svigans niður. Annað hlutverk svigans, að innihald hans geti komið í stað orða utan svigans, virðist ekki vera algengt.

(33) a. **smástrákur** (lítill) ungur drengur
b. **stráklingur** lítill, ungur drengur
c. **prammi** (lítill eða) stór flatbotna bátur [...]
d. **sproti** (lítill eða allvænn) silungur
e. **sprotasilungur** (lítill) silungur
f. **branda** (lítill) fiskur
g. **skítseiði** lítill fiskur

Í (33a) gæti *smástrákur* þýtt til dæmis a. *lítill og ungur drengur*, b. *lítill eða ungur drengur* (orð í sviga kemur í stað orðs utan hans) eða c. *ungur drengur, yfirleitt lítill.* Til samanburðar má líta á skýringuna við *stráklingur* og leiða að því getur að sviginn í (33a) sé í raun óþarfur. Í (33c) virðist lýsingarorðunum vera ofaukið því væntanlega getur *prammi* líka lýst bát sem hvorki er stór né lítill, *flatbotna* er hér lýsingarorðið sem skiptir máli. Sama á við í (33d) en hér er sviginn settur utan um bæði lýsingarorðin. Í greiningu merkingarvensla getur komið til mótsagna þegar skýringar eru orðaðar á þennan hátt: EIGINLEIKI (prammi, lítill), EIGINLEIKI (prammi, stór). Eins og áður segir er þessum skýringum sleppt í greiningunni eins og er. Annað dæmi um sviganotkun er þegar yfirheiti og undirheiti er í raun skeytt saman: *(ára)bátur, (vín)sopi* eða orðum eins og *breyting – tilbreyting: (til)breyting, skipti – umskipti (um)skipti*. Þetta skapar ákveðin vandamál við vélræna greiningu þar sem textanum er skipt upp í tóka fyrir mörkun. Þá stendur eitt orð eða eitt greinamerki í hverri línu. Fyrir orðið *(til)breyting* lítur það svona út:

(
til
)
breyting

Ekki er lengur hægt að sjá hvort orð í sviga tilheyra eftirfarandi orði (eins og í dæmunum hér á undan), undanfarandi orði eða eru sjálfstæð orð. Mögulega er hægt að skrifa reglur sem leysa úr þessu en það hefur ekki verið kannað til hlítar.

Þó hér hafi verið bent á dæmi um form skýringartexta, sem skapa vandamál við vélræna greiningu, er rétt að benda aftur á að 92,61% skýringarhlutanna voru greindir af MerkOr (sjá töflu 1).

6 Lokaorð

Í þessari grein var lýst tilraun til sjálfvirkrar greiningar merkingarvensla í *Íslenskri orðabók*. Greiningartólið MerkOr, sem þróað var í þessum tilgangi, greindi merkingarvensl milli nafnorðaflettna og orða úr skýringartextum þeirra. Alls voru greind merkingarvensl í 74.633 merkingarliðum sem jafngildir 96,45% heildargagnanna. Nákvæmni í prófunarsetti var 82,13% ef einungis eru taldir merkingarliðir sem hlutu fullnægjandi greiningu en 94,77% ef allir merkingarliðir eru meðtaldir sem voru greindir villulaust. Þessar tölur benda til þess að með áframhaldandi þróun aðferðarinnar væri hægt að ná mjög góðum árangri.

Tenging á milli flettna og orða, sem standa í merkingarvenslum við þær, býður upp á margvíslega möguleika við uppsetningu orðabóka á tölvutæku formi. Einnig verður notagildi orðabókarinnar margþættara og má nefna sem dæmi að með hjálp niðurstaðna MerkOr er hægt að fá lista yfir allar löggildar iðngreinar sem er lýst í ÍO.

Stefnan í áframhaldandi þróun MerkOr er þríþætt: a. að kanna forsetningarliði í skýringartextum með það í huga að greina fleiri merkingarvensl, b. að kanna möguleikana á greiningu merkingarvensla fyrir sagnorðaflettur og lýsingarorðaflettur og c. að kanna hvort ekki sé hægt með lítilli fyrirhöfn að greina aðrar orðabækur og uppflettirit með aðstoð MerkOr.

Heimildir

Agirre, Eneko o. fl. 2000. *Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar*.
http://arxiv.org/abs/cs.CL/0010025. sótt: 21.01.2005

Alshawi, Hiyan. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *Computational Linguistics* Vol. 13, Nr. 3–4: 195–202.

Beygingarlýsing íslensks nútímamáls. Á vefsíðu Orðabókar Háskólans:
http://www.lexis.hi.is/beygingarlysing/

BÍN = *Beygingarlýsing íslensks nútímamáls.*

Cruse, Alan. 1986. *Lexical Semantics.* Cambridge: Cambridge University Press.

Geyken, Alexander og Rainer Ludwig. 2003. *Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache*.
www.dwds.de/help/pages/ExtrHyp.pdf. sótt: 28.05.2005

Herbst, Thomas og Michael Klotz. 2003. *Lexikografie.* Paderborn: Ferdinand Schöningh.

ÍO = *Íslensk orðabók*.

Íslensk orðabók. 2002 (3. útgáfa). Ritstjóri: Mörður Árnason. Reykjavík: Edda.
Jón Hilmar Jónsson. 2005. Aðgangur og efnisskipan í íslensk-erlendum orðabókum — vandi og valkostir. *Orð og tunga* 7: 21–40.
Kristín Bjarnadóttir. 1998. Um skýringarorðaforðann. *Orð og tunga* 4: 33–43.
Mörður Árnason. 1998. Endurútgáfa „Íslenskrar orðabókar". Stefna — staða — horfur. *Orð og tunga* 4: 1–8.
Storrer, Angelika. 2001. Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. Í: Ingrid Lemberg, Berhard Schröder og Angelika Storrer (ritstj.). *Chancen und Perspektiven computergestützter Lexikographie,* bls. 53–69. Tübingen: Max Niemeyer.
Svensén, Bo. 1993. *Practical Lexicography. Principles and Methods of Dictionary-Making.* Oxford/New York: Oxford University Press.
Vefbækur Eddu: Íslensk orðabók; Ensk orðabók: edda.is/vefbaekur. sótt: 31.08.2006
Þórdís Úlfarsdóttir. 2006. Málfræðileg mörkun orðasambanda. *Orð og tunga* 8: 117–144.

Lykilorð

merkingarfræði orða, merkingarvensl, orðabókaskýringar, orðflokkamynstur

Keywords

lexical semantics, semantic relations, dictionary definitions, POS-patterns

Abstract

In the design of electronic dictionaries it is possible to organize the information due to the meaning of the lexems. In this article a method for automatic extraction of semantic relations from dictionary definitions is demonstrated. Definitions of all noun lexems of a monolingual Icelandic dictionary, *Íslensk orðabók*, were analyzed. First, the definitions were tagged with Brants TnT-tagger which had been trained on an Icelandic corpus. From the tagged data the POS-patterns of the definitions were extracted and rules for extracting the semantic relations were developed. The rule algorithm was implemented in Smalltalk, resulting in the tool MerkOr. The results of the analyzis were promising. The test was made with a random set of lexemes, about 1,34% of the data. For each lexeme the result could be completely right, that is all semantic relations from the reference data was found in the analyzis of MerkOr, or it could be partly right, that is MerkOr did not find every relation compared to the reference data, but did nevertheless not extract any wrong word or relation. The precision was 82,13% (completely right analyzed) up to 94,77% (completely right plus partly right).

Anna Björk Nikulásdóttir
Seminar für Computerlinguistik
Ruprecht Karls-Universität Heidelberg
anna.b.nik@gmx.de

Ásta Svavarsdóttir

Talmál og málheildir — talmál og orðabækur

1 Inngangur

Tilkoma rafrænna texta og textasafna hefur hleypt nýju lífi í rannsóknir á ýmsum sviðum málvísinda og gert mögulegt að kanna þætti í máli og málnotkun sem illmögulegt er að rannsaka án aðgangs að slíkum söfnum. Innan *gagnamálfræði* (e. corpus linguistics) er m.a. fengist við tíðnirannsóknir á orðum, orðmyndum, orðasamböndum, setningagerðum o.fl. og niðurstöðurnar hafa orðið grundvöllur að lýsingu og samanburði á textum og textategundum (sjá t.d. Biber, Conrad & Reppen 1998:1 o.áfr., Teubert 2004:96 o.áfr.). Einnig hafa rafræn textasöfn valdið róttækum breytingum á sviði orðfræði og orðabókafræði og slík söfn eru orðin ein mikilvægasta undirstaðan undir gerð orðabóka. Þar nýtast þau bæði við efnissöfnun og efnisval og við greiningu á merkingu og notkun orða og orðasambanda (sjá t.d. Landau 2001:296 o.áfr.). Á undanförnum áratugum hafa verið gefnar út orðabækur sem byggja einungis á slíkum söfnum, s.s. ensku COBUILD-orðabækurnar (Sinclair 1987).

Rafræn textasöfn eru nú til fyrir fjöldamörg tungumál. Flest eru þau samtímaleg og geyma einkum texta úr nútímamáli. Einnig hafa verið sett saman söfn með eldri textum, bæði söfn sem ná yfir texta frá ýmsum tímum og sérhæfð textasöfn fyrir ákveðin tímabil. Nærtækt dæmi um það síðartalda eru *Íslendinga sögur* (1996) sem gefnar hafa

Orð og tunga 9 (2007), 25–50.

verið út á rafrænu formi, bæði textinn í heild (útgáfa Svarts á hvítu) og orðstöðulykill sem unninn er upp úr honum. Eiríkur Rögnvaldsson (1990, 1996 o.v.) hefur fjallað um gerð hans og nýtingu, þ.á m. við orðabókagerð, og hann hefur nýtt þennan efnivið ásamt fleiri textum í rafrænu formi til ýmissa rannsókna (sjá t.d. Eiríkur Rögnvaldsson 1994–5, 2002).

Í greininni verður fjallað um gerð stórra rafrænna málsafna sem ætluð eru til margvíslegra verkefna, einkum hagnýtra og fræðilegra rannsókna í málvísindum og tungutækni. Mest er fjallað um samsetningu slíkra safna, einkum val á textum. Sérstök áhersla er lögð á hlut óútgefins efnis, bæði óformlegra ritmálstexta af ýmsu tagi og efnis úr talmáli. Rætt er um gildi þess að málsöfn rúmi slíkt efni ekki síður en fjölbreytilega ritmálstexta og jafnframt er gerð grein fyrir ýmsu sem greinir söfnun, úrvinnslu og frágang talmálsefnis frá efnisöflun úr ritmáli. Um þessi atriði fjalla annar og þriðji kafli greinarinnar, sá fyrri um svonefndar *málheildir* (e. corpus) almennt, samsetningu þeirra og notagildi, en hinn síðari um hlut talmálsins í slíkum söfnum og um öflun og úrvinnslu talmálsefnis. Í fjórða kafla er sjónum svo beint að orðabókum, hlutverki talmálsefnis við gerð þeirra og áhrifum sem slíkt efni getur haft á orðlýsinguna.

Greinin á rætur að rekja til aðildar höfundar að tveimur stórum verkefnum sem nú er unnið að. Annað þeirra er *Mörkuð íslensk málheild* (MÍM) [1] sem er í smíðum við Orðabók Háskólans (nú orðfræðisvið Stofnunar Árna Magnússonar í íslenskum fræðum). Hitt er rannsóknarverkefnið *Tilbrigði í setningagerð*[2] en einn þáttur þess beinist að því að draga saman og ganga frá textum úr talmáli til notkunar í rannsókninni. Þótt verkefnin séu ólík fela þau bæði í sér viðamikla efnissöfnun og samvinna hefur tekist milli þeirra um samnýtingu á gögnum og verkaskiptingu við öflun þeirra og úrvinnslu til hagsbóta fyrir bæði verkefnin. Stór hluti talmálsefnisins er reyndar fenginn úr eldri verkefnum og sumarið 2006 var unnið að söfnun óformlegra rit-

[1] Verkefnið er unnið fyrir styrk úr tungutækniáætlun Menntamálaráðuneytisins undir stjórn Sigrúnar Helgadóttur (sjá nánar: http://www.lexis.hi.is/malheild.htm).

[2] Verkefnisstjóri er Höskuldar Þráinsson. Að verkefninu stendur hópur málfræðinga við Háskóla Íslands, Stofnun Árna Magnússonar í íslenskum fræðum (áður Orðabók Háskólans) og Kennaraháskóla Íslands. Verkefnið tengist stærra norrænu verkefni, *ScanDiaSyn* (Scandinavian Dialect Syntax; sjá: http://uit.no/scandiasyn/scandiasyn/). Íslenska rannsóknin nýtur öndvegisstyrks frá Rannís 2005–2007.

málstexta í verkefninu *Blogg og bréfaskriftir. Óútgefin skrif Íslendinga*[3]. Síðar er ætlunin að fella þann efnivið inn í MÍM og hann nýtist líka ágætlega í tilbrigðarannsókninni. Samvinnan teygir því anga sína í ýmsar áttir og það er m.a. tilgangur minn með greininni að hvetja sem flesta til samstarfs um efnisöflun og gagnavinnslu.

2 Textasöfn og málheildir

Textasafn er hér notað um hvers kyns söfn rafrænna texta en hugtakið *málheild* (d. korpus, e. corpus) er aftur á móti haft um textasöfn með völdum textum sem hafa verið greindir málfræðilega auk þess að gengið er frá textunum á ákveðinn hátt (sbr. Sigrún Helgadóttir 2004a:67). Biber, Conrad & Rippen lýsa muninum á textasafni og málheild þannig: „A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language“ (1998:246). Til að tiltekið safn rafrænna texta geti kallast málheild þarf það m.ö.o. að uppfylla ákveðin skilyrði og eru þau helstu talin upp í (1).

(1)
1. Textasafnið þarf að vera skipulega samansett miðað við fyrirfram ákveðnar forsendur sem miða að því að safnið geti talist endurspegla málið eða afmarkaðan hluta þess á tilteknum tíma eða tímabili.
2. Textarnir þurfa að vera greindir m.t.t. ýmissa málfræðilegra atriða og markaðir á skipulegan hátt.
3. Upplýsingar um uppruna og eðli textanna þurfa að liggja fyrir og vera skráðar á skipulegan hátt.
4. Umbúnaður safnsins þarf að vera þannig að auðvelt sé að leita í því og nýta það til margvíslegra fræðilegra og hagnýtra verkefna.
5. Upplýsingar um gerð og samsetningu safnsins þurfa að vera aðgengilegar.

Með hugtakinu *markaður* (e. tagged) í lið (1)2 er átt við það að hvert einstakt lesmálsorð hafi verið greint m.t.t. tiltekinna málfræðiatriða, oftast orðflokks og beygingarmyndar, og tengt við ákveðið uppflettiorð (lemmu eða les). Með því móti er gerður greinarmunur á tví- eða

[3]Eyrún Valsdóttir vann að þessu verkefni sumarið 2006 undir stjórn Sigrúnar Helgadóttur og Ástu Svavarsdóttur, m.a. fyrir styrk úr Nýsköpunarsjóði námsmanna.

margræðum orðmyndum þannig að t.d. sé ljóst hvort *langar* sé birtingarmynd so. *langa* eða lo. *langur* og hvort *vinna* sé nafnorð eða sögn. Mörkunin gerir notendum jafnframt mögulegt að leita að dæmum og flokka þau eftir málfræðilegum einkennum óháð orðunum sem þau birtast í (sjá dæmi um þetta í grein Eiríks Rögnvaldssonar í þessu hefti). Mörkun stórra textasafna er unnin vélrænt með ákveðnum aðferðum en slík úrvinnsla verður ekki rædd frekar hér (sjá um það efni t.d. Sigrún Helgadóttir 2004b og grein hennar í þessu hefti). Ekki verður heldur fjallað um ýmis tæknileg atriði varðandi frágang slíkra safna, sbr. lið (1)4.[4]

Þegar allt er talið eru til talsvert mörg textasöfn með íslenskum textum en þau eru misstór og mjög mismunandi að innihaldi og umbúnaði. Einungis hluti þeirra er öllum opinn og hin eru misaðgengileg til nota í málrannsóknum. Auk eiginlegra textasafna má nefna að veraldarvefurinn er í vaxandi mæli notaður sem n.k. textasafn með hjálp leitarvéla eins og *Google* eða *Emblu*. Stærsta afmarkaða safnið með íslenskum textum er án efa *Textasafn Orðabókar Háskólans* sem lengi hefur komið að góðum notum við orðabókagerð og rannsóknir en það geymir nú um 60 milljónir lesmálsorða. Aðgangur að safninu í heild er takmarkaður við starfsfólk stofnunarinnar og aðra sem fengið hafa sérstakt leyfi til að nota það. Önnur stór og gagnleg íslensk textasöfn eru t.d. *Gagnasafn Morgunblaðsins* og *Lagasafnið* sem bæði eru aðgengileg á vefnum og opin til leitar að orðum eða orðhlutum.

Ekkert ofangreindra safna getur þó talist vera málheild í þeim skilningi sem hér er lagður í hugtakið, hvorki hvað varðar innihald né umbúnað. Í rauninni fullnægir bara eitt textasafn með íslensku nútímamáli flestum þeim skilyrðum sem talin eru í (1) og það er textasafnið sem lagt var til grundvallar við gerð *Íslenskrar orðtíðnibókar* (1991). Safnið er lítið á nútímamælikvarða (um 500 þúsund orð) en það er sett saman úr textabrotum af tiltekinni gerð og í ákveðnum hlutföllum (sbr. (1)1), textarnir eru fullmarkaðir m.t.t. orðflokks og beygingar (sbr. (1)2) og í formála bókarinnar er gerð ítarleg grein fyrir samsetningu safnsins (sbr. (1)3 og 5). Aftur á móti er safnið ekki að fullu

[4]Benda má á grein Renouf (1987) þar sem lýst er gerð málheildarinnar sem lögð var til grundvallar COBUILD og tekið á ýmsum þáttum sem varða innihald hennar, efnisöflun, frágang og umbúnað textanna. Einnig á heimasíðu BNC (British National Corpus) þar sem finna má skýrslur og greinagerðir um flest það sem lýtur að gerð þeirrar málheildar.

aðgengilegt til leitar (sbr. (1)4) en það stendur til bóta.[5] Textasafn Íslendingasagna, sem áður hefur verið nefnt, getur líka kallast málheild í þeim skilningi að það nær til allra tiltækra texta af tilteknu tagi og þeir hafa verið greindir m.t.t. uppflettiorðs og orðflokks (sjá t.d. Eirík Rögnvaldsson 1990), upplýsingar um gerð safnsins liggja fyrir og það er aðgengilegt til leitar, því það hefur bæði verið gefið út á geisladiski (*Íslendinga sögur* 1996) og á vefsetri Eddu útgáfu undir nafninu *Sagnalykill*.

Víkjum nánar að samsetningu málheilda. Tiltölulega litlar málheildir sem taka til skýrt afmarkaðs sviðs fela sumar í sér alla texta sem um er að ræða og gefa þar með heildarmynd af málinu á viðkomandi sviði. Þetta eru einkum sérhæfðar málheildir sem spanna t.d. ákveðna textategund eða tiltekið tímabil í sögu málsins sem tiltölulega fáir textar tilheyra. Sem dæmi um söfn af þessu tagi eru íslenska lagasafnið og textasafn Íslendingasagna[6] sem áður eru nefnd (þótt hvorugt þeirra geti talist fullgild málheild m.t.t. áðurgreindra viðmiða). Málið vandast hins vegar í málheildum sem er ætlað að endurspegla málnotkun á sviði sem útilokað er að ná utan um í heild sinni, t.d. málheild með íslensku ritmáli á 20. öld eða íslensku samtímamáli. Í slíkum tilvikum verður að velja hæfilegt magn úr öllum þeim textum sem völ er á með það markmið í huga að efniviðurinn sé dæmigerður fyrir málnotkun almennt og endurspegli málið innan þess ramma sem málheildinni er settur.

Mikið hefur verið rætt og ritað um val á textum þannig að málheildir birti eðlilegt þversnið af því tungumáli sem þær geyma. Forsenda þess að draga megi almennar ályktanir um mál og málnotkun af niðurstöðum rannsókna sem byggðar eru á tiltekinni málheild er að samsetning hennar endurspegli raunverulega málnotkun í öllum sínum fjölbreytileika (sjá t.d. Teubert & Čermáková 2004:112–118, Biber, Conrad & Reppen 1998:246–250 og Landau 2001:323 o.áfr.). Hér

[5]Sumarið 2006 vann Sigríður Andrea Ásgeirsdóttir að því undir stjórn Sigrúnar Helgadóttur að ganga þannig frá textasafninu að hægt sé að leita í því með hugbúnaðarpakkanum Xaira. Að því búnu getur bæði textinn sjálfur og mörkin nýst að fullu. Ef leyfi fæst frá rétthöfum textanna er ætlunin að opna aðgang að safninu í því formi á vefsíðu stofnunarinnar. Verkefnið var styrkt af Nýsköpunarsjóði námsmanna.

[6]Hér er horft fram hjá textafræðilegum atriðum varðandi form textanna. Textarnir í safninu eru allir með nútímastafsetningu og það takmarkar notkun þess við atriði þar sem ritháttur skiptir litlu máli. En það dugar t.d. ágætlega við rannsóknir á tíðni orða, orðaröð og setningagerð í fornu máli, a.m.k. í máli sagnanna.

koma ýmis sjónarmið til álita: Á textavalið fyrst og fremst að beinast að því sem er gefið út, birt eða flutt opinberlega, t.d. í bókum, blöðum og öðrum fjölmiðlum? Eða á jafnframt að taka með ýmiss konar óútgefin skrif, s.s. sendibréf og tölvuskeyti, og óopinbert eða hálfopinbert talmál, t.d. samtöl og viðræður á lokuðum fundum? Hvernig á að ákvarða hlutfall mismunandi texta? Á hlutfall mismunandi texta að miðast við hlutfallslega skiptingu þess sem er gefið út eða birt opinberlega á tilteknu tímabili eða á líka að taka tillit til þess að útbreiðsla texta og textategunda er mismikil? Dagblöðum er t.d. dreift til þúsunda áskrifenda dag hvern en margar bækur og tímarit höfða til þröngs hóps og slíkir textar eru bara lesnir af fáeinum tugum eða hundruðum málnotenda. Og hvaða mælistiku er hægt að leggja á eðlilegt hlutfall milli opinberlega útgefins efnis og efnis sem tilheyrir einkalífi fólks og daglegum samskiptum? Slíkar vangaveltur eru gagnlegar og þeir sem setja saman málheildir verða að velta slíkum spurningum fyrir sér en þegar kemur að því að safna textum vega ýmis hagnýt atriði ekki síður þungt. Þar má einkum nefna það hversu auðvelt er að nálgast texta af ákveðnu tagi, hvort leyfi fæst til að nota þá og hversu mikil vinna felst í öflun textanna og úrvinnslu þeirra. Almennt séð munu þó flestir sammála um mikilvægi eftirfarandi atriða þegar í hlut eiga málheildir til almennra nota (stundum nefndar *reference corpus* á ensku): (1) að það tímabil og þau svið málsins sem málheildinni er ætlað að spanna (m.t.t. textategunda, efnis, uppruna o.þ.h.) séu skýrt afmörkuð; (2) að samsetning safnsins sé skipulögð fyrirfram, þ.m.t. hlutfall texta af hverri gerð, og að fjölbreytni safnsins sé sem best tryggð; og (3) að það sé gerð nákvæm grein fyrir afmörkun og innihaldi málheildarinnar þannig að notendur geti sjálfir metið hvort og hvernig hún nýtist þeim og hvaða ályktanir þeir geti leyft sér að draga af niðurstöðum sínum.

Eins og fram hefur komið er tekið tillit til ýmissa þátta við val á textum og flokkun þeirra. Einn þeirra er uppruni textanna og Tafla 1 sýnir samsetningu bresku málheildarinnar BNC (British National Corpus) eftir uppruna. Hún sýnir m.a. að talmálstextar eru minna en fimmtungur af BNC. Hlutur talmálsins er því áberandi lítill miðað við ritmál og að því leyti endurspeglar málheildin tæplega almenna málnotkun. Þetta er þó einkenni á velflestum almennum málheildum og ástæður þessa ójafnvægis verða ræddar í þriðja kafla.

Ritmál		**83%**
Bækur, blöð og tímarit	73%	
Ýmislegt ritmál (útgefið og óútgefið)	8%	
Til upplestrar (ræður, handrit o.fl.)	2%	
Talmál		**17%**
Persónuleg samtöl	10%	
Talað mál við ólíkar aðstæður	7%	

Tafla 1 Hlutfall texta í BNC eftir uppruna þeirra (Burnard 2000)

Í ritmálshluta BNC eru 75% nytjatextar (e. non-fiction) af ýmsu tagi og á ýmsum sviðum en 25% textanna eru skáldverk (e. fiction). Flestir textarnir eru yngri en frá 1975. Með hverjum texta eru skráðar upplýsingar um atriði eins og efni hans, höfund, (ætlaðan) lesendahóp o.fl. og því má flokka textana og leita í þeim eftir slíkum atriðum. BNC er ein af helstu fyrirmyndum íslensku málheildarinnar (MÍM) og gert er ráð fyrir að samsetning hennar verði svipuð m.t.t. textategunda og hlutfallsins milli þeirra. Íslenska málheildin verður þó miklu minni en sú breska því áætlað er að MÍM verði um 25 milljónir lesmálsorða á móti 100 milljónum lesmálsorða í BNC. Til samanburðar má nefna að í textasafni *Íslenskrar orðtíðnibókar* (500 þúsund orð) eru skáldverk u.þ.b. 60% textanna (frumsamdar og þýddar skáldsögur auk barnabóka) en aðeins 40% nytjatextar. Þá eru ævisögur og endurminningar flokkaðar með nytjatextum þótt það geti orkað tvímælis og sumir textar af því tagi eigi ekki síður heima með bókmenntaverkum.

Í framhaldi af þessu er vert að líta aðeins nánar á hlutverk og notkun málheilda. Eins og áður segir er þeim fyrst og fremst ætlað að vera heimild um málið eða tiltekinn hluta þess og nýtast sem efniviður í hagnýtum og fræðilegum verkefnum. Þar má í fyrsta lagi nefna tungutækniverkefni af ýmsu tagi, ekki síst þróun og gerð tóla og hjálpargagna sem tengjast máli og málnotkun. Þar má nefna leiðréttingar- og þýðingarforrit, leitarvélar, ýmis hjálpartól fyrir fatlaða og kennsluforrit. Í öðru lagi eru málheildir mikilvæg stoð við samningu orðabóka og annarra handbóka um mál og málnotkun og við gerð kennsluefnis. Í slíkum verkefnum nýtast málheildir vel við greiningu og lýsingu á orðaforða, beygingum og setningagerð því ýmiss konar mynstur í málnotkun, t.d. orðastæður og orðasambönd, birtast mjög skýrt þegar hægt er að draga saman mikinn fjölda dæma og raða þeim innbyrðis.

Stórar málheildir gefa líka góða hugmynd um það hvað er algengt og hvað er sjaldgæft í máli og málnotkun og geta þannig stutt ákvarðanir um áherslur í kennslu og kennsluefnisgerð. Þá eru málheildir og textasöfn mikilvæg uppspretta notkunardæma um orðanotkun, setningagerðir o.fl. (sjá t.d. Landau 2001:296–323). Í þriðja lagi geta málheildir verið dýrmætur efniviður til hvers kyns málrannsókna, t.d. í málvísindum, orðabókafræði og tungutækni. Meðal þess sem lesa má úr málheildum eru ýmiss konar tíðniupplýsingar, t.d. um hlutfallslega tíðni orða og orðmynda, orðflokka, beygingarmynda og beygingarflokka, orðasambanda og setningagerða; einnig vitneskju um orðaforða tiltekinna texta og textategunda svo og vitnisburð um form, notkun og merkingu orða, orðasambanda og orðastæðna, og um setningagerðir af ýmsu tagi (sbr. Sigrún Helgadóttir 2004).

Niðurstöður tíðnirannsókna á íslensku hafa m.a. birst í *Íslenskri orðtíðnibók* (1991). Eins og áður hefur verið lýst ræðst almennt gildi niðurstaðna m.a. af því hvernig samsetningu málheildar er háttað en einnig af stærð hennar. Stærð safnsins og fjölbreytileiki textanna getur skipt verulegu máli í sambandi við orðaforða og tíðni orða. Þetta sést vel á tíðnitölum einstakra orða, t.d. eru *vúdú-iðkun* og *votviðri* jafnalgeng og útbreidd orð samkvæmt *Íslenskri orðtíðnibók* (1991:527) sem er eins og fyrr segir byggð á tiltölulega litlu textasafni. Þetta er vel þekkt vandamál. Ýmis önnur svið málsins hafi ekki verið rannsökuð jafn mikið á grundvelli textasafna og málheilda en gera má ráð fyrir að um þau gegni svipuðu máli, sérstaklega þar sem hinar stærri einingar málsins eiga í hlut, s.s. orðasambönd og setningagerðir, en síður varðandi beygingarfræðileg atriði eins og t.d. hlutfallslega tíðni einstakra falla.

3 Talmál og talmálsheimildir

3.1 Talmál og ritmál

Málnotendur skynja margvíslegan mun á talmáli og ritmáli og oft má heyra fullyrðingar um að eitthvað „komi bara fyrir í daglegu tali“ eða „sjáist aðallega í rituðu máli“. Þótt slíkar staðhæfingar byggi fyrst og fremst á tilfinningu fólks er ljóst að það er raunverulegur munur á dæmigerðu töluðu máli og rituðum texta sem sést vel ef borin er saman skráning á raunverulegu samtali (sjá t.d. dæmi undir (3) í kafla 3.3

hér á eftir) og samtali í skáldsögu. Hins vegar er lítið vitað um umfang og eðli þessa munar því samanburðarrannsóknir á talmáli og ritmáli eru fáskrúðugar. Það sama má segja um sambandið á milli talmáls og ritmáls ef frá eru talin tengsl framburðar og stafsetningar.

Ein þeirra spurninga sem vakna um mun talmáls og ritmáls og sambandið þar á milli er hvers eðlis munurinn sé. Er fyrst og fremst um stílmun að ræða sem birtist þá í ólíku vali á orðum og orðalagi eða er einhver formgerðarmunur á talmáli og ritmáli? Og ef svo er, í hverju felst þá þessi munur? Þetta er að verulegu leyti ókannað svið, ekki bara í íslensku heldur almennt í tungumálum, og ástæðan er ekki síst sú að heimildir um talmál hafa lengstum verið af skornum skammti. Þetta kann að þykja undarleg staðhæfing þegar haft er í huga að talmálið er sínálægt í daglegu lífi. Tal er hins vegar hverfult fyrirbrigði. Að vísu er hægt að leggja einstakar framburðarmyndir og einstök orð á minnið eða skrá þau hjá sér en til þess að hægt sé að rannsaka stærri einingar eða fá yfirlit um útbreiðslu og tíðni tiltekinna einkenna þarf að koma talinu í það form að vinna megi með það. Með vaxandi áhuga á sérkennum talmálsins hefur þetta verið gert í auknum mæli eins og talmálshlutinn af BNC ber m.a. vitni um og gagnamálfræðingar hafa staðið fyrir samanburðarrannsóknum á tal- og ritmáli á grundvelli málheilda, einkanlega í ensku. Niðurstöður sýna greinilegan stílmun milli talmáls og ritmáls. Hann birtist t.d. í mismunandi tíðni tiltekinna einkenna en rannsóknir benda jafnframt til þess að fleira spili þarna inn í, ekki síst mismunandi aðstæður og þar með málsnið, og myndin sé margbreytilegri en gefið er í skyn með einfaldri tvískiptingu milli talmáls og ritmáls (sjá t.d. Biber 1988, Finegan & Biber 2001). Á grundvelli málheilda og rannsókna á þeim hafa einnig verið gerðar mállýsingar og samdar málfræðihandbækur þar sem tekið er tillit til sérkenna talmáls ekki síður en ritmáls, a.m.k. hvað varðar ensku, og má nefna sem dæmi málfræðibækur sem gefnar hafa verið út á vegum Longman (sjá einkum Biber, Johansson, Leech, Conrad & Finegan 1999).

Íslenskt talmál hefur sáralítið verið rannsakað og þar bíða því mörg forvitnileg athugunarefni. Talað mál hefur alla tíð verið fyrirferðarmikið í einkalífi og persónulegum samskiptum fólks en vægi talmáls í opinberu lífi hefur verið vaxandi á undanförnum áratugum, m.a. vegna breytinga í fjölmiðlun þar sem mikilvægi talmiðla hefur smám saman aukist á kostnað dagblaða. Þetta kallar á það að talmál

fái aukið vægi í móðurmálskennslu og í hjálpargögnum eins og orðabókum og öðrum handbókum um mál og málnotkun. Fjöldi þeirra sem læra íslensku sem annað mál fremur en erlent tungumál kallar líka á breyttar áherslur í kennslu. Þarna er um að ræða fólk sem hefur sest að á Íslandi og lærir íslensku til að geta tekið eðlilegan þátt í samfélaginu. Fyrir slíka nemendur er mikilvægast að ná tökum á daglegu máli, töluðu jafnt sem rituðu. Rannsóknir á töluðu máli og einkennum þess eru því ekki bara fræðilega áhugaverðar heldur hafa niðurstöður slíkra rannsókna líka hagnýtt gildi því þær eru forsenda þess að hægt sé að draga fram sérkenni talmálsins á skipulegan hátt.

3.2 Heimildir um talmál

Upphaflegar áætlanir gerðu ekki ráð fyrir talmálstextum í MÍM. Slíkur efniviður liggur síður á lausu en ritmálstextar og fjárhags- og tímarammi verkefnisins leyfði ekki þá vinnu sem öflun og úrvinnsla slíkra texta krefst. Þó var ljóst að þeir myndu auka mjög gildi verksins þar sem talmálsefni er forsenda fyrir því að málheildin gefi raunsanna mynd af íslensku samtímamáli. Það rættist þó úr þegar samvinna tókst á milli MÍM og tilbrigðaverkefnisins um samnýtingu á textum og verkaskiptingu við öflun þeirra og úrvinnslu. Nú er fyrirsjáanlegt að í íslensku málheildinni verði talmálshlutinn 2–3% af safninu. Þetta er mun lægra hlutfall en í BNC en þótt hlutfallið sé lágt er talmálsefnið samt rúmlega 500 þúsund lesmálsorð, þ.e.a.s. heldur meira en allt safnið sem orðtíðnibókin byggir á.

Almennt séð verður að gera sömu kröfur um fjölbreytileika texta úr talmáli og ritmáli en viðmiðin eru að nokkru leyti önnur. Helstu tegundir talmálstexta eru samtöl, viðtöl og eintöl, t.d. ræður og fyrirlestrar, og málsnið talmálstexta er mismunandi eftir aðstæðum og málumhverfi. Ólíkt ritmálstextum sem flestir eru verk eins höfundar eru þátttakendur í samtölum og viðtölum fleiri. Fjöldi þeirra og innbyrðis tengsl hafa áhrif á málsniðið, t.d. það hvort viðmælendur þekkjast vel eða eru ókunnugir; einnig getur aldur þeirra, kyn, uppruni o.fl. skipt máli. Taka þarf tillit til alls þessa við val texta þannig að málheildin endurspegli fjölbreytileika talmálsins.

Talmálshluti BNC skiptist t.d. í tvennt. Helmingur efnisins eru persónuleg samtöl frá rúmlega hundrað manns: körlum og konum á ýmsum aldri, úr ólíkum þjóðfélagshópum og frá ýmsum landshlutum.

Hinn helmingurinn eru viðtöl og eintöl sem tengjast ákveðnum sviðum þjóðlífsins: menntun og fræðslu, viðskiptum, opinberu lífi (þingræður, predikanir o.fl.) og frítíma (t.d. íþróttalýsingar) (sjá Burnard 2000). Upplýsingar um eðli, uppruna og önnur einkenni talmálstextanna eru skráðar til þess að hægt sé að sjá úr hvers konar textum tiltekin dæmi eru runnin. Þær upplýsingar gagnast líka við flokkun textanna eftir tilteknum einkennum og til þess að takmarka leit við texta af ákveðnu tagi.

Stærstum hluta talmálsefnisins sem verður hluti af MÍM er ekki safnað á vegum málheildar- og tilbrigðaverkefnanna sjálfra heldur er um að ræða endurnýtingu á efni sem var safnað í öðrum tilgangi. Íslensku textarnir eru tæplega eins fjölbreytilegir og talmálsefnið í BNC og ekki eins skipulega valdir m.t.t. efnis og notkunarsviðs. Meðal þeirra eru þó samtöl, viðtöl og eintöl karla og kvenna á ýmsum aldri og úr ýmsum áttum. Efnið sem um er að ræða er talið í (2)1-4.

(2) 1. Persónuleg samtöl sem hljóðrituð voru fyrir ÍSTAL-verkefnið[7] árið 2000 (u.þ.b. 20 klst.)
2. Viðtöl við hópa fólks sem hljóðrituð voru sumarið 2002 fyrir rannsókn á aðkomuorðum í norrænum málum (MIN)[8] (u.þ.b. 10 klst.)
3. Umræður á Alþingi[9] frá árunum 2004 og 2005 (u.þ.b. 20 klst.)
4. Samtöl ungs fólks um ákveðið efni, ýmist við jafnaldra sína eða eldra fólk, hljóðrituð sumarið 2006[10] (u.þ.b. 4 klst.)

[7] Markmið ÍSTAL var að safna efni í íslenskan talmálsbanka til notkunar við málsrannsóknir og tungutækniverkefni af ýmsu tagi. Að verkefninu stóð hópur málfræðinga við Háskóla Íslands, Kennaraháskóla Íslands og Orðabók Háskólans og verkefnisstjóri var Þórunn Blöndal. Verkefnið var unnið fyrir styrk úr tæknisjóði Rannís (markáætlun um upplýsingatækni og umhverfismál). Nánari upplýsingar um verkefnið eru á heimasíðu þess: http://www.hi.is/~eirikur/istal/ (sjá líka Þórunn Blöndal 2005:108–110).

[8] MIN var norrænt samstarfsverkefni undir stjórn Helge Sandøy. Viðtölin, sem Hanna Óladóttir og Halldóra Björt Ewen stjórnuðu, eru úr þeim hluta rannsóknarinnar sem beindist að viðhorfum málnotenda til erlendra aðkomuorða í íslensku. Upplýsingar um verkefnið er að finna á vefsíðunni http://moderne-importord.info/.

[9] Umræðurnar voru hljóðritaðar á vegum þingsins sem veitti verkefninu aðgang að þeim til nota við rannsóknar- og þróunarverkefni.

[10] Sigrún Ammendrup hljóðritaði þessi samtöl og vann úr þeim á vegum tilbrigða-

Alls er þetta nálægt 55 klukkutímum af hljóðrituðu efni sem búið var að skrá misnákvæmlega á vegum þeirra sem létu það í té.

3.3 Öflun og úrvinnsla talmálsefnis

Efnisöflun úr talmáli er um margt flóknari og kostnaðarsamari en öflun ritmálstexta sem flestir eru þegar til í rafrænu formi. Forsendan fyrir því að hægt sé að nýta talmálsefni í málheildum er að það sé hljóðritað og talmál verður heldur ekki rannsakað nema að takmörkuðu leyti nema með hljóðritun. Eðlilegt talað mál sprettur ekki síst upp í einkalífi fólks og ekki er sjálfgefið að það leyfi hljóðritun á samtölum við vini, kunningja og vinnufélaga og notkun á upptökunum til rannsókna. Það sama á við um margvísleg önnur samskipti jafnvel þótt þau séu ekki alveg eins persónuleg, s.s. það sem fram fer í skólastofu milli kennara og nemenda eða samtöl innan stofnana og fyrirtækja, t.d. milli starfsfólks og viðskiptavina eða á fundum. Auðveldara er að afla efnis sem flutt er opinberlega, t.d. í fjölmiðlum, en eitt sér gefur það mjög takmarkaða mynd af talmáli almennt. Hljóðrituninni fylgja ýmis tæknileg og aðferðafræðileg úrlausnarefni sem ekki er ástæða til að tíunda hér en snerta m.a. gæði upptökunnar og það hvernig takast má að ná fram eðlilegu tali þrátt fyrir áhrif þess að verið sé að taka það upp (þetta er hinn vel þekkti „Observer's Paradox" eða þversögn athugandans, sbr. t.d. Feagin 2002:20, Þórunn Blöndal 2005:106).

Björninn er þó engan veginn unninn þótt tekist hafi að útvega hljóðritað talmálsefni því ekki er hægt að nýta hljóðupptökurnar beint. Áður en hægt er að fella efniviðinn inn í málheild eða rannsaka tiltekin einkenni sem þar birtast þarf að umrita talið, þ.e.a.s. skrá allt sem sagt er og breyta þannig talinu í texta. Þetta er mikið nákvæmnisverk og tímafrekasti hluti úrvinnslunnar. Hægt er að fara ýmsar leiðir við umritun og hún getur verið mjög misítarleg. Mikilvægast er að skrá allt sem sagt er orðrétt eftir hljóðritinu og í réttri röð. Sjaldgæft er að talmálsefni sé umritað með hljóðritunartáknum en stundum er farin sú leið að líkja eftir tilteknum framburðaratriðum í umrituninni, skrifa t.d. *oní* ('ofan í') og *ettir* ('eftir') þar sem það á við. Íslenska talmálsefnið sem sagt var frá hér á undan (sbr. (2) í kafla 3.2) er aftur á móti skráð með hefðbundinni stafsetningu og umritunin er því ónákvæm m.t.t. framburðar og annarra hljóðþátta (áherslur, tónfall) þótt sums

verkefnisins, m.a. fyrir styrk úr Nýsköpunarsjóði námsmanna.

staðar séu settar inn athugasemdir um ýmiss konar sérkenni í framburði. Á hinn bóginn þótti venjuleg stafsetning tryggja best þá stöðlun sem er nauðsynleg til að leit í efninu gangi greiðlega og skili öllum þeim dæmum sem notendur sækjast eftir. Auk orða og orðmynda eru ákveðin atriði sem varða framvindu samtalsins eða ræðunnar skráð kerfisbundið. Þar má nefna ófullgerð orð og setningar, þagnir, framígrip, skörun o.fl. Slík atriði geta skipt máli fyrir túlkun dæma, t.d. er munur á því hvort sagt er *ég dreymdi* í einni samfelldri lotu eða *ég # dreymdi* með hiki eða þögn á undan sögninni því þögnin gæti verið merki um að sá sem talar hafi ætlað að segja eitthvað annað en séð sig um hönd í miðjum klíðum.

Undir (3) er sýnt dæmi um umritun; textinn er brot af samtali úr ÍSTAL-gögnunum.

(3) *Dæmi um umritun úr* ÍSTAL[11]

A: er þetta já þetta er vélin sem G gaf þér
B: já
A: hún er ofsalega skemmtileg
B: já hún er <A>rosalega</A>
A: maður</B> ýtir bara á <B>er það ekki</B> já
B: jú bara
B: þú bara sérð og
A: halló mamma
D: <A>æðislegt</A>
A: (þessi er)</D> flott (.)
D: maður þarf líka að eiga bara svona <A1>einhverja svona bara</A1> imba sem (maður) bara <A2>smellir</A2>
A1: en svo hérna
A2: þetta er</D> bara (engi-) þetta er í rauninni enginn imbi sko <!D> því hún er með fókus og allt=

Oft er notast við venjuleg ritvinnsluforrit við skráningu talmálsefnis en einnig eru til ýmis forrit sem flýta fyrir umritun og tryggja betra samræmi í skráningu. Ef hljóðritin eru í stafrænu formi geta slík forrit yfirleitt líka tengt saman hljóð og texta með e.k. kóðun. Með því

[11] A, B og D tákna þátttakendur í samtalinu. Bókstafir innan oddklofa sýna hvar skörun eða framígrip á sér stað, t.d. sýnir strengurinn „B: já hún er <A>rosalega </A> A: maður </B> ýtir bara á" (í línu 4 og 5) að A byrjar að tala áður en B hættir og orðin "rosalega" (hjá B) og "maður" (hjá A) skarast. Svigi utan um orðmynd táknar að hún heyrist illa og því sé um n.k. tilgátu að ræða. Bandstrik í enda orðs (t.d. engi-) táknar að því hafi ekki verið lokið.

að koma samtengdum hljóð- og textaskrám fyrir í gagnagrunni er þá hægt að leita í gögnunum og rata ekki bara inn í umritað dæmi heldur líka á réttan stað í hljóðskránni. Í MÍM verða þó einungis textaskrár[12] en textarnir verða markaðir þannig að bæði verði hægt að nýta orðmyndir og mörk til leitar auk þeirra einkenna á textunum sjálfum sem eru skráð (textategund, uppruni o.fl.). Eigi að síður er stefnt að því að tengja saman hljóð og texta í íslenska efninu til þess að greiða fyrir aðgangi að hljóðskránum en ekki er enn ljóst hvernig búið verður um safnið í þeirri mynd og hvar það verður vistað.

4 Talmál og orðabækur

4.1 Orðaforði talmáls og ritmáls

Oft er gert ráð fyrir því að orðaval og orðanotkun í tali og riti sé mismunandi að einhverju marki. Þetta hefur ekki verið rannsakað ítarlega en samanburðarathugun á textadæmum úr talmáli og ritmáli bendir til þess að þetta eigi við rök að styðjast. Orðaforðinn í hluta ÍSTAL-samtalanna (u.þ.b. helmingur safnsins) var borinn saman við orðaforðann í tveimur litlum textasöfnum úr ritmáli sem hvort um sig var álíka stórt og talmálssafnið. Í öðru safninu voru tiltölulega óformlegir og persónulegir textar (dagbækur o.fl.) en textarnir í hinu voru formlegri og ópersónulegri (dagblaðatextar). Samanburðurinn sýndi að ákveðin orð og orðmyndir eru sérstaklega einkennandi fyrir talmálstextana í samanburði við hina en hann leiddi jafnframt í ljós að talsverður innbyrðis munur er á ritmálssöfnunum tveimur (Ásta Svavarsdóttir 2003). Dæmi um nokkur þeirra orða sem einkenna talmálstextana eru sýnd í Töflu 2.

Taflan sýnir að sumar orðmyndirnar koma alls ekki fyrir í ritmálstextunum (*ókei*, *heyrðu*) og aðrar einungis í óformlegu textunum (*svoleiðis*, *rosalega*, *sko*). Allar koma þær miklum mun sjaldnar fyrir í ritmálsefninu en í talmálsgögnunum og í sumum tilvikum er munurinn sláandi (*bara*). Sem kunnugt er byggist *Íslensk orðtíðnibók* (1991) ein-

[12] Umritunarskrám úr því talmálsefni sem hér hefur verið lýst hefur nú flestum verið komið fyrir í textasafni Orðabókar Háskólans. Þar er til bráðabirgða hægt að leita í textanum ómörkuðum (undir flokknum *Talmál* og undirflokkum hans á leitarsíðunni *http://www.lexis.hi.is/corpus/leit.pl*) en notkunarmöguleikar þeirra munu síðar aukast með tilkomu málheildarinnar.

göngu á ritmálstextum en þeir eru bæði miklum mun fleiri og annars konar en þeir textar sem hafðir voru til samanburðar í athuguninni. Allar orðmyndirnar í Töflu 2 koma þar fyrir nema boðháttarmyndin *heyrðu* en tíðni þeirra er lág, t.d. er bara eitt dæmi um *ókei*. Þrjú orðanna koma fyrir í meira en helmingi textanna og það eru jafnframt þau orð sem flest dæmi eru um. Þetta eru sömu orðin og þau sem reyndust algengust í ritmálsefninu í samanburðarathuguninni (*bara*, *kannski* og *einmitt*).

	ÍSTAL	Óformlegt ritmál	Formlegt ritmál
bara	757	100	7
svoleiðis	65	6	0
kannski	128	57	7
einmitt	79	18	2
náttúrulega	134	6	0
rosalega	44	2	0
ókei	16	0	0
heyrðu	65	0	0
meina	80	1	1
sko	670	5	0

Tafla 2 Samanburður á dæmafjölda nokkurra orðmynda í talmáli og ritmáli

Orðmyndirnar í Töflu 2 eru teknar úr samhengi og því er óvarlegt að fullyrða neitt um merkingu þeirra eða hlutverk. Í fljótu bragði virðast þó flestar þeirra vera áhersluorð, hikorð eða annars konar orðræðuagnir sem algengar eru í tali. Athugunin bendir líka til þess að ákveðin orð af öðru tagi komi frekar fyrir í talmáli en ritmáli, þ.á m. upphrópanir (t.d. *almáttugur*, *andskoti*) og ýmis lítt aðlöguð tökuorð (t.d. *partí*, *frílans*, *breinstorma* o.fl.; Ásta Svavarsdóttir 2003:46).

4.2 Talmálsorð í orðabókum

Málheildir og önnur textasöfn gegna m.a. hlutverki við orðabókagerð. Til þessa hafa íslenskar orðabækur fyrst og fremst verið byggðar á ritmálsheimildum auk þekkingar og máltilfinningar ritstjóranna sjálfra.

Forvitnilegt er að athuga hvaða áhrif það hefði á lýsingu orða sem einkum eru bundin við talmál ef einnig væri stuðst við það talmálsefni sem nú er tiltækt.

Hér verða nokkrar orðalýsingar úr *Íslenskri orðabók* skoðaðar í þessu ljósi. Dæmi eru tekin af nokkrum orðanna sem koma lítt eða ekki fyrir í ritmáli samkvæmt Töflu 2 (sjá kafla 4.1). Einkum er litið á nýjustu útgáfu bókarinnar (ÍO-2002) en einnig eru tekin dæmi úr þeim eldri (ÍO-1963 og ÍO-1983). Til samanburðar eru svo birt dæmi úr ÍSTAL-efninu. Orðin sem skoðuð eru voru valin út frá Töflu 2 (sjá kafla 4.1).

Lítum fyrst á orðmyndina *heyrðu*. Formlega er þetta boðháttarmynd af sögninni *heyra* og í eldri útgáfum orðabókarinnar er hún ekki sjálfstætt uppflettiorð. Þetta breyttist við nýjustu endurskoðun hennar og í ÍO-2002 er *heyrðu* orðið sérstakt uppflettiorð og greint sem upphrópun eins og sést í (4).

(4) *ÍO-2002* (hluti orðlýsingar)

> **heyra** *–ði* s 1 greina, skynja hljóð [...] *heyrðu mig (mér)* komdu og talaðu við mig...
> ...
> **heyr|ðu** UH • notað til að ná athygli viðmælanda, oftast í upphafi samtals eða frásagnar ▷ *Heyrðu! Áttu nokkuð sígarettu?* / **heyrðu mig (mér)** hlustaðu nú á mig sbr. heyra

Í talmálsefninu í heild eru rúmlega 240 dæmi um orðmyndina *heyrðu*, allt boðháttarmyndir. Sé rýnt í dæmin úr ÍSTAL-samtölunum virðist orðmyndin þó ekki notuð í eiginlegri boðháttarmerkingu, sbr. dæmin í (5). Áberandi er að *heyrðu* stendur mjög oft í upphafi lotu, þ.e.a.s. þar sem nýr þátttakandi fær orðið (sbr. línu 1, 3, 5 o.fl). Orðmyndin virðist einkum gegna því hlutverki að kalla á athygli og vera e.k. yfirlýsing um að nú taki mælandinn orðið, oft til þess að koma að innskoti eða útúrdúr í samtalinu. Í þessu samhengi felur boðháttarmyndin ekki í sér boð eða fyrirmæli um að 'nema hljóð, hlýða á'. Því má segja að *heyrðu* gegni fremur tilteknu hlutverki í framvindu samtalsins en að hún tjái ákveðna merkingu.

(5) ÍSTAL *(hluti dæma)*

1 yrði að vera fimmtándi eða B:	heyrðu <A1>hvernig hvernig</A1> stendur á helgi þá sko getum við
2 allt safnast saman hjá þér	heyrðu já átjánda og þú átt að fara að já nú sé ég af hverju þú varst svona
3 B: ert þú ekki tilvalin í það A:	heyrðu hún spurði einmitt <B>bara (lang-) langar þig ekki ((dynkir)) að
4 og sagði hvenær og og þá já	heyrðu þá er best að ég geri það líka heldurðu að ég geti fengið þær eftir
5 síðan í morgun maður var= A: =	heyrðu hvað= B: =ég veit ekki hvernig stendur á því þetta er svona langur
6 skrifað dagsetningu á já	heyrðu Kári var hérna þetta gamall þegar já hann var í þessum fötum ú ég
7 guð hvað á ég að segja núna	heyrðu hvernig er þetta nú særi ég hana og (blabla) A: já B: en hún getur
8 tengdamamma allt í einu svona	heyrðu já Ella er ekki einhver sem þú þekkir sem heitir Jóhannes Freyr
9 hefðirðu getað farið að skoða A:	heyrðu já það minnir mig ég ætla að athuga bækurnar þarna kannski B: já
10 jæja gjörið þið svo vel D: já A:	heyrðu Guðrún Ösp kemst ekki C: (x) þau voru <B>hérna áðan</B> B:
11 C: ja hann fékk sér bara aðeins	heyrðu <!D> rjóminn D: já A: ég ætlaði (x) að fara að segja (x) rjómann
12 alla vega bæði á r D: já B:	heyrðu ég man alltaf eftir sögunni af hérna (.) sem amma þín sagði mér (.)
13 C:	heyrðu þetta er allt</B> <A>þetta er allt</A> þetta er allt prjónað fast
14 þessa uppskrift ### A:	heyrðu Inga þú mátt eiga þetta því að <B1>## frosna</B1> grænmetið
15 B:	heyrðu hvað er að frétta af Einari A: neinei ég kem bara með þér ##

Í *Íslenskri orðabók* er *heyrðu* annars vegar lýst í orðasambandinu *heyrðu mig* (eða *mér*) ‘komdu og talaðu við mig’ og hins vegar sem „upphrópun til áherzlu, til að vekja athygli eða til að láta í ljós hrifningu“ eins og það var orðað í ÍO-1963. Orðasambandið *heyrðu mig/mér* kemur alls ekki fyrir í talmálsefninu en síðari skýringin felur greinilega í sér viðleitni til að lýsa þeirri notkun sem virðist ríkjandi í talmáli, a.m.k. í samtölum. Eins og sjá má í (4) hefur þessi skýring verið þróuð áfram í síðari útgáfum og í ÍO-2002 er t.d. vikið að dæmigerðri stöðu hennar.

Atviksorðið *rosalega* er ekki í ÍO-2002, einungis lýsingarorðið *rosalegur* ‘svakalegur, ofsalegur (m.a. um fólk)’. Í textasafni Orðabókarinnar eru tæplega 700 dæmi um orðið en langstærstur hluti þeirra er annaðhvort úr óformlegu ritmáli (bloggtextum) eða talmáli. Dæmin í talmálsefninu eru t.d. hátt í 300, álíka mörg og í öllum ritmálstextunum í safninu að blogginu frátöldu, enda þótt fjöldi texta úr ritmáli sé margfaldur á við talmálið. Heimilda um merkingu og notkun atviksorðsins er því greinilega fyrst og fremst að leita í gögnum um talað mál og í (5) eru sýnd nokkur dæmi um notkun þess í samtölum.

(6) ÍSTAL (hluti dæma)

1 nefnilega það <B5>breyti ekki</B5> B1: þetta er	rosalega girnilegt
2 Kópavogi B: jájá var það ## skemmtilegt A: já	rosalega það var bara
3 B: mm= A: =ef þú fengir nú= B: =þetta er	rosalega gott A: ha
4 núna <!B1> ## hann hefur verið að taka alveg	rosalega <B2>fín</B2> ## B: nei A: jú B: nei það er
5 flott= A: =já	rosalega <!B1> góð fara í þýsku hann hefur heyrt að
6 þetta væri svo	rosalega erfitt A: já B: já ## A: ## B: já ## A: þær eru
7	rosalega stressaðar þetta betra A: æ jájá þakka þér fyrir
8 þetta er	rosalega huggulegt
9 veslast upp af netskorti</H> já A: nei ég hef	rosalega litla samúð tólf og mér ## á mínum tíma ##= A:

10 =hann var	rosalega klár í meira kaffi ## #### B: #### A: já mikið
11	rosalega er ég fegin í fermingarveislum hérna <!B1> mér
12 fannst þetta	rosalega skrítið að kunni ekki við þetta ## hm #### en
13 þetta var	rosalega fínt ## nema B: nei ég held ekki A: nei ## ég
14 keypti svo	rosalega flotta 1 1:1 sjö hundruð eða sjöþúsundkall ha A:
15 jájá það er	rosalega <C>fínt</C> minn ekki svona niður B: já ##
16 #### mikið	rosalega var mikið af við fórum út að borða í gærkvöld
17 #### á alveg	rosalega skemmtilegan

Dæmin í (6) sýna að *rosalega* er fyrst og fremst notað með lýsingarorði eða atviksorði til áherslu. Stærstur hluti dæmanna í talmálsefninu í heild er af þessu tagi og úrvalið gefur því ágæta mynd af notkun orðsins í talmáli. Nánari athugun sýnir að orðið stendur mjög oft í sambandinu *alveg rosalega* (með lo. eða ao. á eftir; sbr. línu 4 og 17). Á grundvelli talmálsefnisins mætti því setja saman orðabókarlýsingu á merkingu og hlutverki orðsins og taka þaðan dæmi til skýringar.

Þriðja orðið sem tekið er sem dæmi, orðið *ókei*, verður að teljast dæmigert talmálsorð. Það er tiltölulega nýtt í málinu og kemur nánast ekki fyrir í ritmáli. Í öllu textasafni Orðabókar Háskólans (rúmlega 50 milljón orð) eru innan við 100 dæmi úr útgefnum ritmálstextum en aftur á móti eru heldur fleiri dæmi í talmálsefninu þótt það sé u.þ.b. hundrað sinnum minna að vöxtum. Þarna hlýtur því aðgangur að talmálsheimildum að skipta verulegu máli fyrir orðlýsinguna. Samanburður á lýsingu orðsins í ÍO-1963 til ÍO-2002 leiðir í ljós forvitnilega þróun eins og sjá má í (7).

(7) *ÍO-1963*

?**ókei** uh, allt í lagi

ÍO-2002

ókei [...] UH óforml. **1** (sem svar eða kveðja) allt í lagi **2** (við útskýringu eða frásögn annars sem samþykki eða boð um skilning, eða sem spurning um skilning annars) þannig já, það er einmitt það, ég skil • (sem spurning um skilning eða samþykki annars) skilið? allt í lagi?

Eldri greinin er mjög stutt en sú nýrri er mun margbrotnari og ítarlegri. Líklegt má telja að tíðni *ókei*, sem var tiltölulega nýtt orð í íslensku í upphafi 7. áratugarins (elsta þekkta dæmi um orðið er frá 1942 samkvæmt RMS; sjá Þórunni Blöndal 2002/2006), hafi aukist og notkun þess breyst á þeim 40 árum sem liðu milli útgáfanna. Munurinn á lýsingunum tveimur kann því að endurspegla breytingar á

notkun orðsins, a.m.k. að einhverju leyti. Yngri greinin ber þó með sér að þar sé gerð tilraun til fyllri orðlýsingar þar sem tekið er tillit til málaðstæðna og hlutverks orðsins ekki síður en merkingar. Mörg dæmigerð talmálsorð kalla einmitt á slíka lýsingu og við samanburð útgáfanna hefði óneitanlega verið forvitnilegt að hafa aðgang að gömlu talmálsefni til samanburðar við það nýja.

Ritstjórar orðabókarinnar hafa tæplega haft aðgang að beinum heimildum um talmálið og hafa því þurft að styðjast við eigin málkennd og það sem þeir heyrðu í umhverfi sínu. Lýsingin í ÍO-2002 virðist þó í meginatriðum fara nokkuð nærri raunverulegri notkun orðsins. Þórunn Blöndal (2002/2006) hefur skoðað dæmi um *ókei* í samtölum og niðurstaða hennar er m.a. sú að málaðstæður hafi mikið að segja um notkun orðsins og hlutverk. Hún greinir fimm afbrigði í merkingu og notkun *ókei* í gögnunum sem hún styðst við. Greining hennar svarar að mörgu leyti til þeirrar lýsingar sem birtist í orðabókinni en Þórunn nefnir einnig notkunarafbrigði sem ekki koma fram í ÍO-2002, t.d. notkun *ókei* sem eins konar inngangs að beinni ræðu (sjá Þórunn Blöndal 2006:20,1). Þetta bendir til þess að enn mætti bæta orðlýsinguna með því að hafa hliðsjón af raunverulegu talmálsefni og þangað mætti líka sækja notkunardæmi til stuðnings lýsingunni, en eins og sést í (7) eru engin dæmi birt í orðabókargreininni.

Síðasta talmálsorðið sem hér verður tekið dæmi af er *sko*. Það er tíunda algengasta orðið í ÍSTAL-efninu (sbr. Þórunn Blöndal 2005:36) og því miklu algengara í talmáli en í ritmáli þar sem það er í 870. sæti samkvæmt *Íslenskri orðtíðnibók*. Eins og fram kemur í (7) má sjá svipaða þróun í lýsingu þessa orðs frá einni útgáfu *Íslenskrar orðabókar* til annarrar og þá sem kom fram í lýsingunni á *ókei*. Þar sem *sko* er gamalgróið orð í íslensku ber hún þó væntanlega fremur vott um viðleitni til að bæta orðlýsinguna og gera hana ítarlegri en að hún endurspegli verulegar breytingar á merkingu, hlutverki og notkun orðsins.

(8) *ÍO-1963*

> **sko** uh, bh af **skoða**
> . . .
> **skoða** [. . .] **4** bh: [. . .] *sko* (= *skoðaðu*): *sko manninn* sjáðu manninn, þarna er maðurinn, *sko til* líttu á, þarna geturðu séð.

ÍO-1983

> **sko** UH (einsk. BH af skoða) **1** sjáðu, lítið á, *s. til* þarna sérðu ☞ *skoða* 4. **2** notað sem AO (sbr. dönsku *sgu*), sannarlega, svo sem*: hann er sko ekkert flón*.

ÍO-2002

> [1]**sko** UH • (ábending um að veita athygli) sjáðu, lítið á ▷ *sko tunglið!* • lýsir viðurkenningu ▷ *sko strákinn* sjáðu strákinn, lítið á strákinn; sei, sei, þetta gat hann! • notað sem hikorð ▷ *ég, sko, fór og talaði við hann, sko, en... / sko til* **1** þarna sérðu **2** líttu á, þarna geturðu séð **3** nú, þetta var þá hægt! sbr. skoða (4)
> [2]**sko** AO • svo sem, reyndar, svo sannarlega, aldeilis ▷ *hann er sko ekkert flón / nei, það gengur sko ekki vel / það mundi ég sko ekki gera!*

Þarna sést að lýsing orðsins verður smám saman nákvæmari og ítarlegri. Í fyrstu útgáfunni er greinin varla nema vísun til so. *skoða*. Þar er orðmyndin *sko* felld undir boðhátt sagnarinnar (jafnvel þótt uppflettiorðið *sko* sé merkt sem UH) og notkun hennar er lýst með dæmum og skýringu á þeim. Millivísunin til so. *skoða* er enn til staðar í annarri útgáfu en þar er lýsingin undir uppflettiorðinu *sko* öll fyllri: henni hefur verið skipt í tvennt, merkingarskýringum og notkunardæmum hefur verið bætt við og þar að auki er vísað til danska orðsins *sgu* sem hliðstæðu þótt orðsifjaupplýsingar séu annars mjög fátíðar í *Íslenskri orðabók*. Og í nýjustu útgáfunni eru tvær flettur með uppflettimyndinni *sko*, upphrópun og atviksorð. Lýsingin er orðin enn ítarlegri og sundurgreining meiri, sérstaklega í fyrri flettunni, og skírskotun til málaðstæðna er orðin áberandi þáttur í orðlýsingunni líkt og í lýsingu orðsins *ókei*. Auk þess hefur notkunardæmum verið fjölgað. Þarna eru m.ö.o. stigin skref til raunsannari lýsingar á hlutverki og notkun orðsins í töluðu máli og með hliðsjón af talmálsgögnum sem sýna dæmi um raunverulega notkun orðsins mætti sannreyna og bæta lýsinguna. Mynd 1 sýnir hluta dæmanna í ÍSTAL en lesendum er látið eftir að bera þau saman við orðabókarlýsinguna. Einnig skal bent á athugun og greiningu Helgu Hilmisdóttur og Camillu Wide (1999) á notkun *sko* í tali ungs fólks.

Mynd 1. Hluti dæma um orðmyndina *sko* í ÍSTAL

```
1 þú ge̗tur þvegið gĺugga án þess að                sko nota nokkuð efni A: er það Lí:SI: ur
2 hafa Frontpage sjálfur                             sko A: já A: já af því að annars getur ekkert sko þú getur sett þau inn á síðuna þína B: jájá
3 rosalega klár því hann var ofsalega montinn        sko yfir því að hún hafi komist inn og hérna en hann ætlar bara að f
4 dauð C: það er hætt við því D: en ég er            sko þú veist en ég hugsaði með mér ég ætla ekki að vera svo lengi í R
5 bundið A: tímabundið út af því að nú þurfa þau     sko B: losa þau þau þurfa að fara þau ætla ekki að vera hér sko þ
6 svona fílingur þarna (.) út af því að ég get ekkert sko ég á mitt torg B: m A: og ég veit ekkert hvernig ég á að þú veist sé
7 úr vinnunni A: út af því að hann var langur        sko þess vegna komstu svona miklu í verk ((hlær)) B: hann var svo ros
8 leitar fyrst í því B: ((tungusmellur)) (meta tækið) sko já að það var það sem ég var að experimenta með þú veist alv
9 í munninum alveg frá því að ég smakkaði hana       sko B: ((hlær við)) B: já A: hún gerði tvær sko og það var önnur sem va
10 eða B: nei A: nokkuð í því B: en í rauninni       sko er hefur þetta hálfa starf ekki verið kennslan hún hefur verið aukavin
11 en það er búið að breyta því núna er bara kennt   sko þú þarft að þekkja mismunandi framburð eftir landshlutum D
12 sem sagt maður hefði ekki neitað því upp á upp á sko þetta hefði kitlað mann að gera þetta C: já C: en það hérna HarpaI
13 verður sjálfsagt nálægt því eða gæti orðið það   sko B: takt þú þetta blað hérna (.) ef að hann fer að einhver hérna eitthv
14 barnabætur eða hvað B: af því að hún er orðin     sko komin yfir þrítugt A: já B: hún er ómenntuð (.) og með (fj-) fjögur
15 ((sýgur upp í nefið)) af því að það (x) alveg     sko bara skari og mjög slæmt ((sýgur upp í nefið)) að skíða krakkarnir
16 A: það er líka af því að það er röddunin          sko af því að í þýsku og dönsku að þá er raddað ef það kemur hérna n
17 ef að það eru einhverjar líkur á því að það standist ekki sko (.) og og alls ekki í skammtíma huggun eða einhverju svoleiðis
18 hafði verið langt á milli þeirra því að <!A6> (x) saman sko en Einar Jóhann hafði leyft sér að ((dyr opnast, blástur hættir)) að k
19 aftur í hana draslið A: nei ég nenni því ekki (.) alls ekki sko (.) enda þarf ég á þessu litla sem að er í (.) ((skark)) vélinni að hald
20 þurftu auðvitað að vera oft að þvo D: nei náttúrulega líka sko því þetta þetta er gamalt fólk þannig að þau þvo ekki eins mikið og
21 og svo kannski hékk þvotturinn upp á snúru í     sko tólf daga D1: (oj) D2: kominn skítafýla af honum D3: já D: já end
22 veist að (.) vilji er fyrir= D: =ja þið sjáið það alla vega sko alla vega sé ég það núna að (.) sko við verðum að fá niunda bekk til okkar
23 D: já sko það verður náttúrlega ekki í níunda bekk sko það er ekki velja sér ekki náttúrufræðibraut núna sko A: en þetta er þetta e
24 er eitthvað sem þau horfa fram eru standa frammi fyrir sko þegar þau koma í tíunda bekk= D: =koma í tíunda <A>bekk sko</A> A: fyrir</D>
25 að (.) <C>(.) fannst þetta</C> alveg (fár-) fáránlegt sko = D: m C: þetta er (ó-) D: =já C: en þetta er alveg óhemjulega vitlaust að lá
26 ja það er það er spáð leiðindaveðri núna á páskunum sko A: nú er það B: já A: já (.) B: <A>gæti</A> orðið einhver (norð-) norðannepj
27 því að það var hérna bara svört slydda í B: (x) B: sko eftir matinn í gær <!A> þá hérna eða svona um (nónleyt-) nónleytið að þá hér
28 svarta B: en svo aftur um (sv-) svona um fimm sexleytið sko þá var orðið fínasta veður hérna A: já (.) það var eins og í morgun þá var s (istal2)
29 A: já (.) B: og svona lágrenningur og og yfir dalinn sko A: já ((gespar)) (.) jájá B: þannig er nú málið (.) A: heimsækirðu aldrei M (
30 nei A: nú var hún í órétti sem sagt B: jájá bíllinn kemur sko hann heldur áfram og bíllinn kom bara beint A: jájá B: beint á (.) A: já það
```

4.3 Talmálsgögn og orðabókagerð

Af umfjölluninni í kafla 4.2 og dæmunum sem þar eru rakin má ljóst vera að greiður aðgangur að talmálsefni, t.d. sem hluta af almennri málheild, kæmi að góðu gagni við orðabókagerð. Augljósust er nytsemd slíks efnis þegar í hlut eiga orð sem eru fyrst og fremst notuð í daglegu tali en síður í ritmáli eins og þau sem skoðuð voru hér að framan. Fjölbreytt talmálsgögn sem sýna raunverulega málnotkun við mismunandi aðstæður eru grundvallarheimild við greiningu á merkingu og hlutverki slíkra orða í málinu og þau eru einnig mikilvæg uppspretta raunverulegra notkunardæma eða sem fyrirmynd við samningu tilbúinna dæma sem birt eru til skýringar og stuðnings við orðabókarlýsinguna.

Enda þótt orð séu notuð jafnt í töluðu og rituðu máli má vera að merking sumra þeirra og notkun sé ekki að öllu leyti eins í tali og riti. Eitt dæmi um slíkan mun er sérstakt hlutverk boðháttarmyndarinnar *heyrðu* í talmáli sem fjallað var um í síðasta kafla. Með notkun textasafna eða málheilda með fjölbreytilegum textum úr talmáli jafnt og ritmáli má leiða í ljós hvers kyns tilbrigði í notkun og merkingu einstakra orða og láta dæmi úr textunum styðja greiningu og lýsingu orðanna. Textasöfn hafa ekki síst sannað gildi sitt í því að þau gefa góða mynd af dæmigerðu samhengi orða, ekki síst orðastæðum og ýmiss konar föstum orðasamböndum (t.d. *alveg rosalega* + *lo./ao.*, sbr. 4.2), og þar gæti verið munur í talmáli og ritmáli.

Sé stuðst við nægilega stóra og fjölbreytilega málheild gefur hún líka mikilvæga vitneskju um tíðni orða og orðasambanda og útbreiðslu þeirra í mismunandi textum eða við ólíkar aðstæður. Slíkar upplýsingar koma að gagni við efnisval, t.d við val á flettiorðum, merkingarafbrigðum, orðasamböndum og notkunardæmum. Þær geta líka verið til mikils stuðnings við efnisskipan í orðabókum, ekki síst í innskipan einstakra orðsgreina þar sem því almennasta og algengasta er gjarnan skipað fremst. Þá geta upplýsingar um tíðni og útbreiðslu orðið grundvöllur að traustari ábendingum um notkunarsvið og stílgildi orða og orðasambanda.

Þeim orðabókaverkum sem eru gefin út í rafrænu formi fer fjölgandi. Það form gefur nýja möguleika í vali á upplýsingaþáttum og framsetningu þeirra og auk þess eru slíkum verkum ekki settar jafn þröngar skorður um samþjöppun efnisins og prentuðum orðabókum. Þar er því t.d. rúm fyrir mun fleiri notkunardæmi og ekki eru jafnríkar kröfur um að takmarka lengd þeirra. Eins og áður er nefnt má finna fyrirmyndir að góðum skýringardæmum í textasöfnum og málheildum en þangað má einnig sækja raunveruleg notkunardæmi til birtingar í orðabókum, hvort sem er úr töluðu eða rituðu máli. Í veforðabókum er jafnvel hægt að hafa beina tengingu úr orðabókargreinunum í málheild eða textasafn þannig að notendur geti sjálfir leitað að dæmum til viðbótar þeim sem birt eru í sjálfri orðabókinni.

Eitt einkenni margra rafrænna orðabóka er að þar er framburður orða og orðasambanda gefinn með hljóðdæmum sem bætast við eða koma í stað hefðbundinnar hljóðritunar (sjá t.d. nýlegar enskar orðabækur, MED og MEDO). Gagnasöfn með raunverulegu töluðu máli geta þjónað vel sem fyrirmynd að framburðardæmum og stuðlað að því að sá framburður sem gefinn er í orðabókum sé í samræmi við það sem gerist í eðlilegu talmáli. Þetta krefst þess að þannig sé búið um gögnin að greiður aðgangur sé að hljóðskránum og að auðvelt sé að finna viðeigandi dæmi í þeim. Ef upptökugæði á talmálsefninu leyfa það má jafnvel hugsa sér að hljóðdæmi til „birtingar“ í rafrænni orðabók séu sótt beint í gagnasafn.

5 Samantekt og niðurlag

Hér hefur verið fjallað um stór rafræn textasöfn og málheildir sem ætlaðar eru til almennra nota í margvíslegum hagnýtum og fræðilegum

verkefnum. Einkum var rætt um samsetningu slíkra málsafna með tilliti til þess að þau gefi nægilega góða mynd af málnotkun innan þess ramma sem þeim er settur. Áhersla var ekki síst lögð á hlut talmáls í slíkum söfnum. Jafnframt var gerð grein fyrir ýmsum atriðum sem valda því að það er flóknara og tímafrekara að afla nothæfs efnis úr talmáli en að safna ritmálstextum. Þá var ávinningurinn af því að nota talmálsefni og gefa því rúm í málsöfnum skoðaður með hliðsjón af orðabókarlýsingum á orðum sem eru fyrst og fremst notuð í talmáli.

Sú krafa er gerð til rafrænna málsafna sem ætlað er víðtækt hlutverk í rannsóknum og ýmsum hagnýtum verkum að þau endurspegli raunverulega málnotkun, t.d. á tilteknum tíma eða á ákveðnu notkunarsviði. Einnig er ljóst að stærð slíkra safna getur skipt verulegu máli og rannsóknir á ýmsum stærri einingum málsins, t.d. orðaforða og setningagerð, krefjast mjög stórra málsafna og það gildir ekki síður um ýmis fyrirbæri sem eru tiltölulega sjaldgæf í málinu. Það er því augljóslega hagkvæmt að setja saman eitt öflugt málsafn fyrir íslensku sem er öllum opið og nýtanlegt til margvíslegra verkefna líkt og nú er verið að gera með *Markaðri íslenskri málheild* (MÍM). Þá er hægt að byggja upp þekkingu, reynslu og aðstöðu á einum stað og þróa þar aðferðir við greiningu og nýtingu safnsins (sjá nánar um greiningar- og leitaraðferðir í greinum Sigrúnar Helgadóttur (2007) og Eiríks Rögnvaldssonar (2007) í þessu hefti).

Annað mikilvægt atriði varðar viðhald og eflingu slíkra safna. Málheild sem er ætlað að endurspegla samtímamálið úreldist fljótt. Í MÍM verða t.d. textar frá árunum 2000–2006, og til þess að málheildin haldi gildi sínu sem heimild um íslenskt samtímamál er nauðsynlegt að bæta við textum þegar fram líða stundir. Því væri æskilegt að skipuleggja slíkt safn til framtíðar sem gagnabanka þar sem bæði mætti leggja inn og taka út. Þótt miðlæg málheild sé til og öllum opin getur hún aldrei svarað öllum þörfum sem fram kunna að koma og ýmis sérhæfð verkefni munu eftir sem áður kalla á sjálfstæða efnisöflun. Ef þannig væri búið um hnútana að hægt væri að taka við slíku efni og bæta því við þann efnivið sem fyrir er væri það ein leið til að halda safninu við. Gera má ráð fyrir að margir textar sem þannig bærust væru einmitt textar sem talsvert hefur verið haft fyrir að safna og vinna úr. Þar má nefna ýmiss konar talmálsefni, óútgefið efni af ýmsu tagi og eldri texta sem ekki voru áður til í rafrænu formi. Slíkt efni er að sjálfsögðu mjög eftirsóknarverð viðbót við safnið.

Ef hvetja ætti fræðimenn og aðra til að leggja efni sem þeir safna inn í slíkan gagnabanka þyrftu að liggja fyrir skýrar og aðgengilegar reglur um frágang gagnanna og formið sem þau þurfa að vera í til að þau nýtist innan málheildarinnar. Ávinningur einstakra „viðskiptavina“ gæti falist í því að þeir fengju gögnin sín greind með þeim tólum sem þróuð hefðu verið í tengslum við bankann þannig að þau nýttust þeim sjálfum betur. Samvinna MÍM og tilbrigðaverkefnisins um öflun og úrvinnslu talmálsefnis er einmitt á þessum nótum.

Heimildir

Ásta Svavarsdóttir. 2003. Ordbogen og den daglige tale. Om den islandske talesprogsbank (ISTAL) og dens betydning i ordbogsredaktion. Í: Hansen, Zakaris Svabo, og Anfinnur Johansen (ritstj.). *Nordiske studier i leksikografi* 6, bls. 43–48. Tórshavn: Nordisk forening i leksikografi, Nordisk sprogråd og Fróðskaparsetur Föroya.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

BNC = *British National Corpus*. Vefsetur: *http://www.natcorp.ox.ac.uk* (10. nóvember 2006).

Burnard, Lou (ritstj.). 2000. *Reference Guide for the British National Corpus (World Edition)*. *http://www.natcorp.ox.ac.uk/docs/userManual* (10. nóvember 2006).

Eiríkur Rögnvaldsson. 1990. Orðstöðulykill Íslendinga sagna. *Skáldskaparmál* 1:54–61.

Eiríkur Rögnvaldsson 1994–5. Breytileg orðaröð í sagnlið. *Íslenskt mál og almenn málfræði* 16–17:27–66.

Eiríkur Rögnvaldsson. 1996. Brugen af et gammelislandsk tekstkorpus i leksikografisk arbejde. *LexicoNordica* 3:19–34.

Eiríkur Rögnvaldsson. 2002. *ÞAÐ* í fornu máli — og síðar. *Íslenskt mál og almenn málfræði* 24:7–30.

Eiríkur Rögnvaldsson. 2007. Textasöfn og setningagerð: Greining og leit. *Orð og tunga* 9 (þetta hefti).

Feagin, Crawford. 2002. Entering the Community: Fieldwork. Í: Chambers, J.K., Peter Trudgill & Natalie Schilling-Estes (ritstj.), *The Handbook of Language Variation and Change*, bls. 20–39. Oxford: Blackwell Publishing.

Finegan, Edward, & Douglas Biber. 2001. Register variation and social dialect variation: the Register Axiom. Í: Eckert, Penelope, & John R. Rickford (ritstj.), *Style and Sociolinguistic Variation*, bls. 235–267.

Gagnasafn Morgunblaðsins. Vefsetur: *http://www.mbl.is/mm/gagnasafn* (12. janúar 2007).

Helga Hilmisdóttir & Camilla Wide. 1999. sko — en mångfunktionell diskurspartikel i isländskt ungdomsspråk. Í: Kotsinas, Ulla-Britt, Anna-Brita Stenström & Eli-

Marie Drange (ritstj.). *Ungdom, språk og identitet*. Rapport fra et nettverksmöte, bls. 101–121. Nord 1999: 30. Kaupmannahöfn: Nordisk Ministerråd.

ÍO-1963 = *Íslensk orðabók handa skólum og almenningi.* Ritstj. Árni Böðvarsson. Reykjavík: Bókaútgáfa Menningarsjóðs.

ÍO-1983 = *Íslensk orðabók handa skólum og almenningi*. 2. útgáfa aukin og endurbætt. Ritstj. Árni Böðvarsson. Reykjavík: Bókaútgáfa Menningarsjóðs.

ÍO-2002 = *Íslensk orðabók.* 3. útgáfa. Ritstj. Mörður Árnason. Reykjavík: Edda.

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

Íslendinga sögur. Orðstöðulykill og texti. (1996) Geisladiskur og handbók. Ritstjórar orðstöðulykils: Bergljót S. Kristjánsdóttir, Eiríkur Rögnvaldsson (aðalritstjóri), Guðrún Ingólfsdóttir og Örnólfur Thorsson. Ritstjórar texta: Bragi Halldórsson, Jón Torfason, Sverrir Tómasson og Örnólfur Thorsson. Reykjavík: Mál og menning.

Lagasafn. Á vefsetri Alþingis: *http://www.althingi.is/vefur/lagasafn.html* (12. janúar 2007).

Landau, Sidney I. 2001. *Dictionaries. The Art and Craft of Lexicography*. 2. útgáfa. Cambridge: Cambridge University Press.

MED = *Macmillan English Dictionary for Advanced Learners*. 2002. (Rafræn útgáfa á geisladiski fylgir bókinni.) Oxford: Macmillan.

MEDO = *Macmillan English Dictionary Online*: *http://www.macmillandictionary.com/online* (15. janúar 2007).

Renouf, Antoinette. 1987. Corpus Development. Í: Sinclair, J.M. (ritstj.), *Looking up*, bls. 1–40.

Sagnalykill. Vefbækur Eddu: *http://edda.is/vefbaekur* (10. nóvember 2006).

Sigrún Helgadóttir. 2004a. Mörkuð íslensk málheild. Í: *Samspil tungu og tækni,* bls. 65–71. Reykjavík: Menntamálaráðuneytið.

Sigrún Helgadóttir. 2004b. Markari fyrir íslenska texta. Í: *Samspil tungu og tækni*, bls. 55–64. Reykjavík: Menntamálaráðuneytið.

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga* 9 (þetta hefti).

Sinclair, J.M. (ritstj.). 1987. *Looking up. An account of the* COBUILD *Project in lexical computing*. London/Glasgow: Collins.

Teubert, Wolfgang. 2004. Language and corpus linguistics. Í: Halliday, M.A.K., W. Teubert, C. Yallop & A. Čermáková. *Lexicology and Corpus Linguistics*, bls. 73–112. London/New York: Continuum.

Textasafn Orðabókar Háskólans á vefsetri OH: *http://www.lexis.hi.is* (10. nóvember 2006).

Teubert, Wolfgang, & Anna Čermáková. 2004. Directions in corpus linguistics. Í: Halliday, M.A.K., W. Teubert, C. Yallop & A. Čermáková. *Lexicology and Corpus Linguistics*, bls. 113–165. London/New York: Continuum.

Þórunn Blöndal. 2002/2006. „...og við alveg bara ókei...“ Vangaveltur um tíðni og hlutverk ókei í íslensku talmáli. *Fátt mun ljótt á Baldri Sigurðssyni fimmtugum,* bls. 78-84. Ritstjórar: Sigurður Konráðsson og Baldur Hafstað. Reykjavík. / *Skíma* 2006(2):17–20 (örlítið stytt gerð).

Þórunn Blöndal. 2005. *Lifandi mál*. Inngangur að orðræðu- og samtalsgreiningu. Reykjavík: Rannsóknarstofnun Kennaraháskóla Íslands.

Lykilorð:

málheild, talmál, orðabækur

Keywords:

corpus, spoken language, dictionaries

Abstract

The topic of this article is the design of text archives and corpora for general use in various theoretical and practical projects. It is emphasized that the spoken language should be represented in such text collections and the advantage of including speech data as well as written texts is discussed, considering some examples of dictionary entries which describe words that have been shown to belong mainly to the spoken language. With respect to the extra work in obtaining speech data compared to written texts the author urges those who collect such material for their projects to share them with others and that the data is gathered in a central corpus, which is open and accessible for research and development in theoretical and applied linguistics, lexicography and language technology.

Ásta Svavarsdóttir
Stofnun Árna Magnússonar í íslenskum fræðum / Háskóla Íslands
Neshaga 16
IS-107 Reykjavík
asta@lexis.hi.is

Eiríkur Rögnvaldsson

Textasöfn og setningagerð: greining og leit

1 Inngangur

Rannsóknir á flestum sviðum málvísinda byggjast með einum eða öðrum hætti á máldæmum. Setningafræðilegar rannsóknir eru þar engin undantekning. Það er hins vegar mjög misjafnt hvernig þessi dæmi eru fengin. Sumir setningafræðingar búa dæmi sín til sjálfir, en aðrir safna þeim úr rituðu eða töluðu máli. Til skamms tíma endurspegluðust mismunandi grundvallarviðhorf til viðfangsefnisins í þessum mismunandi uppsprettum dæma, en nú eru mörkin þarna á milli að verða óskýrari.

Þessi grein fjallar um setningafræðileg dæmi og leit að þeim í textasöfnum. Í byrjun er fjallað lítillega um texta og textasöfn sem setningafræðilegar heimildir. Bent er á að undanfarna áratugi hafa verið uppi mjög mismunandi viðhorf til gildis textasafna í setningafræðilegri umræðu og röksemdafærslu, en þann ágreining má að verulegu leyti rekja til mismunandi skoðana á því hvert viðfangsefni málfræðinnar sé; hvort hún eigi að lýsa málinu sjálfu (málbeitingunni) eða málkerfi málnotenda (málhæfninni). Einnig er fjallað nokkuð um margvíslegan vanda við túlkun þeirra upplýsinga sem textasöfn veita – ekki síst túlkun á þögn textanna um tilteknar setningagerðir.

Meginhluti greinarinnar fjallar um möguleika á setningafræðilegri dæmaleit í mismunandi greindum textasöfnum; hráum texta án nokk-

Orð og tunga 9 (2007), 51–73.

urra sérmerkinga, texta með beygingarlegri greiningu, og texta þar sem helstu setningarliðir og setningafræðileg hlutverk hafa verið greind. Sagt er frá tilraunum til að safna dæmum um tilteknar setningagerðir í íslenskum textasöfnum; einkum grunnskrám *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) og ÍSTAL-safninu (sjá Þórunni Blöndal 2005 og grein Ástu Svavarsdóttur (2007) í þessu hefti). Við slíka dæmasöfnun væri æskilegt að hafa aðgang að textum sem hafa verið greindir setningafræðilega, en engin söfn slíkra texta íslenskra eru til.

Nú er hins vegar hægt að greina íslenska texta beygingarlega (marka þá) á vélrænan hátt (sjá grein Sigrúnar Helgadóttur (2007) í þessu hefti). Í ljósi þess að verulegar setningafræðilegar upplýsingar felast í hinum málfræðilegu mörkum þótti ómaksins vert að athuga hvort hægt væri að nýta þau í setningafræðilegri leit. Í ljós kom að beygingarlega mörkunin gagnast mjög vel við leit að ýmsum setningagerðum, og gerir setningafræðilega dæmasöfnun margfalt fljótlegri og markvissari en áður. Í lokin er drepið stuttlega á ólokið verkefni þar sem unnið er að vélrænni hlutaþáttun (setningafræðilegri greiningu) sem gæti auðveldað setningafræðilega dæmaleit að mun.

2 Textasöfn sem heimild um setningagerð

Á fyrri hluta 20. aldar var talsvert gert af því í ýmsum löndum að nota stór textasöfn sem grundvöll mállýsinga. En viðhorfin til textasafna og megindlegra athugana á tungumáli gerbreyttust svo að segja á einni nóttu fyrir hálfri öld, eins og McEnery og Wilson benda á (1996:4):

> [...] we can pinpoint a discontinuity in the development of corpus linguistics fairly accurately in the late 1950s. After this period the corpus as a source of data underwent a period of almost total unpopularity and neglect. Indeed it is no exaggeration to suggest that as a methodology it was widely perceived as being intellectually discredited for a time.

Það sem olli þessum straumhvörfum var tilkoma málkunnáttufræði (generatífrar málfræði) Chomskys (1957). Chomsky gaf lítið fyrir gildi textaathugana en byggði málfræði sína þess í stað á máltilfinningu og

dómum málnotenda. Áhrif Chomskys voru mjög mikil og næstu áratugina þótti fæstum setningafræðingum ástæða til að leggjast í dæmasöfnun úr textum til að rökstyðja kenningar sínar, heldur bjuggu sjálfir til dæmi sín og dæmdu þau tæk eða ótæk. Þessi aðferð þykir vissulega enn góð og gild, en á seinni árum hafa menn aftur horfið til dæmasöfnunar úr textum og láta aðferðirnar vinna saman og bæta hvora aðra upp. Í þessum kafla er drepið á nokkrar forsendur þess ágreinings sem hefur verið um gildi textadæma og rætt sérstaklega um það hvernig mismunandi fræðilegar forsendur geta leitt til mismunandi túlkunar þess vitnisburðar sem textarnir gefa.

2.1 Heimild um málhæfni eða málbeitingu?

Ein meginröksemd Chomskys fyrir því að textasöfn væru gagnslaus í setningafræðilegri greiningu og röksemdafærslu var sú að þau væru ævinlega og óhjákvæmilega takmörkuð, endanleg, og tilviljanakennd (sjá t.d. Chomsky 1957:13-17). Auðvelt er t.d. að tilfæra ýmis dæmi um setningar og setningagerðir sem sjaldan eða aldrei finnast í textasöfnum, jafnvel mjög stórum, en málhöfum ber þó saman um að séu tækar. Þetta hefur oft verið notað sem rök fyrir því að málhæfnin sé að verulegu leyti meðfædd; menn geti ekki hafa lært slíkar setningar af öðrum, heldur hljóti að hafa einhverja meðfædda þekkingu á þeim reglum sem um þær gilda.

Skiptar skoðanir um þessi mál leiddu til hálfgerðs stríðs milli málkunnáttufræðinga (generatífista) og þeirra sem fengust við gagnamálfræði (corpus linguistics). Chomsky talaði víða óvirðulega um gagnamálfræði, og í ritum gagnamálfræðinga er að finna mörg og beitt skot á Chomsky og fylgismenn hans (sjá um þetta t.d. McEnery og Wilson 1996:4-17, 61-66 o.v.). Hér er þó rétt að halda því til haga að þarna er að verulegu leyti um sýndarágreining að ræða – meðvitað eða ómeðvitað. Menn voru nefnilega ekki að tala um sama hlutinn. Chomsky var að tala um málhæfni (competence) en gagnamálfræðingar skoða málbeitingu (performance) (sjá t.d. Chomsky 1965:4). Chomsky var sem sé að tala um málfræðina, málkerfið, en gagnamálfræðingar skoða afurð kerfisins – málið sjálft. Þarna á milli er flókin víxlverkun sem ekki hefur verið kortlögð til fulls, en meginatriðið er að báðar aðferðirnar eiga fullan rétt á sér og eru nauðsynlegar – en þær svara mismunandi spurningum.

Textasöfn eru þannig gagnleg til að finna ýmsar setningagerðir og átta sig á þeim. Það er t.d. hægt að nota þau, að vissu marki, til að úrskurða tiltekna setningagerð tæka. Það er hins vegar ekki hægt að nota þau til að úrskurða setningagerð ótæka. Þótt hún komi ekki fyrir í þeim textum sem við skoðum getur það verið tilviljun. Eðli málsins samkvæmt getur textasafn okkar aldrei innihaldið allar hugsanlegar setningar. Ef við erum að lýsa málinu (ekki málkerfinu) gerir þetta ekkert til. Textasafnið sem við höfum undir afmarkar þá viðfangsefni okkar, og ef tiltekin setningagerð kemur ekki fyrir í safninu er hún ekki hluti viðfangsefnisins og kemur okkur þess vegna ekkert við.

En ef við erum að lýsa málkerfinu sjálfu horfir málið öðruvísi við. Það málkerfi sem við lýsum á að gera okkur kleift að mynda allar málfræðilega tækar setningar en ekki aðrar. Þess vegna nægir okkur ekki að vita hvers konar setningar eru tækar – við þurfum líka að vita hvers konar setningar væru ótækar. Og því svarar textasafnið ekki – það er vitaskuld ekki hægt að takmarka mengið „tækar setningar“ við þær setningar sem fyrir koma í tilteknu safni, hversu stórt sem það er; „it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in his field work“, segir Chomsky (1957:15) og hnykkir enn á því síðar:

> [...] though “probability of a sentence (type)” is clear and well defined, it is an utterly useless notion, since almost all highly acceptable sentences (in the intuitive sense) will have probabilities empirically indistinguishable from zero and will belong to sentence types with probabilities empirically indistinguishable from zero. Thus the acceptable or grammatical sentences (or sentence types) are no more likely, in any objective sense of this word, than the others (Chomsky 1965:195).

2.2 Hversu marktækir eru textarnir?

Í samtímalegri setningafræði er hægt að snúa sig út úr þessum vanda með þeim einfalda hætti að spyrja málnotendur. Þá erum við ekki háð afmörkuðu textamengi, heldur getum búið til texta eftir þörfum, ef svo má segja, og borið þá undir málnotendur og fengið dóma þeirra

um hvort tiltekin setning sé tæk eða ekki. Þeir sem fást við sögulega setningafræði eiga aftur á móti ekki þessa útleið – þeir verða að reiða sig algerlega á textana (sjá umræðu um þetta hjá Eiríki Rögnvaldssyni 1998). Stundum hafa menn reynt að bæta sér það upp með einhverjum ráðum, eins og t.d. því sem kallað hefur verið „lögmál ónýttra tækifæra" (Principle of missed opportunities) og er orðað svo:

(1) If a certain syntactic form is used regularly in a given function or type of context C in a living language L, and if F is absent in C at an earlier stage of the language, OL, then there is good reason to assume that F does not exist in OL (Faarlund 1990:17).

Þetta getur þó aðeins verið viðmið sem verður að beita af mikilli varfærni. Hvernig getum við t.d. fullyrt að eitthvað sé „absent [...] at an earlier stage of the language" – hvernig skilgreinum við „language" þarna? Við höfum ekki annað til að miða við en þá texta sem varðveittir eru eða við höfum aðgang að – og þeir eru ekki alltaf miklir. En við þurfum líka að gæta þess að skoða þá alla áður en við fullyrðum nokkuð, gæta þess að ekki komi eitthvað annað til sem geti valdið því að viðkomandi setningagerð finnst ekki á eldra málstigi, o.s.frv.

Hér á undan var sagt að hægt væri – að vissu marki – að nota textasöfn til að úrskurða tiltekna setningagerð tæka. En þar verður líka að hafa fyrirvara. Því fer nefnilega fjarri að allar setningar sem koma fyrir í textasöfnum séu tækar í raun og veru, þ.e. samræmist málkerfi flestra málnotenda. McEnery og Wilson (1996:13) hafa eftir Chomsky að allt að 95% allra segða (utterances) séu í raun málfræðilega ótækar. Þar er væntanlega miðað við talmál og hlutfallið örugglega mun lægra í rituðu máli – og McEnery og Wilson vilja líka meina að tala Chomskys sé alltof há. En jafnvel þótt við lítum framhjá mállýskumun er ljóst að í rituðu máli er nokkuð um setningar sem flestir myndu telja ótækar – setningar þar sem fyrir koma ýmiss konar pennaglöp, mistök í ritvinnslu, prent- og ásláttarvillur, einstaklingsbundið málfar, o.s.frv. Dæmi um þetta má sjá í eftirfarandi setningu af mbl.is:

(2) Alls voru um 179 tonn af hvalaúrgangi af þeim sem sjö langreyðum sem veiddust við landið í haust urðað að Fífilholtum á Mýrum. [Undirstrikun mín, ER]
http://www.mbl.is/mm/frettir/frett.html?nid=1245425

Þarna eru – að flestra mati, geri ég ráð fyrir – tvær villur; *sem* ofaukið og sambeygingu vantar. Þegar við lesum leiðréttum við þetta með

sjálfum okkur þegjandi og hljóðalaust, í samræmi við málkennd okkar. En hvenær getum við leyft okkur það? Hvað með þá sem fást við eldri málstig og geta ekki beitt eigin málkennd á textana? Verðum við að líta á allar setningar sem finnast í textum sem jafnréttháar? Yfirleitt gera menn það ekki í raun; „one must be ready to characterize certain unattested sentences as well-formed and some attested sentences as ill-formed“, segir Lightfoot (1979:6.) En þarna eru menn vissulega á hálum ís, og oft getur verið freisting að láta fræðikenningar taka af sér ráðin; hafna setningum sem koma fyrir ef þær falla ekki að þeirri kenningu sem maður vinnur með, en gera ráð fyrir öðrum sem ekki finnast dæmi um, vegna þess að kenningin segir að þær ættu að geta komið fyrir.

2.3 Ályktanir af þögn textanna

Gott dæmi um það hvernig mismunandi fræðileg afstaða kemur fram í mismunandi túlkun á þögn textanna má taka úr lífseigri deilu um það hvort aukafallsfrumlög hafi verið til í fornu máli. Flestir málfræðingar fallast núorðið á að nútímaíslenska hafi aukafallsfrumlög, en margir halda því fram að tilkoma þeirra sé sérískensk þróun og samsvarandi liðir hafi ekki haft stöðu frumlags í fornu máli. Það hefur m.a. verið rökstutt með því að í fornmáli komi viðkomandi aukafallsliðir ekki fyrir í öllum sömu setningagerðum og í nútímamáli, t.d. ekki í setningum af þessu tagi:

(3) a. Ég vonast til að vanta ekki efni í ritgerðina. (Höskuldur Þráinsson 1979)
b. Hann vonast til að leiðast ekki. (Halldór Ármann Sigurðsson 1989)

Það skiptir að margra mati verulegu máli hvort einhver fornmálsdæmi finnist um slíkar setningar; Falk (1995:203) nefnir t.d. þessa setningagerð sem fullnaðarsönnun fyrir því að nútímaíslenska hafi aukafallsfrumlög. Mørck (1992) segist hafa leitað sérstaklega að dæmum á við (3) í fornum textum, en án árangurs, „så jeg meiner at vi får holde fast ved at lik-NP-stryking bare virker på nominativledd, inntil noe anna kan dokumenteres“ (Mørck (1992:71). Ég hef hins vegar bent á (Eiríkur Rögnvaldsson 1996) að setningar á við (3) eru mjög sjaldgæfar í nútímamáli, að því er virðist – finnast ekki einu sinni þótt leitað

sé með *Google* í öllu því textamagni sem er að finna á netinu, og er auðvitað margfalt það sem til er á fornískensku.

Hver er þá niðurstaðan? Ég fæ ekki betur séð en hér geti hver trúað því sem hann vill. Það væri vissulega betra fyrir mig ef dæmi á við (3) fyndust í fornu máli, en ég get samt haldið því fram – með vísun til þess sem haft er eftir Chomsky hér að framan – að við því sé alls ekki að búast að þau finnist, og fjarvera þeirra segi ekkert um það hvort þau hafi verið tæk að fornu. Þeir sem vilja hafna tilvist aukafallsfrumlaga í fornu máli geta líka sagt: Fyrst engin dæmi af þessu tagi finnast þá höfum við enga sönnun fyrir því að þessi setningagerð hafi verið tæk í fornu máli, og meðan við höfum enga slíka sönnun getum við ekki leyft okkur að gera ráð fyrir aukafallsfrumlögum.

2.4 Breytt viðhorf til textadæma

Viðhorf margra málfræðinga til texta og textasafna hefur breyst á seinni árum. Nú þykir miklu eðlilegra en fyrir fáum árum að rökstyðja setningafræðilegar greiningar með dæmum úr töluðu eða rituðu máli. Hrint hefur verið af stað viðamiklum rannsóknarverkefnum til að kanna setningafræðilegan mállýskumun og safna setningafræðilegum dæmum, s.s. norræna verkefninu *Scandinavian Dialect Syntax* (*http://uit.no/scandiasyn*) og „dótturverkefnum“ þess, m.a. íslenska verkefninu *Tilbrigði í setningagerð* sem Höskuldur Þráinsson stýrir (sjá Ástu Svavarsdóttur 2006). Þetta hefði tæpast getað gerst fyrir 20 árum eða svo.

Að hluta til má skýra þessa þróun með því að máldæmi, einkum ritmálsdæmi, eru orðin mun auðfengnari en áður. Með tilkomu sístækkandi rafrænna textasafna er nú orðið auðvelt að safna fjölbreyttum textadæmum af ýmsu tagi. Vefurinn hefur svo gert mönnum kleift að komast í margs konar texta sem áður voru óaðgengilegir og einnig hafa þar orðið til nýjar textategundir sem margar hverjar standa nær talmálinu en hefðbundnu ritmáli. Leitarvélar á vefnum, eins og *Google* og *Embla*, hafa svo auðveldað mönnum dæmasöfnun úr þessum textum.

En þessi þróun býður líka hættunni heim og það verður að fara varlega við notkun og túlkun þeirra dæma sem aflað er með leit á netinu. Þótt þar finnist setningagerð sem menn þekktu ekki áður þarf það ekki að tákna að hún sé ný í málinu – eins gæti verið að hún hafi

lengi verið til, en tilheyri hins vegar málsniði sem ekki hefur áður verið notað í (aðgengilegu) ritmáli. Það þarf líka að hafa í huga að dæmi á netinu eru ekki endilega úr nútímamáli. Það er talsvert af fornum textum á netinu (t.d. allar *Íslendingasögur*, *Heimskringla* og *Fornaldarsögur Norðurlanda* hjá Netútgáfunni, *http://www.snerpa.is/net*). Þegar ég var að skoða samband fornafnanna *sjálfur* og *sinn* í fyrra fann ég á netinu allnokkur dæmi um *sjálfrar sinnar*; en þegar að var gáð reyndust þau flest vera úr eldri textum.

Eins þarf að gæta þess að talsvert er af málfræðigreinum á netinu og í þeim eru stundum dæmi sem annaðhvort eru beinlínis ótæk og eiga að vera það, eða koma sjaldan fyrir í venjulegum textum og eru því ekki marktæk sem dæmi um málnotkun. Hér að framan var því haldið fram að jafnvel í leit á netinu með *Google* fyndust engin dæmi á við (3) en það er ekki alveg rétt; í raun finnur *Google* fjögur dæmi um sambandið *vonast til að vanta ekki* og tvö dæmi um *vonast til að leiðast ekki*. En þegar dæmin eru skoðuð kemur í ljós að þau eru öll úr dæmasetningum málfræðinga.

2.5 Textasöfn í tungutækni

Hér er ekki ætlunin að gera ítarlega úttekt á kostum þess og göllum að nota textasöfn í setningafræðirannsóknum. Enginn vafi er á því að textasöfn geta komið að miklu gagni á því sviði, en hitt er jafnljóst að þau svara ekki öllum spurningum og nauðsynlegt er að gæta varúðar í túlkun þeirra. En þegar litið er á gildi textasafna frá sjónarhóli tungutækni er viðhorfið annað. Þar er sjónarhornið hagnýtt fremur en fræðilegt – ekki verið að leita upplýsinga um málkerfið, heldur greina textana og vinna úr þeim upplýsingar sem síðan er hægt að nota til að „hanna eða útbúa einhvern hugbúnað eða tæki sem nýtist mönnum í starfi eða leik“, eins og segir í skilgreiningu orðsins *tungutækni* í *Orðabanka Íslenskrar málstöðvar* (*http://herdubreid.rhi.hi.is:1026/wordbank/-search*). Þá skiptir ekki endilega máli hvers vegna tiltekin setningagerð kemur ekki fyrir – hvort það er vegna þess að hún er sjaldgæf, eða vegna þess að hún sé alls ekki hugsanleg í málinu; málfræðilega ótæk. Ef hún kemur ekki fyrir í stóru textasafni er ekki líklegt að mörg dæmi um hana komi fyrir í öðru safni sambærilegra texta. Því er ekki líklegt að hugbúnaður okkar eða tól þurfi að glíma við hana, nema þá í mjög litlum mæli. Jafnvel þótt setningagerðin komi fyrir, og hugbún-

aður okkar greini hana rangt eða alls ekki, hefur það þá ákaflega lítil áhrif á heildarframmistöðu búnaðarins.

3 Leit að setningagerðum í textasöfnum

Það er til lítils að koma upp safni af textum úr töluðu og rituðu máli ef ekki eru til aðferðir til að vinna úr þessum söfnum. Það þarf að vera hægt að leita í þeim að dæmum um tilteknar setningagerðir. Við þá leit má beita tveimur ólíkum aðferðum. Önnur er sú að lesa textana frá upphafi til enda og skrá dæmi úr þeim. Ókostur aðferðarinnar er vitanlega sá að hún er mjög seinleg, auk þess sem alltaf er hætta á að dæmi fari fram hjá lesandanum. Til skamms tíma var þetta þó eina aðferðin sem völ var á, en það hefur breyst á síðustu 20-25 árum með tilkomu rafrænna texta. Það væri því mikill kostur ef hægt væri að leita að dæmum á skipulegan hátt í tölvu. Bæði væri slík leit mjög fljótleg, og eins ætti hún að geta verið tæmandi – sé leitað á réttan hátt. Forsendur fyrir slíkri leit eru einkum tvær; að til séu tölvutækir textar, og að þeir séu málfræðilega greindir á þann hátt að hægt sé að leita að setningafræðilegum fyrirbærum.

Í þessum kafla er fjallað um mismunandi aðferðir við setningafræðilega dæmaleit í textum; frá einfaldri textaleit yfir í leit í beygingarlega mörkuðum textum, og að lokum um leit í setningafræðilega mörkuðum textum. Sagt er frá nokkrum tilraunum sem ég hef gert til að nýta beygingarlega mörkun í setningafræðilegum tilgangi og hafa gefið góða raun.

3.1 Textaleit

Einfaldasta form leitar er það sem öll ritvinnsluforrit bjóða upp á; að slá inn streng (eitt orð eða fleiri) og leita að honum, nákvæmlega eins og hann er ritaður. Smávægileg tilbrigði eru möguleg (t.d. að tilgreina hvort hástafir og lágstafir skipta máli), og stundum er hægt að nota algildisstafi (e. *wildcard characters*) til að leita að hvaða staf sem er. Í *Word* finnur b^?r til dæmis *bar*, *ber*, *byr*, *bor*, *bær*, *býr* o.s.frv. Í UNIX-stýrikerfinu er hægt að nota reglulegar segðir (e. regular expressions) við leitina og tilgreina þannig flókin leitarmynstur. Þannig finnur [iy]n[gk][^jíeæ] strengina *ing*, *yng*, *ink* og *ynk*, en þó því aðeins að enginn stafanna *j*, *i*, *í*, *e*, *æ* komi næst á eftir. Ýmis sérhæfð texta-

vinnsluforrit bjóða upp á sérhæfðari möguleika. Í *WordCruncher* er t.d. hægt að leita að orðum sem koma fyrir nálægt hvort öðru, með tilgreindum hámarksstafafjölda á milli. Þar er líka hægt að hlaða orðum inn í bálka og leita að dæmum þar sem eitthvert orð úr öðrum bálkinum kemur fyrir í grennd við eitthvert orð úr hinum. Svona mætti lengi halda áfram að telja upp þá möguleika sem finnast í ýmsum forritum og auðvelda manni leitina.

Leit af þessu tagi er þó alltaf þeim takmörkunum háð að hún er bundin við orð. Það veldur því að erfitt er að nýta hana við leit að tilteknum setningagerðum – það er því aðeins hægt að unnt sé að tengja setningagerðir við ákveðin orð. Þannig er t.d. hægt að nýta slíka leit að vissu marki til að skoða afturbeygingu, með því að leita að myndum afturbeygða fornafnsins *sig/sér/sín*, og afturbeygða eignarfornafnsins *sinn/sín/sitt*. En jafnvel hér er þessi aðferð ófullnægjandi. Í fyrsta lagi vegna þess að sumar myndir afturbeygða fornafnsins og afturbeygða eignarfornafnsins falla saman við beygingarmyndir annarra orða (*sér* getur verið 3. pers. et. fh. nt. af *sjá*, og *sinn* getur verið hvorugkynsnafnorðið *sinn*). Því þarf að fara gegnum öll dæmin sem finnast við slíka leit og vinsa úr þeim. Það er seinlegt en ekki frágangssök. En til að fá góða mynd af notkun afturbeygingar þarf líka að skoða setningar þar sem afturbeygð fornöfn eru ekki notuð, heldur persónufornöfn. Þá vandast málið; því að persónufornöfn eru vitanlega notuð við miklu fjölbreyttari aðstæður. Þau eru svo algeng að það borgar sig ekki að leita að þeim; það væri svo mikið verk að vinsa úr leitarniðurstöðunum að það er alveg eins gott að lesa bara textann í heild.

Ég hef mikla reynslu af því að nota textaleit, einkum með hjálp *WordCruncher*, í setningafræðilegum rannsóknum á fornískenskum textum. Þeir textar sem ég vann með voru að mestu leyti ógreindir, þótt vissulega megi hafa nokkurt gagn af greiningunni í *Orðstöðulykli Íslendinga sagna* (Bergljót S. Kristjánsdóttir o.fl. 1996). Textaleitin hefur vissulega skilað ágætum árangri í mörgum tilvikum. Þannig hef ég t.d. leitað að dæmum um aukafallsfrumlög (Eiríkur Rögnvaldsson 1996) og boðháttarsagnir (Eiríkur Rögnvaldsson 2000). Í fyrra tilvikinu lá fyrir hverjar væru helstu sagnir sem kæmu til greina að tækju aukafallsfrumlög, og því var leitað að þeim; í því síðara var leitað að dæmum um boðháttarmyndir algengra sagna. Í báðum tilvikum dugði að finna (sem flest) dæmi; ekki var ætlunin að setja fram neina tölfræði á grundvelli þeirra, og því var aðferðin fullnægjandi.

En ég hef líka notað þessa aðferð til að skoða orðaröð í sagnlið og leita þar að tilteknum orðaraðarmynstrum (Eiríkur Rögnvaldsson 1994-95). Þar reyndi ég að sýna fram á að tiltekin mynstur kæmu fyrir innan flókinna sagnliða (einkum liða sem hafa að geyma tvær fallhátt-arsagnir og tvö andlög) en önnur ekki, og það væri kerfi í því hvað kæmi fyrir og hvað ekki. Í þessu tilviki var ekki einfalt að nota einstök orð í leitinni. Ég nýtti mér það að tveggja andlaga sagnir eru ekki óendanlega margar, og leitaði að dæmum um fallhætti eins margra þeirra og ég gat. Með því móti hafði ég fjölmörg dæmi upp úr krafsinu, og sú leit skilaði niðurstöðum sem ég þóttist sjá kerfi í.

En hvort sem ég hef nú rambað á rétta niðurstöðu í þessu tilviki eða ekki (ég hef ekki rekist á neitt síðan sem kollvarpi henni) þá er ljóst að þessi aðferð er ófullnægjandi. Ein ástæða er sú að hún er afskaplega seinleg; ég þurfti að slá inn margar sagnir og leita hvað eftir annað. Önnur ástæða er sú að mikil hætta er á villum; leitin skilar mörgum dæmum sem flest koma ekki málinu við og þegar farið er yfir þau má búast við að eitthvað fari fram hjá manni. Þriðja ástæðan er svo sú að þau orðaraðarmynstur sem ég fann engin dæmi um gæti verið að finna hjá einhverri sögn sem mér datt ekki í hug að leita að.

3.2 Trjábankar

Enn vandast málið ef við ætlum að leita að dæmum um setningagerðir á við kjarnafærslu andlags í aukasetningum, þ.e. dæmum þar sem andlag stendur næst á eftir aukatengingu, eins og í (4):

(4) Ég veit að þennan mann þekkir þú ekki.

Vissulega tengist þessi setningagerð afmörkuðum hópi orða, þ.e. aukatengingum, en vandinn er sá að dæmin um þær eru gífurlega mörg og ef þarf að skoða þau öll jafngildir það því að fara gegnum allan textann. Auðvitað er hægt að þrengja leitina með því að tilgreina tiltekin andlög, en þar með verður leitin líka mjög tilviljanakennd. Þarna dugir orðaleitin því ekki, heldur þyrftum við að geta leitað eftir setningafræðilegu hlutverki – leitað að *andlagi í upphafi aukasetningar*. En til að leita á þann hátt þyrftum við að hafa setningafræðilega greinda texta eða málheild.

Slíkar málheildir eru sums staðar til og mjög víða í smíðum um þessar mundir. Þær eru yfirleitt kallaðar *treebanks*, trjábankar, með vís-

un til þeirrar vel þekktu aðferðar að sýna setningafræðilega formgerð með trjám eða hríslum. Þessir trjábankar eru þó með ýmsu móti og því fer fjarri að í þeim öllum sé formgerð greind á sama hátt og gert er í hríslum. Í sumum trjábönkum (t.d. þeim búlgarska, sjá Osenova og Simov 2003) byggist greiningin á ákveðnu setningafræðilegri kenningakerfi – margir trjábankar sem nú eru í smíðum byggjast t.d. á HPGS, head-driven phrase structure grammar, eða hausastýrðri liðgerðarmálfræði ef við þýðum það á íslensku. Það er sama kenningakerfi og byggt var á í setningagreiningarverkefni því sem unnið var að hjá Friðriki Skúlasyni í nokkur ár (sjá Maren Albertsdóttur og Stefán Einar Stefánsson 2004). Aðrir trjábankar byggjast á venslamálfræði (dependency grammar), t.d. nýr danskur trjábanki (sjá Kromann 2003). Í enn öðrum trjábönkum er áhersla lögð á að hafa greininguna óháða setningafræðilegum teoríum, og þá verður hún yfirleitt ekki jafn smámunasöm (sjá t.d. Nivre 2002).

Ástæðan fyrir því að nú er víða verið að koma upp trjábönkum er sú að úr þeim má vinna mjög margvíslegar upplýsingar um setningagerð – upplýsingar sem ekki fást á annan hátt. Þessar upplýsingar eru ekki síst notaðar í ýmsum verkefnum innan tungutækni, s.s. við málfarsleiðréttingar, vélrænar þýðingar o.fl., en vitanlega nýtast þær einnig við setningafræðirannsóknir, í orðabókagerð o.s.frv. Vandinn er hins vegar sá að setningafræðileg greining samfelldra texta er mjög snúin og tímafrek og stofnkostnaður trjábanka því mjög hár. Reyndar var um tíma í gangi norrænt samstarfsnet um trjábanka, *Nordic Treebank Network*. Í því var gerð tilraun með setningafræðilega greiningu og samanburð á fyrstu köflunum úr *Veröld Soffíu* eftir Jostein Gaarder – Gunnar Hrafn Hrafnbjargarson sá um greiningu á íslenska textanum. En ekki er fyrirsjáanlegt neitt framhald á þessari tilraun hér á landi.

3.3 Setningafræðileg nýting málfræðilegrar mörkunar

3.3.1 Setningafræði í beygingargreiningunni

En þótt setningafræðilega greind íslensk málheild sé ekki til hefur mikill árangur náðst í málfræðilegri (þ.e., einkum beygingarlegri) greiningu íslenskra texta, eins og m.a. kemur fram í grein Sigrúnar Helgadóttur (2007) í þessu hefti. Hér má sjá dæmi um málsgrein sem búið er að marka.

(5) ég `fp1en` stökk `sfg1eþ` á `aa` eftir `aþ` strætó `nkeþ` og `c` veifaði `sfg1eþ` , , vagnstjórinn `nkeng` sá `sfg3eþ` mig `fp1eo` og `c` stoppaði `sfg3eþ`. . ég `fp1en` tautaði `sfg1eþ` takk `au` og `c` brosti `sfg1eþ` til `ae` hans `fpkee` um `ao` leið `nveo` og `c` ég `fp1en` lét `sfg1eþ` miðann `nkeog` detta `sng`. .

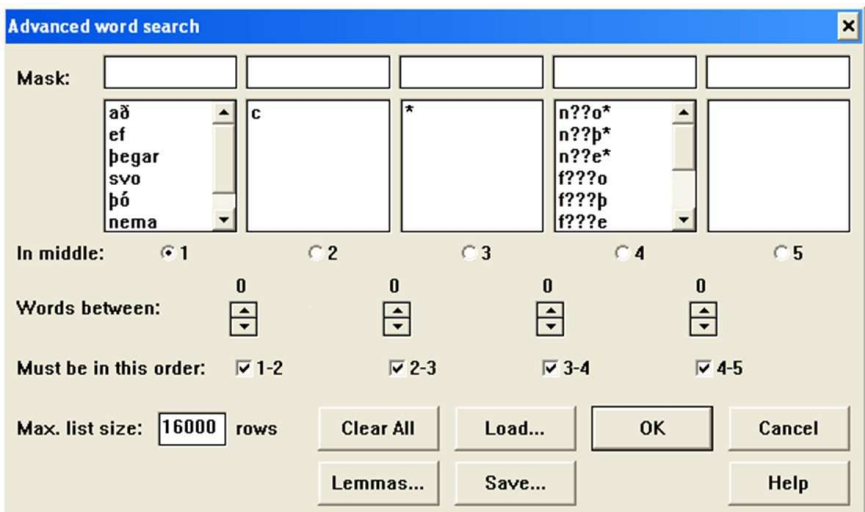
Á eftir hverju orði kemur mark þess eða greiningarstrengur. Hver stafur í strengnum stendur fyrir eitt málfræðilegt atriði. Fyrsti stafurinn stendur alltaf fyrir orðflokk – *f* er fornafn, *s* er sögn, *a* er atviksorð/forsetning, *n* er nafnorð *og* c er samtenging. Aðrir stafir tákna síðan greiningarþætti orðflokkanna. Í *stökk* táknar *s* þannig sögn, *f* framsöguhátt, *g* germynd, *1* 1. persónu, *e* eintölu, og *þ* þátíð. Í *strætó* táknar *n* nafnorð, *k* karlkyn, *e* eintölu og *þ* þágufall. Þótt beygingarlega greiningin taki eingöngu til eiginleika einstakra orða gefa greiningaratriðin mjög oft vísbendingar um vensl orða í setningu; íslensku greiningarstrengirnir gefa miklu meiri setningafræðilegar upplýsingar en þeir ensku t.d. Fallorð innan nafnliðar standa í sama kyni, tölu og falli; frumlag stendur í nefnifalli (nema með skilgreindum hópi sagna) en andlag í aukafalli; o.s.frv. Þess vegna kom sú hugmynd upp að athuga hvort og þá að hvaða marki hægt væri að láta málfræðilegu greiningunna koma í stað setningafræðilegrar greiningar.

Í dæminu hér að framan koma mörkin inn í textann og standa þar eins og hver önnur orð. Það getur oft komið sér vel því að iðulega er hægt að leita að tilteknum setningagerðum með því að tilgreina einhvers konar samband af orðum og greiningarstrengjum.

Ég hef prófað að nota forrit sem heitir *WinCord* til að vinna með þessa mörkuðu texta. Þetta er einfalt forrit sem hægt er að fá ókeypis á netinu, er mjög þægilegt í notkun og hefur gagnast mér vel. En vitanlega væri einnig hægt að nota fjölmörg önnur forrit af svipuðu tagi, eða nota mynsturleit með reglulegum segðum í UNIX eða öðrum tólum. *WinCord* býður upp á samsetta leit (*Advanced Word Search*), þar sem hægt er að slá orð inn í allt að fimm leitarreiti hvern á eftir öðrum, eins og sýnt er hér á eftir. Hér nær 'orð' einnig yfir greiningarstrengi, þar sem þeir koma inn í textann og forritið gerir engan mun á þeim og venjulegum orðum.

3.3.2 Leitað að kjarnafærslu í aukasetningum

Lítum nú á hvernig við förum að því að leita að dæmum um kjarnafærslu í aukasetningum með hjálp WinCord.



Hér höfum við sett sex dæmigerðar aukatengingar inn í fyrsta reitinn, en vitanlega væri hægt að hafa þær fleiri. Í næsta reit er svo mark þessara tenginga, *c*. Þetta er hvort tveggja nauðsynlegt. Ef við notum bara markið en sleppum orðunum fáum við líka allar aðaltengingar með og það viljum við ekki. Ef við notum bara orðin en sleppum markinu fáum við dæmi þar sem orðin í fyrsta reit eru annað en tengingar. Í þriðja reit er stjarna sem stendur fyrir hvað sem er. Það er vegna þess að við viljum ekki þurfa að tilgreina nein tiltekin orð þarna – bara greiningarstreng þeirra sem kemur í fjórða reit. Þar setjum við þrjú mynstur nafnorða, eitt fyrir hvert fall. Spurningarmerki stendur fyrir einn staf, og við höfum tvö spurningarmerki á eftir *n* vegna þess að hvorki kyn né tala skipta máli. Aftast setjum við svo stjörnu sem stendur fyrir ótiltekinn fjölda stafa. Það er vegna þess að ekki skiptir máli hvort orðið hefur viðskeyttan greini, er mannsnafn eða staðarnafn, o.s.frv. Þarna eru líka þrjú mynstur fornafna, þar sem undirflokkur, kyn og tala skiptir ekki máli; og einnig eru þarna (þótt það sjáist ekki á myndinni) þrjú mynstur lýsingarorða, eitt fyrir hvert aukafall.

Leit með þessu mynstri í textum *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) skilar á fimmta hundrað dæmum – hér má sjá nokkur þeirra.

(6) a. ég held að henni leiðist vinnan
b. og af hverju ekki ef mér leyfist að spyrja?
c. það var óneitanlega léttir þegar þessu var lokið
d. allt og sumt var að hana hafði dreymt eitthvað um nóttina

Eins og þessi dæmi benda til sýna langflest þeirra dæma sem leitin skilar í raun aukafallsfrumlög – skilgreining okkar dugir ekki til að greina þar á milli. Með því að láta forritið raða dæmunum eftir sögninni er þó mjög fljótlegt að vinsa þessi dæmi frá. Í sumum þeirra dæma sem þá standa eftir er aukafallsnafnliðurinn ekki andlag heldur hefur stöðu atviksliðar, eins og hér eru sýnd dæmi um:

(7) a. og vill svo til að þann dag varð Tryggvi Gunnarsson sjötugur
b. og um leið og þetta upplaukst fyrir henni varð henni ljóst að allar götur frá fyrstu árum hafði hún verið að lemja niður einhvern óskiljanlegan ótta

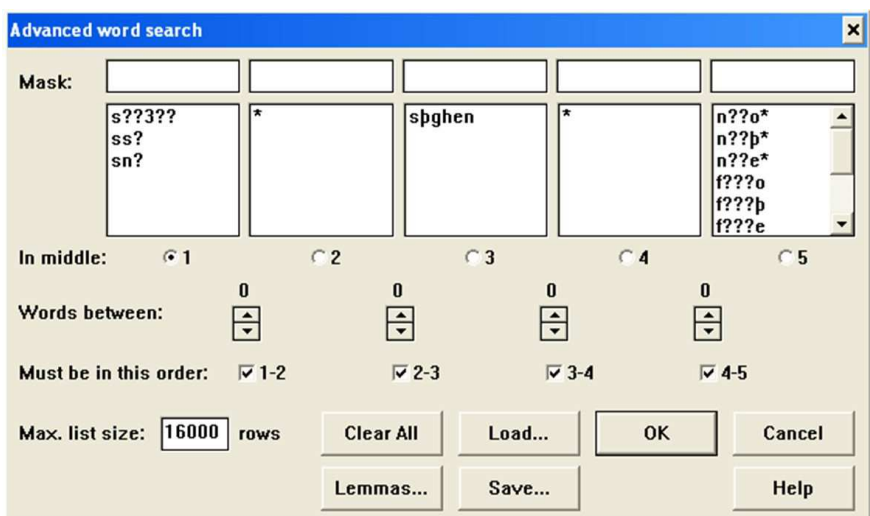
Þegar þessi dæmi hafa líka verið síuð frá koma í ljós fáein dæmi um kjarnafærslu, eins og hér er sýnt.

(8) a. Guð veit að það geri ég líka
b. hún fullyrti að það myndi hún gera
c. ég gerði hins vegar alls ekki ráð fyrir að þessu sæti myndi fylgja seta á Alþingi
d. þeir vissu þó vel að þetta máttu þeir ekki
e. reynt var að leiða henni fyrir sjónir að þetta yrðu allir að gera
f. Vigdís forseti segir að bestu ljóð um vor og sumar hafi karlar ort til kvenna

Þarna er sem sé hægt á fáeinum mínútum að kalla fram öll dæmi um kjarnafærslu andlaga í aukasetningum í 500 þúsund orða texta. Það hefði tekið fleiri daga að finna þessi dæmi í ómörkuðum texta, hvort sem maður hefði leitað að þeim með því hreinlega að lesa allan textann eða reynt að leita að dæmunum í tölvu á einhvern hátt.

3.3.3 Leitað að nýju þolmyndinni

Annað dæmi má taka af hinni svokölluðu „nýju þolmynd“ (Sigríður Sigurjónsdóttir og Joan Maling 2001). Megineinkenni hennar er að andlag germyndarsetningar flyst ekki í frumlagssæti þótt sögnin taki á sig þolmyndarform (þ.e., aðalsögnin standi í lýsingarhætti þátíðar og so. *vera* eða *verða* komi inn), heldur stendur áfram í dæmigerðu andlagssæti. Þar að auki halda þolfallsandlög falli sínu í stað þess að fá nefnifall eins og þau gera þegar þau færast í frumlagssæti í venjulegri þolmynd (þágufalls- og eignarfallsandlög halda aftur á móti alltaf falli sínu). Að þessari setningagerð má leita með þessu mynstri:



Hér táknar * í öðrum leitarreit og *sþghen* í þriðja leitarreit að leitað sé að lýsingarhætti þátíðar af hvaða sögn sem er. Þar á eftir kemur óskilgreint orð, en í fimmta leitarreit er tilgreint að það orð skuli vera nafnorð eða fornafn (eða lýsingarorð, þótt það sjáist ekki á myndinni) í aukafalli. Í fyrsta reit getur svo verið sögn í þriðju persónu (*s??3??*), nafnhætti (*sng*) eða sagnbót (*ssg*). Þannig finnast dæmi eins og *(Það) var barið mig*, *(Það hefur) verið barið mig*, og *(Það mun) verða barið mig*.

Þetta leitarmynstur var keyrt á texta *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) og skilaði 45 dæmum, en fljótlegt er að ganga úr skugga um að ekkert þeirra sýnir nýja þolmynd svo að öruggt sé. Þarna eru t.d. 8 dæmi um sambandið *þegar/er hér var komið sögu*. Eina dæmið sem svipar til nýrrar þolmyndar

er eftirfarandi:

(9) þar með var lokið hvellinum mikla

Lítill vafi leikur þó á því að eðlilegra er að greina þetta dæmi á annan hátt (þ.e. sem frestun þungs nafnliðar eða eitthvað slíkt; sjá Höskuld Þráinsson 2005:587–588).

Sama leitarmynstur var keyrt á ÍSTAL-safnið og skilaði þar fimm dæmum; eitt þeirra gæti hugsanlega verið ný þolmynd:

(10) það var lokað tjaldstæðinu á Þingvöllum

Hér er þó einnig hugsanlegt að um sé að ræða frestun þungs nafnliðar en ekki nýja þolmynd. Til að fá ótvíræð dæmi um nýja þolmynd þyrfti andlagið að vera perónufornafn, því að þeim er ekki hægt að fresta á þennan hátt.

Það kemur ekki á óvart að engin ótvíræð dæmi um nýja þolmynd skuli finnast í þessum tveimur söfnum. Bæði er þessi setningagerð yfirleitt talin frekar nýtilkomin, og þar að auki að mestu bundin við mál barna og unglinga (sjá Sigríði Sigurjónsdóttur og Joan Maling 2001), en allir textar í söfnunum eru frá fullorðnu fólki. En vissulega er gagnlegt að geta fengið staðfestingu á þessu með ítarlegri leit.

3.3.4 Leitað að *það*-lepp með áhrifssögnum

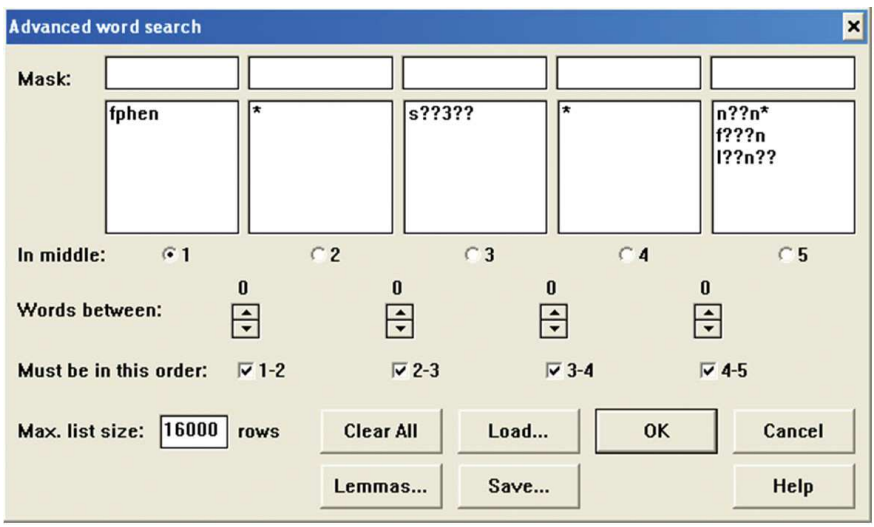
Dæmi má einnig taka af leppnum eða aukafrumlaginu *það* (sjá Eirík Rögnvaldsson 2002). Ég leitaði einu sinni í ÍSTAL-safninu að dæmum um *það* með áhrifssögnum, eins og sést í (11), og sagðist „í fljótu bragði“ ekki hafa fundið nein dæmi af þessu tagi þar (Eiríkur Rögnvaldsson 2002:11nm):

(11) a. Það hefur einhver borðað allan grautinn minn.
b. Það getur enginn svarað þessu.
c. Það stungu einhverjir stúdentar smjörinu í vasann.
d. Það keypti hann eitthvert fífl.

Á þeim tíma átti ég ekki völ á öðru en textaleit; leitaði að dæmum um *það* og fór yfir þau. En þau dæmi eru ákaflega mörg í ÍSTAL (á sjöunda þúsund) þannig að auðvelt er að láta sér sjást yfir einhver dæmi um það sem maður er að svipast um eftir. Vegna þess hversu

seinlegt þetta var lagði ég ekki í að leita í öllu ÍSTAL-safninu í þessari leit.

Mér fannst þess vegna forvitnilegt að gera aðra atrennu að því að leita að þessari setningagerð í ÍSTAL, og notaði til þess eftirfarandi mynstur:



Hér er *fphen* í fyrsta reit; það á eingöngu við *það*. Annar og þriðji reitur skilgreina svo ótilgreina sögn í 3. persónu, og fjórði og fimmti reitur skilgreina fallorð (nafnorð, fornafn eða lýsingarorð) í nefnifalli. Þessi leit skilar 720 dæmum úr ÍSTAL, en í þeim flestum er nefnifallsliðurinn sagnfylling en ekki frumlag. Hann stendur þá næst á eftir einhverri mynd so. *vera* eða *verða*, og þeim dæmum er auðvelt að henda út með því að láta forritið raða dæmunum eftir sögninni. Þá standa eftir milli 30 og 40 dæmi, sem sum hver sýna ótvírætt þá setningagerð sem leitað var að:

(12) a. það átti enginn skap saman
b. það þekkja allir Rósu
c. það vita allir hver Rósa er
d. það heldur enginn að þú sért hommi

Í stað þess að lesa allt ÍSTAL-safnið, eða fara gegnum á sjöunda þúsund dæma um *það*, dugir því að skoða þessi 30-40 dæmi. Í stað margra tíma þreytandi yfirlegu þar sem hætta á mistökum er veruleg

kemur 10-20 mínútna vinna þar sem tækifæri gefst til að skoða hvert dæmi vandlega og meta hvort það falli undir þá setningagerð sem leitað er að.

3.4 Setningafræðileg þáttun

Þessar tilraunir sýna ótvírætt að hægt er að hafa verulegt gagn af hinni beygingarlegu mörkun í setningafræðilegri dæmaleit. Vissulega er ekki hægt að leita að öllum setningagerðum á þennan hátt. Það á m.a. við um setningagerðir þar sem vensl milli orða ná yfir ótiltekinn orðafjölda, langdræg vensl. Það væri t.d. erfitt að nota þessa aðferð til að skoða vísun afturbeygðra fornafna svo að dæmi sé tekið. *WinCord* forritið býður að vísu upp á að tilgreint sé hversu mörg orð megi koma á milli orðanna í leitarreitunum – jafnvel er hægt að leyfa ótiltekinn orðafjölda. En gallinn við það er að þá kemur óhjákvæmilega með mikill fjöldi dæma sem ekki koma málinu við, og mjög tímafrekt getur verið að hreinsa frá.

Nú er að opnast annar möguleiki á setningafræðilegri leit í íslenskum textum. Verið er að vinna að gerð hlutaþáttara (e. shallow parser) fyrir íslensku. Hlutaþáttun er setningafræðileg greining þar sem ekki er stefnt að því að sýna fullkomna formgerð setninga eða öll vensl liða. Þess í stað er lögð áhersla á að greina helstu setningarliði – flokka saman orð sem eiga saman. Slík greining getur nýst vel í ýmsum tilvikum, og hentar stundum betur en full greining (e. deep parsing). Einnig eru helstu setningafræðileg hlutverk greind. Samið hefur verið sérstakt þáttunarskema til að skilgreina hvaða liðir og hlutverk eru greind, og við hvað skuli miðað í greiningunni (Hrafn Loftsson og Eiríkur Rögnvaldsson 2006). Hér er sýnt dæmi um setningu sem greind hefur verið eftir þessu skema.

(13) {***SUBJ**> [*NP* augnaráðið `nheng` *NP*] ***SUBJ**>}
[*VP* negldist `sfm3eþ` *VP*]
[*PP* við `ao` [*NP* [*AP* gráa `lkeovf` AP] jakkann `nkeog` *NP*] *PP*]
[*SCP* sem `ct` *SCP*]
{***SUBJ**> [*NP* hann `fpken` *NP*] ***SUBJ**>}
[*VPb* var `sfg3eþ` *VPb*]
[*VPi* að `cn` klæða `sng` *VPi*]
{***OBJ**< [*NP* sig `fpkeo` *NP*] ***OBJ**<}
[*PP* úr `aþ` *PP*]
[*CP* og `c` *CP*]
[*VPi* hengja `sng` *VPi*]
[*PP* [*MWE_PP* inn `aa` í `ao` *MWE_PP*] [*NP* skáp `nkeo` *NP*] *PP*]

Hér eru beygingarleg mörk með lágstöfum; mörk setningarliða með hástöfum og skáletruð, auk þess sem hornklofar umlykja liðina; og mörk setningafræðilegra hlutverka hefjast á stjörnu og eru með hástöfum og feitletruð, og slaufusvigar afmarka hlutverkin. Mörk liða og hlutverka ættu flest að vera auðskilin. Þó er rétt að nefna að *MWE* stendur fyrir 'multiword expression' og er notað til að marka orðarunur sem í raun eru ígildi eins orðs (einkum fleiryrtar samtengingar og forsetningar). Oddur (> og <) er notaður á frumlög og andlög (og sagnfyllingar) til að vísa á sögnina sem þessir liðir tengjast.

Eins og sjá má er hér ekki gerð tilraun til að tengja liði saman og sýna þannig stigveldisformgerð setningarinnar, nema að litlu leyti (nafnliðir eru sýndir sem hluti forsetningarliða). Áhersla er fremur lögð á að ná mikilli nákvæmni í vélrænu greiningunni, miðað við greiningarskemað sem þáttarinn vinnur með. Fyrstu tilraunir benda til að þáttarinn skili hlutverki sínu mjög vel og villur í greiningunni séu fáar.

Þar með eru komnar nýjar forsendur til setningafræðilegrar dæmaleitar í textum. Með þessari greiningu verður t.d. hægt að leita að röðinni *aukatenging – andlag – sagnliður* og finna þannig dæmi um kjarnafærslu í aukasetningum, í stað þess að byggja leitina á beygingarlegum mörkum eins og gert var hér að framan. Það á svo eftir að koma í ljós hvort leit eftir setningafræðilegum mörkum skilar betri árangri en hin. En hér eru e.t.v. að opnast möguleikar á að koma upp vísi að íslenskum trjábanka.

4 Lokaorð

Í fyrri hluta þessarar greinar var sagt lauslega frá mismunandi viðhorfum til gildis textadæma í setningafræði undanfarna áratugi. Tilkoma generatífrar málfræði fyrir hálfri öld olli því að dæmaleit í textum naut lítillar virðingar um langt skeið, og málfræðingar skiptust í fylkingar sem höfðu mjög andstæðar skoðanir á þessu sviði, þótt sá ágreiningur hafi að verulegu leyti verið sýndarágreiningur og stafað af því að menn voru að bera saman epli og appelsínur. En með tilkomu viðamikilla rafrænna textasafna og málheilda, og ekki síst mikils magns veftexta, hafa skilin milli fylkinganna dofnað og nú þykir ekki lengur neitt að því að safna dæmum úr textum. En ýmislegt er að varast við notkun dæmanna og gæta verður varúðar í túlkun þeirra, eins og bent er á í greininni.

Meginviðfangsefni greinarinnar var að skoða hvernig hægt er að standa að verki við leit að dæmum um tilteknar setningagerðir í rafrænum íslenskum textasöfnum. Bent var á að hægt er að nýta hráa texta, án nokkurrar sérstakrar mörkunar, að vissu marki, en þó því aðeins að leitað sé að setningagerðum sem tengjast ákveðnum orðum. Með tilkomu beygingarlega markaðra málheilda og mörkunarforrita hafa aðstæður til setningafræðilegrar leitar hins vegar gerbreyst. Vegna eðlis íslenska beygingakerfisins má lesa miklar setningafræðilegar upplýsingar út úr hinum beygingarlegu mörkum, og þær upplýsingar má síðan nýta í leit að ákveðnum setningagerðum. Sýnd voru þrjú dæmi um hvernig hægt er á einfaldan og fljótvirkan hátt að leita að dæmum um þrjár setningagerðir; kjarnafærslu í aukasetningum, nýja þolmynd, og *það*-lepp með áhrifssögnum. Í öllum tilvikum skilaði leitin niðurstöðum sem tekið hefði fleiri daga að fá með þeim aðferðum sem áður var völ á, en nú tók leitin aðeins fáeinar mínútur.

Í lok greinarinnar er svo sagt frá verkefni sem enn er ólokið og felst í gerð hlutaþáttara fyrir íslensku. Ef það verkefni skilar tilætluðum árangri er hægt að fara að gera raunhæfar áætlanir um smíði viðamikils íslensks trjábanka sem myndi verða öllum sem fást við rannsóknir á íslenskri setningafræði að ómetanlegu gagni.

Heimildir

Ásta Svavarsdóttir. 2006. Tilbrigði í setningagerð. *Orð og tunga* 8:156–157.

Ásta Svavarsdóttir. 2007. Talmál og málheildir – talmál og orðabækur. *Orð og tunga* 9 (þetta hefti).

Bergljót S. Kristjánsdóttir, Eiríkur Rögnvaldsson, Guðrún Ingólfsdóttir og Örnólfur Thorsson (ritstj.). 1996. *Orðstöðulykill Íslendinga sagna.* [Geisladiskur.] Mál og menning, Reykjavík.

Chomsky, Noam. 1957. *Syntactic Structures.* Mouton, Haag.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Massachusetts.

Eiríkur Rögnvaldsson. 1994-95. Breytileg orðaröð í sagnlið. *Íslenskt mál* 16–17:27–66.

Eiríkur Rögnvaldsson. 1996. Frumlag og fall að fornu. *Íslenskt mál* 18: 37–69.

Eiríkur Rögnvaldsson. 1998. Heimildatúlkun í sögulegri setningafræði. Baldur Sigurðsson, Sigurður Konráðsson og Örnólfur Thorsson (ritstj.): *Greinar af sama meiði*, bls. 317–334. Rannsóknarstofnun Kennaraháskóla Íslands, Reykjavík.

Eiríkur Rögnvaldsson. 2000. Setningarstaða boðháttarsagna í fornu máli. *Íslenskt mál* 22:63–90.

Eiríkur Rögnvaldsson. 2002. *ÞAÐ* í fornu máli – og síðar. *Íslenskt mál* 24:7–30.

Faarlund, Jan Terje. 1990. *Syntactic Change. Toward a Theory of Historical Syntax.* Mouton, Berlín.

Falk, Cecilia. 1995. Lexikalt kasus i svenska. *Arkiv för nordisk filologi* 110:199–226.

Halldór Ármann Sigurðsson. 1989. *Verbal Syntax and Case in Icelandic.* In a Comparative GB Framework. Doktorsritgerð, Lund Universitet, Lund.

Hrafn Loftsson og Eiríkur Rögnvaldsson. 2006. A Shallow Syntactic Annotation Scheme for Icelandic Text. *Technical Report RUTR-SSE06004*, Department of Computer Science, Reykjavik University, Reykjavík.

Höskuldur Þráinsson. 1979. *Complementation in Icelandic.* Garland, New York.

Höskuldur Þráinsson. 2005. *Setningar.* Handbók um setningafræði. (Íslensk tunga III.) Almenna bókafélagið, Reykjavík.

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

Kromann, Matthias Trautner. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. Joakim Nivre og Erhard Hinrichs (ritstj.): *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, bls. 217–220. Växjö University Press, Växjö.

Lightfoot, David S. 1979. *Principles of Diachronic Syntax.* Cambridge University Press, Cambridge.

Maren Albertsdóttir og Stefán Einar Stefánsson. 2004. Beygingar- og málfræðigreinikerfi. *Samspil tungu og tækni*, bls. 16–19. Menntamálaráðuneytið, Reykjavík.

McEnery, Tony, og Andrew Wilson. 1996. *Corpus Linguistics.* Edinburgh University Press, Edinburgh.

Mørck, Endre. 1992. Subjektets kasus i norrønt og mellomnorsk. *Arkiv för nordisk filologi* 107:53–99.

Nivre, Joakim. 2002. What kinds of trees grow in Swedish soil? A comparison of four annotation schemes for Swedish. Erhard Hinrichs og Kiril Simov (ritstj.): *Proceed-*

ings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002), 20–21 September 2002. Sozopol, Bulgaria.
Osenova, Petya, og Kiril Simov. 2003. The Bulgarian HPSG Treebank: Specialization of the Annotation Scheme. Joakim Nivre og Erhard Hinrichs (ritstj.): *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, bls. 129–140. Växjö University Press, Växjö.
Sigríður Sigurjónsdóttir og Joan Maling. 2001. Það var hrint mér á leiðinni í skólann: Þolmynd eða ekki þolmynd? *Íslenskt mál* 23:123–180.
Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga* 9 (þetta hefti).
Þórunn Blöndal. 2005. *Lifandi mál.* Inngangur að orðræðu- og samtalsgreiningu. Rannsóknarstofnun Kennaraháskóla Íslands, Reykjavík.

Lykilorð:

málheildir, dæmasetningar, málfræðileg mörkun

Keywords:

text corpora, example sentences, PoS tagging

Abstract

This paper discusses the use of text corpora in syntactic research, and how to search for example sentences in corpora. During the past few decades, widely divergent views have been expressed as to the value of corpora in syntactic argumentation. It is argued in the paper that this disagreement stems from different views as to the subject of linguistic research. The paper also discusses various problems that arise in the interpretation of the information extracted from corpora – especially in drawing conclusions from the silence of the texts on certain constructions. The main section of the paper discusses the possibilities of searching for certain syntactic constructions in different types of Icelandic corpora; raw untagged text, PoS tagged text, and text where the major syntactic constituents and syntactic functions have been identified. Data-driven PoS taggers have now been trained on Icelandic texts, and it is shown that due to the inflectional character of Icelandic and the richness of the tagset, the resulting PoS tagging is very effective in the search for various syntactic constructions.

Eiríkur Rögnvaldsson
Háskóla Íslands
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is

Sigrún Helgadóttir

Mörkun íslensks texta

1 Inngangur

Í ýmsum tungutækniverkefnum[1] 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, 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, Friðrik Magnússon og Stefán Briem 1991).

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998–1999 (Rögnvaldur Ólafsson, Þorgeir Sigurðsson og Eiríkur Rögnvaldsson 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“. Í anda tillögunnar var gerð málfræðilegs markara fyrir íslensku eitt af þeim verkefnum sem var styrkt af tungutækniverkefni menntamálaráðuneytisins[2] í apríl 2002. Markmið verkefnisins var að finna aðferðir til þess að

[1] Orðið *tungutækni* er hér notað um það sem á ensku nefnist venjulega **language engineering**. Einnig má nota orðið *máltækni*.

[2] Verktakar við verkið voru Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir) og Orðabók Háskólans. Verkefnisstjóri var Eiríkur Rögnvaldsson en Sigrún Helgadóttir mótaði

Orð og tunga 9 (2007), 75–107.

greina íslenskan texta vélrænt í orðflokka og eftir beygingu. Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni.

Verkefnið þróaðist á þann veg að prófaðar voru fjórar aðferðir við mörkun íslensks texta. Í greininni verður gerð grein fyrir málfræðilegri mörkun og nokkrum aðferðum við vélræna greiningu. Greint verður frá tilraun til þess að nota fjórar aðferðir við vélræna mörkun íslensks texta. Við tilraunirnar var notað textasafn sem var búið til vegna *Íslenskrar orðtíðnibókar*. Einnig verður greint frá tilraunum til þess að bæta mörkun, m.a. tilraunum til þess að sameina niðurstöður þriggja markara eftir tilteknum reglum til þess að ná sem bestum árangri við mörkun. Að lokum er greint frá tilraunum til þess að marka texta sem eru ekki hluti af textasafni Orðtíðnibókarinnar. Lokaskýrslu var skilað til menntamálaráðuneytisins í febrúar 2004.

2 Málfræðileg mörkun texta

Með mörkun (e. *tagging*) er átt við það að merkja orð í samfelldum texta á kerfisbundinn hátt, t.d. með málfræðilegum upplýsingum, nefnimynd orðsins og upplýsingum um setningafræðilegt hlutverk. Í þessari grein er orðið *mark* notað um málfræðilegt mark[3]. Málfræðilegt mark er greiningarstrengur sem er tengdur orði í texta og segir til um orðflokk orðsins og önnur málfræðileg atriði, t.d. kyn, tölu og fall fallorða og persónu, tölu og tíð 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.

Elsta aðferð við málfræðilega mörkun er handvirk greining texta eftir orðflokkum og beygingu. Sú aðferð er þó mjög tímafrek og þess vegna hefur lengi verið fengist við að þróa vélrænar aðferðir við mál-

vinnulag við prófun markaranna og vann meginhluta vinnunnar ásamt félögum í Málgreiningarhópnum.

[3] Í ensku eru notuð orðin **POS tag, part-of-speech tag** og **morphological tag** um það sem hér er kallað málfræðilegt mark. Þó að **POS** eða **part-of-speech** sé venjulega notað um orðflokk eru þessi orð oft einnig látin ná yfir beygingarlegar myndir.

fræðilega mörkun. Þetta svið hefur því fengið mikla umfjöllun á undanförnum áratugum hjá þeim sem vinna við máltækni.

Vélrænar aðferðir við mörkun eru venjulega flokkaðar í tvo flokka, regluaðferðir (e. *rule based methods*) og gagnaaðferðir (e. *data-driven methods*). Fyrstu vélrænu aðferðirnar sem var beitt voru regluaðferðir. Orðasafn var notað til þess 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. Forrit sem notuðu reglurnar voru háð því tungumáli sem reglurnar voru gerðar fyrir.

Gagnaaðferðir byggjast allar á því að nota textasafn sem hefur verið markað og mörkunin yfirfarin handvirkt þannig að hún sé eins rétt og kostur er. Forrit er síðan látið læra af gögnunum á tiltekinn hátt. Í þeirri vinnu sem hér er greint frá voru gerðar tilraunir með þrjár mismunandi gagnaaðferðir: tölfræðilegar aðferðir, aðferð sem mætti kalla leiðréttingaaðferð (e. *transformation-based learning*) og minnisaðferð (e. *memory-based method*). Forrit eða kerfi sem nota fyrir fram greint textasafn til þess að læra af mætti kalla námfúsa markara. Í greininni er sagt frá tilraun til þess að láta fimm mismunandi námfúsa markara læra að marka íslenskan texta. Tveir markaranna nota tölfræðilegar aðferðir, tveir nota leiðréttingaaðferð og einn notar minnisaðferð. Í 4. kafla er gerð grein fyrir þessum aðferðum og forritum.

3 Efniviður

Í þeirri vinnu sem hér er lýst var notað textasafn sem var gert fyrir vinnslu *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991) sem Orðabók Háskólans gaf út 1991. Vinna við undirbúning textasafnsins hófst 1985 og er safninu lýst nákvæmlega í formála Orðtíðnibókarinnar. Í textasafninu eru brot úr 100 textum sem voru gefnir út á tímabilinu 1980–1989, hvert með um 5.000 lesmálsorðum. Textarnir voru valdir úr 5 textaflokkum: íslenskum skáldverkum (20 textar), þýddum skáldverkum (20 textar), ævisögum og minningum (20 textar), fræðslutextum (10 á sviði hugvísinda, 10 á sviði raunvísinda) og barna og unglingabókum (10 frumsamdir textar, 10 þýddir textar).

Í formála Orðtíðnibókarinnar er *lesmálsorð* skilgreint sem samfelld röð af bókstöfum og/eða tölustöfum og táknum sem aðgreind eru með stafbili eða greinarmerkjum. Notuð var sú regla að reynt var að fella eins langan stafastreng undir töluorð og kostur var. Plúsar, mínusar og prósentumerki fylgdu þannig lesmálsorðunum. Blendingar tölustafa og annarra rittákna, t.d. efnafræðiformúlur og stærðfræðiformúlur teljast eitt lesmálsorð. Vert er að benda á að skammstafanir eru í flestum tilvikum greindar eins og lesið er úr þeim.

Í textasafninu eru 590.297 lesmálsorð sem birtast í 59.358 mismunandi orðmyndum að meðtöldum greinarmerkjum. Lesmálsorðunum fylgja 639 mismunandi greiningarstrengir að meðtöldum greinarmerkjum. Þegar fengist er við vélræna málfræðilega greiningu er erfiðast að eiga við orðmyndir sem geta haft fleiri en eina greiningu. Í textasafni Orðtíðnibókarinnar hafa 15,9% orðmynda fleiri en eina hugsanlega greiningu. Margræðasta orðmyndin er *minni* sem hefur 24 greiningarstrengi í textasafninu, en fleiri eru mögulegir (ég *minni* þig á það; ég geri þetta eftir *minni*; Nonni er *minni* en Siggi; o.s.frv.)

orð	**nefnimynd**	**mark**	**skýring**
ég	ég	fp1en	f: fn; p: pfn; 1: 1. pers.; e: et.; n: nefnifall
stökk	stökkva	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
á	á	aa	a: ao.; a: stýrir ekki falli
eftir	eftir	aþ	a: ao.; þ: stýrir þágufalli
stræto	stræto	nkeþ	n: no.; k: kk.; e: et.; þ: þgf.
og	og	c	c: samtenging
veifaði	veifa	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
,	,	,	komma
vagnstjórinn	vagnstjóri	nkeng	n: no.; k: kk.; e: et.; n: nf.; g: með greini
sá	sjá	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
mig	ég	fp1eo	f: fn.; p: pfn.; 1: 1. pers.; e: et.; n: þolfall
og	og	c	c: samtenging
stoppaði	stoppa	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
.	.	.	punktur

Mynd 1. Greining orða í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Hverju lesmálsorði var síðan komið fyrir í sérstakri línu. Í þeirri línu var einnig komið fyrir greiningarstreng orðsins eða marki og nefnimynd (flettimynd) þess. Í mynd 1 er sýnd ein setning úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson og hvernig hún er greind. Til glöggvunar er sýnd skýring á greiningarstrengjum.

Í formála Orðtíðnibókarinnar er gerð grein fyrir vélrænni greiningu sem notuð var við gerð bókarinnar (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991). Vélræna greiningin byggist á greiningu 54.000 lesmálsorða sem höfðu verið greind handvirkt og notuð við orðtíðnikönnun (Friðrik Magnússon 1988). Stefán Briem (1990) gerir grein fyrir aðferðum sem var beitt við vélrænu greininguna. Höfundar Orðtíðnibókarinnar telja að um 80% lesmálsorða hafi fengið rétta greiningu að öllu leyti með vélrænu greiningunni. Nokkrum árum seinna var forritið endurbætt á grundvelli greiningar alls textans. Fékkst þá tæplega 90% nákvæmni (Stefán Briem, munnlegar upplýsingar). Athyglisvert er að bera þá niðurstöðu saman við niðurstöðu tilraunarinnar sem hér verður greint frá.

Í greiningu lesmálsorða sem notuð var í Orðtíðnibókinni er greint á milli átta orðflokka: nafnorða, lýsingarorða, fornafna, lauss greinis, töluorða, sagna, atviksorða og samtenginga. Orð sem ekki flokkast í þessa orðflokka voru annað hvort talin erlend orð eða ógreind orð. Helstu frávik frá venjulegri orðflokkagreiningu voru þau að forsetningar voru taldar með atviksorðum. Þess vegna koma fyrir atviksorð sem stýra falli. Upphrópanir voru einnig taldar með atviksorðum. Nafnháttarmerki var talið með samtengingum. Í viðauka A er yfirlit yfir greiningarstrengi sem voru notaðir.

4 Aðferðir og markarar

Í þeirri könnun sem hér er greint frá voru eingöngu prófaðar gagnaaðferðir. Þær byggjast á því að forrit býr til líkan út frá fyrir fram greindu textasafni. Þetta safn kallast þjálfunarsafn. Aðferðin er síðan prófuð á sérstöku prófunarsafni. Til þess að prófa tiltekna mörkunaraðferð þarf að hafa aðgang að nokkuð stóru textasafni sem hefur verið greint í lesmálsorð og hverju lesmálsorði gefinn greiningarstrengur í samræmi við þá greiningu sem óskað er að fá fram. Textasafninu er skipt í tvo hluta og er annar hlutinn kallaður þjálfunarsafn og hinn hlutinn prófunarsafn. Þjálfunarsafnið er oft um 90% af textasafninu

sem er til umráða og prófunarsafnið um 10%. Búið er til líkan með aðstoð þjálfunarsafnsins og það síðan prófað á prófunarsafninu. Þar sem prófunarsafnið er líka fullgreint má reikna út hversu nákvæm aðferðin er.

Prófaðar voru fjórar gagnaaðferðir og fimm forrit eða markarar sem unnt er að þjálfa á íslenskum texta og eru fáanleg án greiðslu. Prófaðir voru tveir tölfræðimarkarar, **TnT** sem byggist á Markovslíkani og **MXPOST** sem byggist á svo kölluðu hámarksóreiðulíkani (e. *Maximum Entropy Model*). Prófaðir voru tveir markarar sem byggjast á leiðréttingaaðferð, μ**-TBL** og **fnTBL.** Einnig var prófaður einn markari, **MBT**, sem byggist á minnistækni. Alla þessa markara má kalla á íslensku gagnamarkara eða námfúsa markara. Markarinn μ-TBL hafði áður verið prófaður á litlum úrdrætti úr textasafni Orðtíðnibókarinnar og fékkst þá viðunandi niðurstaða (Sigrún Helgadóttir 2002). En markarinn virtist ekki ráða við allt textasafn Orðtíðnibókarinnar. Honum eru því ekki gerð frekari skil. Hér á eftir verður lauslega lýst þeim fjórum aðferðum við mörkun sem voru notaðar í tilrauninni og þeim mörkurum sem voru valdir til prófunar á íslenskum texta.

4.1 Falin Markovslíkön

Í þessum flokki var valinn markarinn TRIGRAMS'N'TAGS (TnT) sem Thorsten Brants samdi (Brants 2000a og 2000b). Aðferðinni verður lýst með því að skýra í stórum dráttum hvernig TnT-kerfið starfar.

TnT notar annars stigs Markovslíkön fyrir mörkun. Ástönd (e. *states*) standa fyrir mörk og færslulíkur eru háðar markapörum. Forritið metur færslulíkur og frálagslíkur út frá mörkuðu textasafni. Kerfið notar sennileikalíkur (e. *maximum likelihood probabilities*) sem eru reiknaðar út frá hlutfallslegri tíðni. TnT notar Viterbi-algrím með geislaleit (e. *beam search*) við mörkun til þess að flýta fyrir vinnslu.

TnT vinnur úr óþekktum orðum með því að greina endingar (viðskeyti) eins og lagt er til í Samuelsson (1993) þar sem líkur á mörkum eru stilltar í samræmi við endingar orða. Lengsta ending sem TnT notar er 10 stafa löng (10 er sjálfgildi í forritinu). Líkindadreifing fyrir tiltekna endingu er búin til með því að skoða öll orð í þjálfunarsafni sem hafa sameiginlega endingu af tiltekinni hámarkslengd. Forritið vinnur aðeins úr endingum orða sem hafa tiltekna lágmarkstíðni og var tíðnin 10 valin út frá reynslu. Forritið heldur einnig tvo lista yf-

ir endingar, einn fyrir orð sem hefjast á lágstaf og einn fyrir orð sem hefjast á hástaf.

TnT er beitt á nýtt mál eða nýtt svið í tveimur þrepum:

1. Líkan er búið til
2. Texti er markaður

Líkanið er búið til út frá þjálfunarsafninu. Tvær skrár verða til í því skrefi: skrá með tíðni orða og marka sem þau geta fengið og skrá með tíðni tveggja eða þriggja marka sem standa saman. Þessar skrár eru síðan notaðar þegar forritið markar nýjan texta. Forritið gefur einnig kost á að nota viðbótarorðasafn. Finnist orð ekki í orðasafninu, sem var búið til þegar líkanið var gert, er leitað að því í viðbótarorðasafninu.

4.2 Hámarksóreiðuaðferð

Í þessum flokki var valinn markarinn **MXPOST** (Ratnaparkhi 1996). Í Ratnaparkhi (1997) er inngangur að því hvernig hámarksóreiðulíkön (e. *Maximum Entropy Models*) eru notuð við málgreiningu. Ratnaparkhi segir þar að mörg málgreiningarverkefni megi endurskilgreina sem tölfræðileg flokkunarverkefni. Verkefnið felst í því að meta líkur á að flokkur *a* komi fyrir í „samhenginu“ *b*, eða *p(a,b).* Í málgreiningarverkefnum eru orð venjulega hluti af „samhenginu“. Í sumum verkefnum er „samhengið“ aðeins eitt orð en í öðrum getur *b* verið nokkur orð og greiningarstrengir þeirra. Í stórum textasöfnum fæst nokkur vitneskja um hvenær *a* og *b* koma fyrir saman. En ekkert textasafn hefur nægilegar upplýsingar til þess gefa upplýsingar um *p(a,b)* fyrir öll hugsanleg pör *(a,b)* þar sem orðin í *b* eru sjaldgæf. Vandamálið snýst um að meta á öruggan hátt líkindalíkanið *p(a,b)* með því að nota ófullkomnar upplýsingar um *a*-in og *b*-in.

Þjálfunarsafninu er lýst sem miklum fjölda af sérkennaþáttum (e. *features*). Þessir sérkennaþættir eru tvígild föll af „sögum“ (e. *histories*) (samhengi orða og greiningarstrengja) og greiningarstrengjum. Í útgáfu Ratnaparkhis eru sérkennaþættir orðið sem verið er að fjalla um, næstu tvö orð á undan, næstu tvö orð á eftir og greiningarstrengur (mark) næstu tveggja orða á undan. Sérkennaþættir sjaldgæfra og óþekktra orða (koma ekki fyrir í þjálfunarsafni) hafa einnig fyrstu og síðustu fjóra stafi orðs og upplýsingar um hvort orðið hafi hástaf, bandstrik eða tölustaf. Sérkennaþættir óþekktra orða eru búnir til úr

sérkennaþáttum sjaldgæfra orða. Litið er á sérkennaþætti, sem koma fyrir sjaldnar en 10 sinnum í þjálfunarsafni, sem óáreiðanlega. Markarinn notar geislaleit til þess að finna líklegustu runu marka og sú röð sem hefur hæst líkindi er valin. MXPOST-forritið gefur ekki kost á því að nota viðbótarorðasafn. Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

4.3 Leiðréttingaaðferð

Brill (1994 og 1995) hefur lýst leiðréttingaaðferðinni og hvernig má beita henni við mörkun texta. Með þessari aðferð er málfræðileg þekking skráð í nokkrum einföldum reglum. Fyrst er hverju orði í þjálfunarsafninu gefinn sá greiningarstrengur sem er líklegastur miðað við þjálfunarsafnið sjálft. Þessi mörk eru síðan borin saman við rétt mörk. Forritið lærir leiðréttingareglur sem er beitt til þess að komast nær hinni réttu greiningu. Forritið lærir reglurnar út frá sniðmátum sem lýsa aðgerð (breyta greiningarstreng A í greiningarstreng B) á grundvelli tiltekins umhverfis (orð og mörk í samhengi), þ.e. hvaða orð og mörk eru næst á undan og eftir því marki sem verið er að skoða. Upphaflega gerði Brill ráð fyrir því að aðeins væru skoðuð mörk í næsta nágrenni við markið sem var skoðað. Síðar bætti hann við sniðmátum þar sem gert var ráð fyrir að orð væru skoðuð líka. Forritið sem lærir leiðréttingareglurnar beitir öllum leiðréttingum, telur hversu margar villur hver leiðrétting lagar og velur þá leiðréttingu sem lagar flestar villur. Ákveðið er fyrir fram hver er minnsti fjöldi leiðréttinga sem regla þarf að hafa í för með sér til þess að vera valin. Þegar engar leiðréttingar finnast sem fækka villum um þann fjölda hættir forritið að læra reglur. Á þennan hátt verður til raðað mengi af leiðréttingareglum, hver regla endurspeglar tiltekið sniðmát.

Í fyrstu tilraunum sínum gerði Brill (1994) ráð fyrir því að engin óþekkt orð væru í þeim texta sem átti að marka. Síðar þróaði Brill aðferð til þess að greina óþekkt orð. Aðferðin byggist líka á því að láta forrit læra leiðréttingareglur. Óþekktum orðum eru gefin mörk. Brill gefur óþekktum orðum sem hefjast á lágstaf mark sem venjuleg nafnorð og óþekktum orðum sem hefjast á upphafsstaf mark sem sérnöfn. Síðan eru skilgreind sniðmát. Sniðmát Brills fela í sér að skoðaðir eru fyrstu og síðustu fjórir stafir í orði. Athugað er hvort orðið hafi forskeyti eða viðskeyti sem er eins til fjögurra stafa langt, hvort unnt sé

að taka í burtu eða bæta við eins til fjögurra stafa forskeyti eða viðskeyti og fá út nýtt orð, hvort orðið hafi tiltekinn staf eða hvort tiltekið orð sé til vinstri eða hægri við orðið. Á grundvelli þessara sniðmáta lærir forritið reglur til þess að beita. Í Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir (2002) er góð lýsing á aðferðinni.

Nokkur forrit bjóðast sem nota aðferðir Brills. Valið var að nota forritið **fnTBL** (*Fast Transformation-Based Learning Toolkit*) eftir Florian og Ngai (2002). Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

4.4 Minnistækni

Námfús markari sem byggist á minnisaðferð lærir af mengi dæma sem eru geymd í gagnasafni þar sem hvert dæmi hefur verið flokkað á tiltekinn hátt og markað í samræmi við það. Dæmin eru tekin úr dæmasafni sem hefur verið handmarkað. Þegar flokka á ný dæmi leitar kerfið í gagnasafninu að dæmi eða mengi dæma sem líkjast sem mest nýja dæminu. Þetta má orða þannig að leitað sé að „næsta nágranna“ nýja dæmisins. Aðferðin er dregin af tækni sem kennd er við „*k* næstu nágranna“ og aðeins *k* næstu nágrannar eru skoðaðir. Oft er *k* látið vera 1 en í tilraunum með minnisaðferð er það eitt af markmiðunum að finna besta gildi fyrir *k*.

Til þess að prófa þessa aðferð var valinn **MBT**-markarinn (Daelemans o.fl. 2003) sem notar Tilburg Memory-Based Learner (*TiMBL*, Daelemans o.fl. 2004). MBT-markarinn býr til aðskilin gagnasöfn fyrir þekkt orð og óþekkt orð. *TiMBL*-kerfið notar svokallað IGTREE-algrím fyrir þekkt orð og IB1-algrím fyrir óþekkt orð. Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

5 Mörkun íslensks texta

Í þessum kafla verður gerð grein fyrir tilraunum við að marka texta Orðtíðnibókarinnar. Skipulagi skráa við tilraunina verður einnig lýst og jafnframt hvernig niðurstöður verða metnar.

5.1 Skrár

Til þess að prófa mismunandi aðferðir við mörkun er oft notuð aðferð sem byggist á því að hafa til umráða tíu pör af þjálfunar- og prófunarsöfnum. Í hverri tölvuskrá Orðtíðnibókarinnar er textabútur úr einni heimild. Pörin voru búin til þannig að hverri skrá var skipt upp í tíu nokkurn veginn jafna hluta. Hver þessara tíu hluta myndar eitt prófunarsafn og samstætt þjálfunarsafn hefur að geyma hina hlutana níu í hvert sinn. Stærri skráin er notuð sem þjálfunarsafn og sú minni sem prófunarsafn. Prófunarsöfnin skarast því ekki en þjálfunarsöfnin hafa um 80% sameiginlega texta. Allir markarar voru prófaðir á öllum 10 pörum og fundin meðalnákvæmni (þessi aðferð er kölluð á ensku *ten-fold cross-validation*).

5.2 Mælikvarðar fyrir nákvæmni

Til þess að finna hversu nákvæmlega markari úthlutar mörkum eru mörk hans borin saman við „rétt mörk“ sem hafa verið yfirfarin handvirkt. Þessi réttu mörk eru oft kölluð á ensku „gold standard“. Mjög erfitt er að ná 100% réttri mörkun þar sem ýmis álitamál koma upp. Tveir einstaklingar mörkuðu textasafn Orðtíðnibókarinnar og hafa þeir efalaust borið sig saman. Nákvæmnin hefur ekki verið rannsökuð sérstaklega.

Til þess að geta metið árangur tiltekins markara þarf að hafa viðmiðun um lægstu nákvæmni, *grunnmörkun* (e. *baseline tagging*), sem unnt er að ná án þess að nota markarann. Grunnmörkun er gerð með tilteknu orðasafni sem hefur upplýsingar um orðmyndir og mörk þeirra og með því að nota tiltekna aðferð við mörkun óþekktra orða. Hér er notuð ein af fjórum skilgreiningum á grunnmörkun sem kemur fram í Megyesi (2002:55). Búið er til orðasafn úr hverju þjálfunarsafni og þeim orðmyndum í viðkomandi prófunarsafni sem koma líka fyrir í þjálfunarsafninu gefið algengasta mark þeirrar orðmyndar. Óþekkt orð rituð með litlum staf fá algengustu greiningu nafnorða (*nken*) og óþekkt orð rituð með upphafsstaf fá algengustu greiningu sérnafna (*nken-m*). Þessi mörk eru síðan borin saman við rétt mörk. Meðalnákvæmni slíkrar mörkunar fyrir öll prófunarsöfnin reyndist **76,63%.** Markari sem nær ekki þessari nákvæmni bætir því engu við það sem fæst með grunnmörkun eingöngu.

Frammistaða hvers markara er metin með því að reikna út hittni (e. *accuracy*) miðað við rétta greiningu (handmörkun) og reiknuð sem

> *hittni=(fjöldi rétt greindra lesmálsorða)/(heildarfjöldi lesmálsorða í safni)*

Tökum sem dæmi að í prófunarsafni séu 59.169 lesmálsorð. Tiltekinn markari markar 53.101 lesmálsorð eins og gert var með handmörkun. Hittni markarans fyrir öll orð er því 53.101/59.169 eða 89,74%.

Einnig má athuga hvernig markaranum tekst að greina einstaka greiningarflokka. Þá má nota mælikvarðana nákvæmni (e. *precision*), griphlutfall (e. *recall*) og *F*-gildi. Þessa mælikvarða má nota til þess að kanna hvaða villur markararnir gera. *Nákvæmni* segir til um hversu rétt markarinn greinir tiltekið mark, en *griphlutfall* segir til um hlutfall hvers marks af þeim mörkum sem markarinn finnur (Megyesi 2002: 53–54). Megyesi skilgreinir þessar stærðir þannig fyrir tiltekið mark X:

> *nákvæmni* (*P*)= (*fjöldi rétt greindra lesmálsorða sem hafa mark* X)/(*heildarfjöldi lesmálsorða sem markari greinir með mark* X)
>
> *griphlutfall (R)=(fjöldi rétt greindra orða sem hafa mark X)/(heildarfjöldi orða með mark X í safni)*

F-gildið er vegið umhverfumeðaltal (harmonic mean) af *P* og *R*.
Manning og Schütze (1999: 269) skilgreina *F*-gildi sem
$F=1/(\alpha*(1/P)+(1-\alpha)*(1/R))$
og Megyesi (2002: 32) skilgreinir *F*-gildið sem
$F=(\beta^2+1)*P*R/(\beta^2*P+R)$
Ef β=1 og α=0,5 er *F*-gildið hreint umhverfumeðaltal af *P* og *R*:
$F=P*R/((R+P)/2)=2*P*R/(P+R)$
Í Manning og Schütze (1999: 267–269) er góð lýsing á því hvernig á að finna *P* og *R*.
Tökum sem dæmi að við viljum finna *P*, *R* og *F* fyrir hvernig tiltekinn markari, t.d. TnT, greinir atviksorð í einu prófunarsafni íslenska verkefnisins.
Í prófunarsafninu eru 11.660 atviksorð. TnT greinir 11.451 af þeim rétt (tp, *true positives* samkvæmt Manning og Schütze).
Fjöldi orða sem TnT greinir sem atviksorð = 11.716 (*selected* með orðalagi Manning og Schütze).

Fjöldi orða sem TnT greinir rangt sem atviksorð = 11.716-11.451=265 (fp, *false positives* með orðalagi Manning og Schütze)
Fjöldi orða sem eru atviksorð en TnT greinir sem eitthvað annað = 11.660-11.451=209 (fn=*false negatives* með orðalagi Manning og Schütze)
Þá er
P=tp/(tp+fp)=tp/(valið)=11.451/11.716 = 0,977
R=tp/(tp+fn)=tp/(það sem átti að velja)=11.451/11.660 = 0,982
F=2*P*R/(P+R)= = 0,980
Þessar stærðir má reikna fyrir hvaða greiningarstreng sem er.

Í íslensku hefur ekki skapast sú hefð að gera greinarmun á hittni (*accuracy*) og nákvæmni (*precision*) heldur er orðið *nákvæmni* notað um hvort tveggja. Þar sem ekki er hætta á ruglingi er þeirri hefð fylgt í þessari grein.

6 Prófanir

Allir markararnir sem voru valdir voru þjálfaðir á þjálfunarsöfnunum 10 og prófaðir á samsvarandi prófunarsöfnum. Í upphaflegu tilraununum sem voru gerðar 2002–2004 fengust niðurstöður með þremur mörkurum, TnT, MXPOST og fnTBL. Tilraunin með MBT-markarann var gerð í nóvember 2005 (Sigrún Helgadóttir og Örvar Hafsteinn Kárason 2005).

	Meðalnákvæmni		
Markari	**Óþekkt orð** **%**	**Þekkt orð** **%**	**Öll orð** **%**
fnTBL	54,02	91,36	88,80
MXPOST	62,51	91,04	89,08
TnT	71,62	91,74	90,36
MBT	56,86	89,21	87,00

Tafla 1. Niðurstaða af þjálfun og mörkun 10 para skráa

Niðurstöður prófana eru sýndar í töflu 1. Eins og sést á töflunni eiga markararnir fjórir misjafnlega auðvelt með að greina óþekkt orð, þ.e. orð sem koma ekki fyrir í viðkomandi þjálfunarsafni og þeir hafa því ekki séð áður. Fundið var hlutfall óþekktra orða í hverju prófunarsafni

og reiknað meðaltal fyrir prófunarsöfnin 10 og reyndist það 6,84%. 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 6 atriðum í greiningarstreng sé rangt.

Mismunur á mörkunarnákvæmni TnT og fnTBL er 1,56 prósentustig. Við það að nákvæmni hækkar úr 88,80% í 90,36% fækkar villum um 14%.

Dreifing orðmynda eftir orðflokkum er ólík meðal óþekktra orða og allra orða. Nafnorð, lýsingarorð og sagnir, eru að meðaltali um 44,3% af öllum orðum í prófunarsöfnunum en um 95,9% að meðaltali af óþekktum orðum.

Gert var parað t-próf á hlutfalli rangt greindra orða til þess að kanna hvort tölfræðilega marktækur munur væri á árangri þeirra þriggja markara sem náðu bestum árangri. Niðurstaða prófsins fyrir pörin fnTBL/TnT, MXPOST/TnT og fnTBL/MXPOST er sýnd í töflu 2. Munur á mörkurum er marktækur í öllum tilvikum ($p<0,05$).

Samanburður	**t**	**frítölur**
fnTBL/TnT	40,16	9
MXPOST/TnT	30,94	9
fnTBL/MXPOST	5,37	9

Tafla 2. Parað t-próf á mismuni á hlutfalli rangt greindra orða

7 Greining á niðurstöðum

Niðurstöður þeirra þriggja markara (TnT, MXPOST og fnTBL) sem náðu bestum árangri voru skoðaðar nánar.

Fyrir hvern greiningarstreng var reiknuð nákvæmni (*precision*, *P*), griphlutfall (*recall*, *R*) og *F*-gildi. Niðurstöður útreikninganna eru ekki sýndar hér þar sem þær taka of mikið pláss. Í töflu 3 er sýndur sambærilegur útreikningur fyrir orðflokka.

Markararnir hegða sér á líkan hátt nema fyrir þá orðflokka sem hafa fá orð, þ.e. *e* (erlend orð), *g* (greinir) og *x* (ógreint). TnT fær t.d. hærri nákvæmni en griphlutfall fyrir greininn þar sem markarinn greinir tiltölulega fá orð sem greini en MXPOST fær hærra griphlutfall

en nákvæmni þar sem sá markari greinir fleiri orð sem greini en ættu að fá þá greiningu.

Orðflokkar	Fjöldi í safni	fnTBL			MXPOST			TnT		
		P	R	F (β=1)	P	R	F (β=1)	P	R	F (β=1)
a (atviksorð)	116.112	98,02	98,31	98,16	97,53	98,25	97,89	98,04	98,11	98,07
c (samtengingar)	60.256	98,64	99,05	98,84	98,39	98,95	98,67	98,41	98,92	98,67
e (erlend orð)	411	54,20	37,71	44,48	72,19	56,20	63,20	85,53	63,26	72,73
f (fornöfn)	74.315	98,99	98,71	98,85	99,14	98,25	98,69	98,81	98,84	98,82
g (greinir)	632	82,15	84,49	83,31	78,77	87,50	82,91	94,22	77,37	84,97
l (lýsingarorð)	35.669	89,69	86,15	87,88	88,90	86,00	87,43	93,48	91,74	92,60
n (nafnorð)	122.621	96,31	96,73	96,52	96,63	96,98	96,80	98,48	98,57	98,53
s (sagnorð)	103.136	96,54	97,00	96,77	97,27	97,52	97,39	97,76	98,22	97,99
t (töluorð)	5.901	92,85	95,03	93,93	94,44	93,90	94,17	95,02	93,12	94,06
x (ógreint)	127	63,64	44,09	52,09	70,49	33,86	45,74	58,40	57,48	57,94

Tafla 3. Nákvæmni (P), griphlutfall (R) og F-gildi fyrir orðflokka

Í töflu 4 er griphlutfall greint í sundur eftir því hvort mörkurunum tekst að greina öll atriði í greiningarstreng rétt eða a.m.k. orðflokkinn rétt. Hlutfallstölur eru reiknaðar af heildarfjölda lesmálsorða í orðflokki í safninu.

Fyrsti dálkur fyrir hvern markara sýnir hlutfall rétt greindra strengja af heildarfjölda slíkra strengja í safninu, annar dálkur sýnir hlutfall þar sem orðflokkur er réttur en einhver greiningaratriði röng og síðasti dálkurinn sýnir summu þessara dálka sem er griphlutfallið fyrir orðflokkinn eins og sýnt er í töflu 3. Fyrir utan sjaldgæfa og erfiða orðflokka (*e*, *g* og *x*) virðast allir markararnir eiga í mestum erfiðleikum með að greina lýsingarorð rétt. Þetta virðist eiga við um orðflokkinn sjálfan og einnig virðist erfitt að greina rétt hinar ýmsu greiningarmyndir. Lýsingarorð í íslensku geta fræðilega haft 120 beygingarmyndir. Sumar eru mjög sjaldgæfar þannig að það kemur ekki á óvart að markararnir eigi erfitt með að búa til reglur um hvernig eigi að greina þær.

Orðflokkur	Fjöldi í safni	fnTBL			MXPOST			TnT		
		Grein. str. réttur	Orðfl. réttur	R	Grein. str. réttur	Orðfl. réttur	R	Grein. str. réttur	Orðfl. réttur	R
a (atviksorð)	116.112	93,54	4,77	98,31	92,83	5,41	98,25	92,22	5,89	98,11
c (samtengingar)	60.256	97,71	1,34	99,05	97,09	1,86	98,95	97,14	1,79	98,92
e (erlend orð)	411	37,71	0,00	37,71	56,20	0,00	56,20	63,26	0,00	63,26
f (fornöfn)	74.315	89,38	9,33	98,71	88,15	10,10	98,25	89,46	9,38	98,84
g (greinir)	632	66,77	17,72	84,49	66,14	21,36	87,50	64,72	12,66	77,37
l (lýsingarorð)	35.669	64,09	22,05	86,15	66,99	19,01	86,00	72,88	18,86	91,74
n (nafnorð)	122.621	78,97	17,76	96,73	80,19	16,79	96,98	84,48	14,09	98,57
s (sagnorð)	103.136	91,89	5,12	97,00	92,94	4,58	97,52	92,64	5,58	98,22
t (töluorð)	5.901	69,17	25,86	95,03	71,65	22,25	93,90	73,34	19,78	93,12
x (ógreint)	127	44,09	0,00	44,09	33,86	0,00	33,86	57,48	0,00	57,48

Tafla 4. Sundurliðun griphlutfalls fyrir orðflokka eftir því hvort allur greiningarstrengur er rétt greindur eða a.m.k. orðflokkur. Hlutfallstölur eru reiknaðar af fjölda lesmálsorða í hverjum orðflokki í safni

Niðurstöður mörkunar voru skoðaðar og greindar til þess að finna hvers konar villur markararnir gera og hvernig mætti bæta árangurinn.

Skipta má villum sem markarar gera í tvo flokka. Í fyrsta lagi eru villur sem verða vegna margræðni, þ.e. tiltekin orðmynd getur haft fleiri en eina greiningu. Í öðru lagi verða villur í greiningu óþekktra orða þegar sú aðferð við greiningu óþekktra orða sem markarinn notar gefur ekki rétta greiningu.

Í töflu 5 eru sýndar 20 algengustu villur sem hver markari gerir. Í töflu 6 eru sýndar 20 algengust villur sem allir markarar eru sammála um.

fnTBL				MXPOST				TnT			
markari> rétt mark	Tíðni	%	Safn-tíðni %	markari> rétt mark	Tíðni	%	Safn-tíðni %	markari> rétt mark	Tíðni	%	Safn-tíðni %
aþ>ao	1.568	2,37	2,37	aþ>ao	2.218	3,44	3,44	aþ>ao	1.734	3,05	3,05
ao>aþ	1.522	2,30	4,68	ao>aþ	1.514	2,35	5,79	ao>aþ	1.489	2,62	5,66
nveo>nveþ	830	1,26	5,93	aa>ao	616	0,96	6,75	ao>aa	1.045	1,84	7,50
nveþ>nveo	824	1,25	7,18	ao>aa	599	0,93	7,68	aþ>aa	911	1,60	9,10
sng>sfg3fn	672	1,02	8,19	nveþ>nveo	586	0,91	8,59	nveþ>nveo	887	1,56	10,66
nheo>nhen	594	0,90	9,09	nveo>nveþ	547	0,85	9,44	nveo>nveþ	865	1,52	12,18
nhen>nheo	582	0,88	9,97	sfg3eþ>sfg1eþ	503	0,78	10,22	aa>ao	689	1,21	13,39
sfg3eþ>sfg1eþ	572	0,87	10,84	nhen>nheo	489	0,76	10,98	ssg>sþghen	671	1,18	14,57
aa>ao	562	0,85	11,69	sfg3fn>sng	446	0,69	11,67	nheo>nhen	659	1,16	15,73
nkeo>nkeþ	500	0,76	12,45	c>ct	392	0,61	12,28	nhen>nheo	638	1,12	16,85
aa>aþ	462	0,70	13,14	aa>aþ	378	0,59	12,86	sng>sfg3fn	599	1,05	17,91
ao>aa	449	0,68	13,82	nheo>nhen	371	0,58	13,44	sfg3eþ>sfg1eþ	584	1,03	18,93
lhensf>lheosf	441	0,67	14,49	nkeþ>nkeo	360	0,56	14,00	sþghen>ssg	570	1,00	19,93
nvfo>nvfn	420	0,64	15,13	nvfn>nvfo	337	0,52	14,52	nkeþ>nkeo	509	0,89	20,83
nkeþ>nkeo	412	0,62	15,75	fpkeþ>fpveþ	335	0,52	15,04	lhensf>lheosf	490	0,86	21,69
fohen>foheo	401	0,61	16,36	sng>sfg3fn	334	0,52	15,56	c>aa	437	0,77	22,46
nheog>nheng	392	0,59	16,95	aþ>aa	330	0,51	16,07	nvfo>nvfn	437	0,77	23,23
ct>c	369	0,56	17,51	nkeo>nkeþ	327	0,51	16,58	nkeo>nkeþ	434	0,76	23,99
ssg>sþghen	359	0,54	18,05	ct>c	324	0,50	17,08	nvfn>nvfo	424	0,75	24,73
nvfn>nvfo	356	0,54	18,59	fohen>foheo	321	0,50	17,58	ct>c	393	0,69	25,42

Tafla 5. Tuttugu algengustu villur sem hver markari gerir

Algengustu villur sem markararnir gera eru af fyrri gerðinni, þ.e. orsakast af margræðni. Langalgengustu villurnar felast í því að rugla saman fallstjórn forsetninga. Algengast er að rugla saman þolfalli og þágufalli. Þar sem prófaðir eru gagnamarkarar í þessari rannsókn hafa þeir ekki innbyggðar reglur sem segja til um samræmi í fallstjórn forsetninga og falli eftirfarandi nafnorðs. Markararnir gætu hins vegar búið sér til slíkar reglur út frá gögnunum. Sá þáttur hefur ekki verið kannaður til hlítar. Næstalgengastur er 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*).

Eins og sést af töflu 5 gera markararnir misjafnlega margar villur. Þeir gera einnig misjafnlega fjölbreytilegar villur. TnT gerir 5.373 mismunandi villur, fnTBL gerir 5.897 mismunandi villur og MXPOST gerir 7.115 mismunandi villur. Þar sem markaskrá Orðtíðnibókarinnar er mjög stór er unnt að gera mjög margvíslegar villur. Fræðilega má gera 552*552 = 304.704 mismunandi villur ef tiltekið safn sem á að marka hefur 552 ólík mörk.

Af töflu 5 sést að fyrstu 20 villur sem TnT gerir skýra um 25% af villum sem markarinn gerir, fyrir fnTBL er þessi tala rúmlega 18% og rúmlega 17% fyrir MXPOST.

	Tíðni	**%**	**Safntíðni %**
markari>rétt			
aþ>ao	499	3,82	3,82
sfg3eþ>sfg1eþ	457	3,50	7,32
ao>aþ	361	2,77	10,09
sng>sfg3fn	235	1,80	11,89
nveþ>nveo	214	1,64	13,53
nveo>nveþ	212	1,62	15,15
sfg3eþ>svg3eþ	203	1,55	16,71
ao>aa	190	1,46	18,16
lhensf>lheosf	170	1,30	19,46
fpkeþ>fpveþ	167	1,28	20,74
nhen>nheo	163	1,25	21,99
aa>ao	148	1,13	23,13
fohen>foheo	144	1,10	24,23
ct>c	141	1,08	25,31
c>aa	128	0,98	26,29
nheo>nhen	126	0,97	27,25
nkeo>nkeþ	118	0,90	28,16
nken-m>nkeo-m	112	0,86	29,02
lvensf>lhfnsf	111	0,85	29,87
nkeþ>nkeo	110	0,84	30,71

Tafla 6. Tuttugu algengustu villur sem allir markarar gera

Af töflu 6 sést að næstalgengasta villa sem allir markarar gera sameinginlega er ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum. Þegar litið er á sameiginlegar villur er þessi villa algengari en ruglingur á milli þolfalls og þágufalls eintölu af kvenkynsnafnorð-

um. Ef leitað verður leiða til þess að koma í veg fyrir sumar algengustu villurnar væri e.t.v. skynsamlegt að hafa þennan lista að leiðarljósi.

Mynd 2 sýnir hvernig þrír markarar, TnT, MXPOST og fnTBL greina orð í setningunni sem einnig er sýnd í mynd 1. Þar sjást vel dæmi um helstu atriði sem markararnir eiga erfitt með að greina rétt. TnT og fnTBL markararnir greina báðir rangt fallstjórn orðsins *eftir* og einnig rangt fall nafnorðsins *strætó* en það er eins í nefnifalli, þolfalli og þágufalli. Allir markararnir greina rangt persónu sagnarinnar *veifaði*. Í töflu 6 sást einmitt að önnur algengasta sameiginlega villa allra markaranna er að rugla saman fyrstu og þriðju persónu sagna í þátíð.

orð	**mark**	**tnt**	**mxp**	**fnt**
ég	fp1en	fp1en	fp1en	fp1en
stökk	sfg1eþ	sfg1eþ	sfg1eþ	sfg1eþ
á	aa	aa	aa	aa
eftir	aþ	**ao**	aþ	**aa**
strætó	nkeþ	**nkeo**	nkeþ	**nkeo**
og	c	c	c	c
veifaði	sfg1eþ	**sfg3eþ**	**sfg3eþ**	**sfg3eþ**
,	,	,	,	,
vagnstjórinn	nkeng	nkeng	nkeng	nkeng
sá	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
mig	fp1eo	fp1eo	fp1eo	fp1eo
og	c	c	c	c
stoppaði	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
.	.	.	.	.

Mynd 2. Greining þriggja markara á orðum í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Fundið var hversu oft markararnir þrír voru sammála og fengu annaðhvort rétta eða ranga niðurstöðu. Tafla 7 sýnir hversu oft allir þrír markarar komast að réttri niðurstöðu, hversu oft tveir komast að réttri niðurstöðu og hversu oft aðeins einn hefur rétt fyrir sér. Í 95,47% tilvika hefur a.m.k. einn markari fundið rétt mark. Það er því hæsta fræðilega nákvæmni sem má ná fyrir þann efnivið sem prófaður var með því að sameina niðurstöður tveggja eða fleiri markara.

	Tíðni	**%**	**Safntíðni %**
3 réttir	484.294	82,04	82,04
2 réttir	51.322	8,69	90,74
1 réttur	27.941	4,73	95,47
Enginn réttur	26.740	4,53	100,00

Tafla 7. Hversu margir markarar eru sammála um rétt mark?

Í töflu 8 er sýndur paraður samanburður á mörkurum. Þar sést að TnT og fnTBL eru oftar sammála um rétt mark (og rangt mark) en önnur pör. Það gæti bent til þess að niðurstöður TnT og fnTBL séu með einhverjum hætti líkar þó að TnT gefi umtalsvert betri niðurstöðu. Það gæti því verið að unnt sé að bæta niðurstöðu TnT með niðurstöðu MXPOST.

Par	**Sama mark rétt** %	**Sama mark rangt** %	**Samtals** %
TnT og MXPOST	85,11	3,03	88,14
TnT og fnTBL	85,56	3,64	89,20
MXPOST og fnTBL	84,15	3,14	87,29

Tafla 8. Samanburður á markarapörum

8 Frammistaða markara bætt

Ýmsum aðferðum má beita til þess að bæta niðurstöður mörkunar. Stundum er reynt að bæta frammistöðu einstakra markara og einnig má sameina niðurstöðu tveggja eða fleiri markara. Í þeirri könnun sem hér er greint frá var gerð tilraun til þess að nota orðasafn til þess að bæta frammistöðu einstakra markara. Einnig var beitt tveimur aðferðum við að sameina niðurstöður markara.

8.1 Áhrif aukaorðasafns á mörkun

Við tilraunina voru notuð forritin TnT og fnTBL þar sem þau gefa kost á að nota viðbótarorðasafn.

Búið var til orðasafn sem hefur um helming þeirra orða sem eru óþekkt í hverju prófunarsafni miðað við samsvarandi þjálfunarsafn og það notað sem viðbótarorðasafn við mörkun með TnT og fnTBL.

Orðasafnið var gert þannig að búinn var listi yfir orð í hverju prófunarsafni sem voru óþekkt miðað við samstætt þjálfunarsafn og listarnir síðan sameinaðir í eitt safn. Síðan var tekið annað hvert orð úr þessu safni og notað sem viðbótarorðasafn. Safnið ætti að geyma um helming óþekktra orða í hverju prófunarsafni. Í töflu 9 sést niðurstaða fyrir mörkun með þessu orðasafni. Til samanburðar eru tölur fyrir mörkun án orðasafns hafðar með í töflunni.

	Meðalnákvæmni án orðasafns			**Meðalnákvæmni með orðasafni***		
Markari	**Óþekkt orð %**	**Þekkt orð %**	**Öll orð %**	**Óþekkt orð %**	**Þekkt orð %**	**Öll orð %**
fnTBL	54,02	91,36	88,80	70,44	91,50	90,06
TnT	71,62	91,74	90,36	86,31	91,93	91,54
*Notað er orðasafn sem hefur um helming þeirra orða sem álitin eru óþekkt frá sjónarhóli hvers prófunarsafns						

Tafla 9. Niðurstaða af þjálfun og mörkun 10 para skráa

Mörkun óþekktra orða batnar umtalsvert og hefur það áhrif á heildarniðurstöðu. Mörkun þekktra orða batnar einnig aðeins og er það sennilega afleiðing af bættri mörkun óþekktu orðanna. Þegar fleiri óþekkt orð fá rétta greiningu gefa þau betri vísbendingar um rétta mörkun þekktu orðanna í kring. Heildarnákvæmni með mörkun fnTBL hækkar meira en heildarnákvæmni með TnT. Ástæðan gæti verið sú að fnTBL-markarinn virðist eiga erfiðara með að marka óþekkt orð og þess vegna batnar mörkun óþekktra orða ef orðasafn er til staðar til þess að greina þau. Með því að nota viðbótarorðasafn nær TnT-markarinn **91,54%** nákvæmni og villum fækkar um 12%. Þessar niðurstöður sýna að mörkun ætti að batna ef unnt er að nota orðasafn. Þeir markarar sem voru prófaðir nota orðasöfn sem hafa tiltekið snið. Nauðsynlegt er að í viðbótarorðasafni séu upplýsingar um hlutfallslegt vægi mismunandi greiningarstrengja þeirra orðmynda sem geta haft fleiri en einn greiningarstreng.

8.2 Sameina niðurstöður markara

Nefna má þrjár aðferðir sem koma til greina við að sameina niðurstöður tveggja eða fleiri markara.

1. Kosið er um hvaða markari er valinn

2. Nýr markari er þjálfaður á grundvelli niðurstaðna úr tveimur eða fleiri mörkurum

3. Notaðar eru málfræðireglur

Í þessu verkefni voru prófaðar tvær af þessum aðferðum, þ.e. að kjósa á milli markara og að nota málfræðireglur.

Í Halteren o.fl. (2001) er yfirlit yfir aðferðir við að sameina niðurstöður tveggja eða fleiri markara. Markmiðið er að ná meiri nákvæmni en fæst með þeim einstökum markara sem gefur bestar niðurstöður. Í greininni er gerð grein fyrir tveimur mismunandi aðferðum við að sameina niðurstöður. Í fyrsta lagi er greint frá nokkrum aðferðum við að kjósa á milli markara. Í öðru lagi er greint frá leiðum til þess að þjálfa nýjan markara á grundvelli niðurstaðna markaranna og rétts marks. Í þessari rannsókn voru aðeins prófaðar aðferðir við að kjósa á milli markara.

Í því verki sem hér er lýst var gerð tilraun með fjögur afbrigði af kosningaaðferðinni. Einfaldasta aðferðin byggist á því að velja það mark sem flestir velja. Ef ekki er unnt að velja þannig er notuð slembitala til þess að velja á milli marka. Annað afbrigðið byggist á því að vega með fyrir fram þekktri heildarnákvæmni hvers markara. Í þriðja afbrigðinu er vegið með nákvæmni fyrir hvert mark. Einnig má vega með nákvæmni og griphlutfalli hvers marks. Vegið er með *nákvæmni* sem segir til um hvernig markarinn stendur sig og (*1-griphlutfalli*) sem segir til um hve oft markaranum mistekst að finna rétta markið. Hvert mark fær *nákvæmni* (precision) þess markara sem leggur markið til og (*1-griphlutfall*) marksins hjá þeim mörkurum sem leggja það ekki til.

Hæsta nákvæmni fékkst með því að nota heildarnákvæmni sem vog. Er það sama niðurstaða og fékkst í Halteren o.fl. (2001) fyrir hollenskan texta. Í töflu 10 eru sýndar niðurstöður kosningar. Þar sést að með því að kjósa milli markara og vega með heildarnákvæmni markaranna fæst **91,54**% nákvæmni. Það er marktækt hærri niðurstaða en fæst með því að nota TnT-markarann eingöngu ($p<0{,}001$). Notaðir eru þrír markarar í íslensku tilrauninni. Allar aðferðir þar sem kosningu er beitt felast því í eftirfarandi: Valið er það mark sem tveir eða fleiri eru sammála um. Ef allir eru ósammála er beitt mismunandi aðferðum við að velja markið. Þegar beitt er meirihlutakosningu er mark valið af handahófi. Þegar vegið er með heildarnákvæmni (accuracy) markarans er valið mark þess markara sem hefur hæsta heildarnákvæmni,

í þessu tilviki mark TnT. Þegar vegið er með nákvæmni hvers marks fyrir hvern markara er valið það mark sem fær hæsta nákvæmni. Þegar vegið er með nákvæmni og griphlutfalli er valið það mark sem fær hæsta summu af nákvæmni þess markara sem leggur markið til og (1-griphlutfall) marksins hjá þeim mörkurum sem leggja það ekki til.

	Óþekkt orð		**Þekkt orð**		**Öll orð**	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Alls	40.392	100,00	549.905	100,00	590.297	100,00
MXPOST	25.246	62,50	500.617	91,04	525.863	89,08
fnTBL	21.823	54,03	502.378	91,36	524.201	88,80
TnT	28.919	71,60	504.484	91,74	533.403	90,36
Meirhlutakosning	27.889	69,05	510.903	92,91	538.792	91,27
Vegið með heildarnákvæmni	29.003	71,80	511.348	92,99	540.351	91,54
Vegið með með nákv. marks	27.808	68,85	511.088	92,94	538.896	91,29
Vegið með nákv. og griphlutfalli.	28.738	71,15	511.440	93,01	540.178	91,51
Vegið með heildarnákvæmni*	34.331	84,97	512.044	93,12	546.375	92,56
* Kosið um mörk þegar viðbótarorðasafn er notað við mörkun með TnT og fnTBL						

Tafla 10. Nákvæmni þriggja markara og nákvæmni sem fæst með þremur mismunandi aðferðum við að kjósa á milli niðurstöðu markaranna

Einnig var gerð tilraun til þess að kjósa um mörk þegar viðbótarorðasafn var notað við mörkun með TnT og fnTBL. Neðsta línan í töflu 10 sýnir niðurstöðu þegar vegið er með heildarnákvæmni TnT en þá fæst **92,56%** nákvæmni. Í töflu 9 sést að nákvæmni þegar markað er með TnT og viðbótarorðasafn notað er 91,54%. Með því að kjósa á milli markaranna fækkar villum um 12% frá þeirri niðurstöðu sem fæst með TnT eingöngu.

Í töflu 11 er niðurbrotinn samanburður á mörkurunum þremur. Þar sést að líklegast er að TnT gefi rétta niðurstöðu ef markararnir gefa ólíkar niðurstöður. Af töflunni sést enn fremur að markararnir TnT og fnTBL eru í einhverjum skilningi líkari heldur en TnT og MXOPOST eða fnTBL og MXPOST. Þess vegna er líklegt að nota megi niðurstöðu MXPOST til þess að bæta niðurstöðu mörkunar.

	Tíðni	**%**	**Safntíðni %**
allir eins og réttir	484.294	82,04	82,04
allir eins og rangir	13.055	2,21	84,25
tnt=fnt=rétt, mxp=rangt	20.783	3,52	87,77
tnt=fnt=rangt, mxp=rétt	8.434	1,43	89,20
tnt=mxp=rétt, fnt=rangt	18.112	3,07	92,27
tnt=mxp=rangt, fnt=rétt	4.850	0,82	93,09
fnt=mxp=rétt, tnt=rangt	12.427	2,11	95,20
fnt=mxp=rangt, tnt=rétt	5.479	0,93	96,13
allir ólíkir, tnt=rétt	4.735	0,80	96,93
allir ólíkir, fnt=rétt	1.847	0,31	97,24
allir ólíkir, mxp=rétt	2.596	0,44	97,68
allir ólíkir og rangir	13.685	2,32	100,00
Samtals	590.297	100,00	

Tafla 11. Samanburður á mörkurum

Lars Borin (2000) hefur rannsakað hvernig megi endurnota efnivið og tungutæknitól, sem þegar eru til, á nýjan hátt. Hann skoðar hvernig megi nota tilbúna markara á efni sem þeir voru ekki þjálfaðir fyrir og þar sem ekki er til reiðu þjálfunarsafn. Borin bendir einnig á hvernig sameina megi niðurstöður markara fyrir þýsku með því að nota málfræðilegar reglur þannig að mörkunarnákvæmni sameinaðra markara verði hærri en nákvæmni þess markara sem nær mestri nákvæmni.

Þó að þessar aðstæður eigi ekki fullkomlega við íslenska verkefnið var aðferðin könnuð nánar.

Tvennt þarf að vera til staðar til þess að unnt sé að bæta nákvæmni með því að sameina niðurstöður tveggja eða fleiri markara.

1. Markararnir gera ekki sömu vitleysurnar, þ.e. þeir bæta hver annan upp (*complementarity*)

2. Mismunur er kerfisbundinn en ekki tilviljunarkenndur

Borin flokkar þá aðferð sem hann leggur til sem „knowlege-rich“, þ.e. rannsakendur þekkja gögnin vel. Málfræðilegar reglur eru skilgreindar til þess að nýta mismun markara til þess að sameina niðurstöður þeirra. Borin setti fram þessar tilgátur:

1. Þegar markararnir eru sammála hafa þeir örugglega rétt fyrir sér.

2. Villur sem markararnir gera eru ólíkar. Í mörgum tilvikum hefur annar markarinn rétt fyrir sér en hinn rangt (Borin skoðaði tvo markara). Mikilvægt er að sá markari sem gefur lægri nákvæmni hafi stundum rétt fyrir sér í slíkum tilvikum.

3. Mismunur á milli markaranna er kerfisbundinn á einhvern hátt. Þennan kerfisbundna mismun má nota til þess að bæta mörkun með því að sameina niðurstöður markaranna.

Fyrsta tilgátan var ekki prófuð. Í töflu 11 sést þó að allir þrír markarar voru sammála og höfðu rétt fyrir sér í 82,04% tilvika og voru allir sammála en höfðu rangt fyrir sér í 2,21% tilvika, þegar prófaðir voru þrír markarar (MXPOST, fnTBL og TnT) í íslensku rannsókninni. Það má því ekki ganga út frá því sem gefnu að niðurstaða sé rétt þó að allir markararnir séu sammála.

Gerð var tilraun til þess að líta á niðurstöðu kosningar sem útkomu úr markara. Athugað var hvort nota mætti niðurstöðu MXPOST, fnTBL eða TnT til þess að bæta þá niðurstöðu. Hæsta nákvæmni, 91,54%, fékkst þar sem kosið var um mörk sem þrír markarar höfðu úthlutað og vegið með heildarnákvæmni þess markara sem hafði staðið sig best, í þessu tilviki TnT. Í töflu 12 sést samanburður á þessari niðurstöðu og niðurstöðum markaranna þriggja.

Á töflunni sést að niðurstöður MXPOST myndu bæta mestu við niðurstöðu með kosningu og gefa 96,37% nákvæmni ef tækist að finna reglur til þess að nýta öll tilvik þar sem MXPOST gefur rétta niðurstöðu en kosning ranga. Með kosningu er þegar búið að nýta kosti TnT og því ekki líklegt að unnt sé að gera betur með þeim markara.

Kannað var hvaða reglum mætti beita til þess að nýta þau tilvik þar sem MXPOST getur gert betur en útkoma úr kosningu gefur. Skoðuð voru tilvik þar sem mark sem kosning gefur er ólíkt marki MXPOST. Fundið var hversu oft MXPOST gefur betri niðurstöðu en kosning í þessum tilvikum.

Kosning vs. MXPOST	**Tíðni**	**%**	**Safntíðni**
Bæði mörk rétt	514.833	87,22	87,22
Kosning rétt, MXPOST rangt	28.509	4,83	92,05
Kosning röng, MXPOST rétt	25.518	4,32	96,37
Mörk lík og röng	11.030	1,87	98,24
Mörk ólík og röng	10.407	1,76	100,00
Kosning vs. fnTBL			
Bæði mörk rétt	517.504	87,67	87,67
Kosning rétt, fnTBL rangt	34.230	5,80	93,47
Kosning röng, fnTBL rétt	22.847	3,87	97,34
Mörk lík og röng	6.697	1,13	98,47
Mörk ólík og röng	9.019	1,53	100,00
Kosning vs. TnT			
Bæði mörk rétt	527.924	89,43	89,43
Kosning rétt, TnT rangt	42.237	7,16	96,59
Kosning röng, TnT rétt	12.427	2,11	98,69
Mörk lík og röng	5.479	0,93	99,62
Mörk ólík og röng	2.230	0,38	100,00

Tafla 12. Samanburður á útkomu markaranna og niðurstöðu kosningar þegar vegið er með heildarnákvæmni

Gert var yfirlit yfir þau tilvik þar sem það að velja mark MXPOST fram yfir útkomu úr kosningu fækkar villum. Flestar villurnar lúta að ruglingi milli falla nafnorða og lýsingarorða. Einnig er þar að finna rugling milli greiningarmynda sagnorða. Ákveðið var að nota útkomu MXPOST fyrir tiltekna samsetningu ef MXPOST gæfi rétta greiningu fram yfir kosningu oftar en 5 sinnum. Reglurnar eru í forminu:

> *ef útkoma úr kosningu er mark1 og útkoma MXPOST er mark2*
> *þá skal velja mark2*

Þegar reglur voru valdar þannig að niðurstaða batnaði um meira en 5 mörk við það að beita reglunni fékkst nákvæmni fyrir öll orð 91,81%, nákvæmni fyrir óþekkt orð 72,13% og fyrir þekkt orð 93,25%.

Einnig var gerð tilraun með að beita reglum þegar upprunaleg mörkun með TnT og fnTBL var gerð með aðstoð orðasafns. Í töflu 10 sést að þegar kosið er um mörk markaranna þriggja sem þannig eru fengin fæst 92,56% nákvæmni. Fundnar voru reglur til þess að velja

mark MXPOST umfram útkomu úr kosningu og fékkst þá **92,69%** nákvæmni.

8.3 Áhrif markaskrár

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 eins og sjá má í viðauka A. 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.

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.

	Meðalnákvæmni fnTBL			Meðalnákvæmni MXPOST			Meðalnákvæmni TnT		
	Rétt (fj.)	%	Safntíðni (%)	Rétt (fj.)	%	Safntíðni (%)	Rétt (fj.)	%	Safntíðni (%)
Allur greiningarstrengur réttur	524.201	88,80	88,80	525.863	89,08	89,08	533.403	90,36	90,36
Atviksorð ekki greind	5.533	0,94	89,74	6.286	1,06	90,15	6.837	1,16	91,52
Samtengingar ekki greindar	806	0,14	89,88	1.118	0,19	90,34	1.076	0,18	91,70
Öllum fornöfnum slegið saman	600	0,10	89,98	741	0,13	90,46	782	0,13	91,83
Aðeins orðflokkur réttur	42.900	7,27	97,25	40.310	6,83	97,29	37.197	6,30	98,14
Rangur orðflokkur	16.257	2,75	100,00	15.979	2,71	100,00	11.002	1,86	100,00
Samtals	590.297	100,00		590.297	100,00		590.297	100,00	

Tafla 13. Nákvæmni mörkunar þegar markaskrá er einfölduð

Í töflu 13 er sýnd nákvæmni markaranna þegar mörk eru einfölduð á þennan hátt. Af töflunni sést að með því að sleppa greiningu atviksorða hækkar nákvæmni TnT úr 90,36% í 91,52%, villum fækkar um 12%. Með því að sleppa einnig greiningu samtenginga og slá saman fornafnaflokkum fer nákvæmni TnT í 91,83%.

Ef aðeins er litið á greiningu eftir orðflokkum nær TnT **98,14%** nákvæmni. Í sumum tungutækniverkefnum gæti greining eftir orðflokkum dugað og þá gefur TnT viðunandi niðurstöðu.

9 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 að-

skilin lítil textasöfn voru notuð[4]. 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 9 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.

Í töflu 14 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.

[4] Aðalsteinn Eyþórsson tók saman efni í þessi textasöfn.

	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Gamall bókmenntatexti						
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						
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						
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						
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

Tafla 14. Nákvæmni við mörkun texta sem eru ekki í textasafni Orðtíðnibókar

10 Niðurstöður og umræða

Hér á undan hefur verið greint frá tilraunum við að marka íslenskan texta með ýmsum aðferðum sem hafa verið þróaðar fyrir önnur tungumál. Fjórir markarar voru þjálfaðir og prófaðir á íslenskum texta og reynt var að finna aðferðir til þess að bæta niðurstöðu markaranna. Gerðar voru tilraunir með að nota orðasafn við mörkun, að kjósa á milli markaranna og að beita málfræðilegum reglum til þess að velja tiltekið mark fram yfir annað mark. Einnig var sýnt að með því að einfalda mörk mætti ná betri niðurstöðu. Það virðist skipta máli í hvaða röð aðgerðunum er beitt. Í töflu 15 er gefið yfirlit yfir helstu niðurstöður af því að sameina aðferðir.

	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Orðasafn notað við mörkun með fnTBL og TnT[5]						
fnTBL	28.461	70,44	503.142	91,50	531.603	90,06
MXPOST	25.252	62,50	500.611	91,04	525.863	89,08
TnT	34.859	86,28	505.511	91,93	540.370	91,54
Mörk einfölduð[6]						
fnTBL	28.467	70,46	509.788	92,71	538.255	91,18
MXPOST	25.261	62,52	508.747	92,52	534.008	90,46
TnT	34.863	86,29	513.797	93,44	548.660	92,95
Vegið með heildarnákvæmni	34.336	84,98	517.773	94,16	552.109	93,53
MXPOST fram yfir kosn. m. heildarnkv.	34.013	84,18	518.818	94,35	552.831	**93,65**

Tafla 15. Nákvæmni við mörkun íslensks texta þegar fjórum aðgerðum er beitt í röð til þess að bæta niðurstöðu mörkunar þriggja markara

Sýndar eru niðurstöður miðað við að notað sé orðasafnið sem var búið til þegar markað er með TnT og fnTBL. Hæsta nákvæmni, **93,65%**, fæst með því að nota orðasafn, einfalda mörk markaranna, kjósa á milli einfaldaðra marka og beita síðan reglum sem velja mark MXPOST þegar tilteknum skilyrðum er fullnægt. Villum fækkar um 34% miðað við niðurstöðu mörkunar með TnT eingöngu.

Niðurstaða sem fæst með því að nota hjálparorðasafn við mörkun með TnT og fnTBL sýnir að villum mun fækka þegar orðasafn er notað. Það fer að sjálfsögðu eftir eðli textanna sem á að marka og stærð hjálparorðasafnsins hversu mikið nákvæmni eykst við það. Með þeim efnivið sem hér var til ráðstöfunar er þó ljóst að þær aðferðir sem hafa verið prófaðar geta gefið um 92% nákvæmni fyrir texta sem eru líkir textum Orðtíðnibókarinnar.

[5]Orðasafn hefur u.þ.b. helming óþekktra orða

[6]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.

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, mörk þeirra og hlutfallsleg tíðni einstakra greiningarmynda. Nota má *Beygingarlýsingu íslensks nútímamáls* (Kristín Bjarnadóttir 2004), sem einnig var gerð var 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. Einnig væri æskilegt að kanna frekar hvers konar markaskrá sé heppileg fyrir hin ýmsu verkefni.[7]

Aðferðirnar voru einnig prófaðar á textum sem voru ekki hluti af textasafni Orðtíðnibókarinnar. Þá kom í ljós að TnT-markarinn nær bestum árangri við mörkun allra textanna. Aðrir markarar náðu svo lélegum árangri að ekki reyndist unnt að bæta niðurstöðu mörkunar með því að nýta þá.

Heimildir

Borin, Lars. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. *Second International Conference on Language Resources and Evaluation*, Athens 31 May – 2 June, 2000, bls. 21–26.

Brants, Thorsten. 2000a. TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, bls. 224–231. Seattle, Washington, USA.

Brants, Thorsten. 2000b. TnT - A Statistical Part-of-Speech Tagger. Version 2.2. *http://www.coli.uni-sb.de/~thorsten/tnt/*

Brill, Eric. 1994. Some Advances in Rule-Based Part of Speech Tagging. *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, bls. 722–727. Seattle, Washington.

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, December 1995: 543–563.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. *MBT: Memory-Based Tagger, Reference Guide. ILK Technical Report 03-13, http://ilk.uvt.nl/downloads/pub/papers/ilk.0313.pdf*

[7] Eftir að þessu verki lauk formlega bjó Hrafn Loftsson (2006) til málfræðilegan reglumarkara 2004–2005 og notaði texta *Íslenskrar orðtíðnibókar* við prófun. Hrafn náði 91,471% nákvæmni í mörkun með reglumarkara sínum (*IceTagger*). Með því að sameina niðurstöður fjögurra markara náði Hrafn 92,94% mörkunarnákvæmni.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02, http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf*

Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir. 2002. Vélræn málfræðigreining með námfúsum markara. *Orð og tunga* 6:1–9.

Florian, Radu and Grace Ngai. 2002. Fast Transformation-Based Learning Toolkit. *http://nlp.cs.jhu.edu/~rflorian/fntbl/tbl-toolkit/tbl-toolkit.html*

Friðrik Magnússon. 1988. Hvað er títt? Tíðnikönnun Orðabókar Háskólans. *Orð og tunga* 1:1–49.

Van Halteren, Hans, Jakub Zavrel and Walter Daelemans. 2001. Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems. *Computational Linguistics* 27 (2), bls. 199–230.

Hrafn Loftsson. 2006. Tagging Icelandic text: A linguistic rule-based approach. Technical Report CS-06-04, Department of Computer Science, University of Sheffield.

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

Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. *Samspil tungu og tækni*. Menntamálaráðuneytið, Reykjavík.

Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, Massachusetts. London, England.

Megyesi, Beata. 2002. Data-Driven Syntactic Analysis – Methods and Applications for Swedish. Ph.D.Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Ratnaparkhi, A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Mehods in Natural Lanugage Processing* (EMNLP-96), bls. 133–143. Philadelphia. PA.

Ratnaparkhi, A. 1997. A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical Report 97-08, Institute for Research in Cognitive Scince, University of Pennsylvania.

Rögnvaldur Ólafsson, Þorgeir Sigurðsson, Eiríkur Rögnvaldsson. 1999. *Tungutækni*. Skýrsla starfshóps. Menntamálaráðuneytið.

Samuelsson, Christer. 1993. Morphological tagging based entirely on Bayesian inference. *9th Nordic Conference on Computational Linguistics NODALIDA-93,* bls. 225–238. Stockholm University, Stockholm, Sweden.

Sigrún Helgadóttir. 2002. The Icelandic μTBL Experiment: Learning rules from four different training corpora by using the μ-TBL System – Further developments. Term paper in NLP 1, GSLT.

Sigrún Helgadóttir and Örvar Kárason. 2005. Memory-Based Learning Assignment. Term paper in Machine Learning, GSLT.

Stefán Briem. 1990. Automatisk morfologisk analyse af islandsk tekst. Jörgen Pind og Eiríkur Rögnvaldsson (ritstj.). *Papers from the Seventh Scandinavian Conference of Computational Linguistics Reykjavík 1989*:3–13. Institute of Lexicography, Institute of Linguistics, Reykjavík.

Lykilorð:

mark, mörkun, markari

Keywords:

part-of-speech tag, tagging, tagger

Abstract

This paper gives the results on the automatic tagging of Icelandic text, using a corpus that was prepared for the making of the *Icelandic Frequency Dictionary*. The corpus contains 590,297 running words with 59,358 word forms, including punctuation. Each running word has been supplied with a morphosyntactic tag and the tagset contains 639 tags, including punctuation tags. Five different data-driven taggers, fnTBL, TnT, MXPOST, μ-TBL and MBT were trained on the corpus by using ten-fold cross-validation. The TnT tagger obtained best results for tagging or 90.36% accuracy. The TnT and fnTBL systems allow the use of a backup lexicon. When using such a lexicon TnT reached 91.54% tagging accuracy and fnTBL 90.06%. Methods for combining the results of the taggers were also tested. A voting method where each tagger votes its overall precision gave best result of the voting methods tested or 91.54% accuracy. By utilizing the ability of the MXPOST tagger to distinguish between noun cases, rules were composed to increase tagging accuracy to 91.81%. By using a special strategy for simplifying tags, the TnT tagger gave 91.83% tagging accuracy. Finally, the different strategies for improving tagging accuracy were applied in a certain order. The best result, 93.65% accuracy, was obtained by tagging with a backup lexicon with fnTBL and TnT, simplifying the resulting tags, voting between the simplified tags and applying rules based on MXPOST. Compared with the result obtained with TnT alone, the number of errors is reduced by 34%. By using a lexicon derived from the Morphological Description of Modern Icelandic as a backup lexicon the accuracy can be further increased. Finally an experiment was made in tagging texts that are not a part of the corpus of the *Icelandic Frequency Dictionary*.

Sigrún Helgadóttir
Stofnun Árna Magnússonar í íslenskum fræðum
Neshaga 16
IS-107 Reykjavík
sigrunh@lexis.hi.is

Viðauki A

Skýring skammstafana í greiningarstrengjum Íslenskrar 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

Erla Hallsteinsdóttir

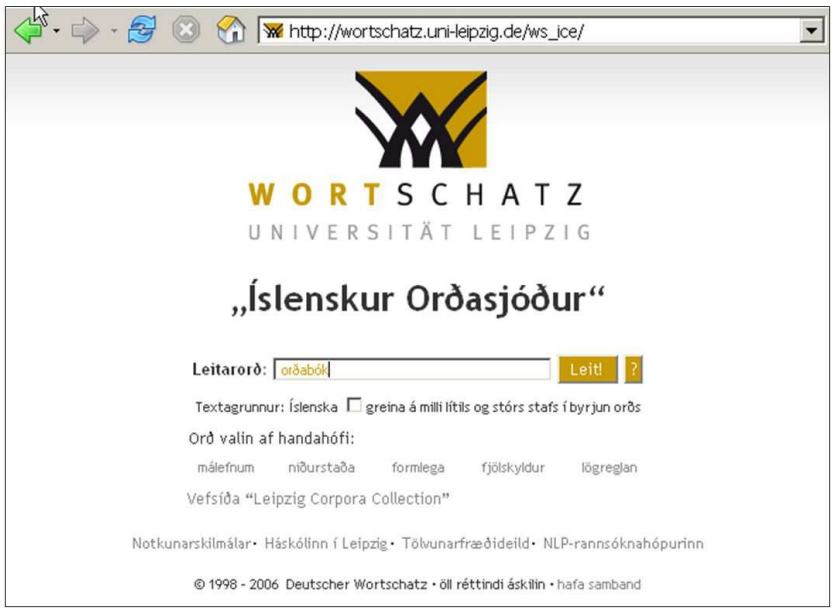
Íslenskur orðasjóður

1 Inngangur

Í þessari grein mun ég lýsa tilurð og eiginleikum íslensks textagrunns – Íslensks orðasjóðs – sem unnið er að við Háskólann í Leipzig. Íslenskur orðasjóður er íslenskur textagrunnur með innbyggðu orðasafni sem samanstendur af u.þ.b. 250 milljónum orða og orðmynda úr íslensku nútímamáli. Íslenskur orðasjóður var unninn við Háskólann í Leipzig sem hluti af rannsóknarvinnu við verkefnið Leipzig Corpora Collection (http://corpora.informatik.uni-leipzig.de/). Textarnir í texta grunninum eru úr vefsíðusöfnun Landsbókasafns Íslands – Háskólabókasafns haustið 2005 og notendaumhverfið var þróað við tölvunarfræðideild Háskólans í Leipzig, sbr. mynd 1 (hér sem þróunarútgáfa með einum árgangi af *Morgunblaðinu*: http://wortschatz.uni-leipzig.de/ws_ice/index.php):

Textagrunnurinn verður birtur á Veraldarvefnum til frjálsrar notkunar. Í þessari grein mun ég lýsa notkunarmöguleikum textagrunnsins sem orðasafns fyrir almenna notendur og sem rannsóknartækis fyrir rannsóknir bæði í hagnýtum og almennum málvísindum.

Orð og tunga 9 (2007), 109–124.



Mynd 1: Vefsíða Íslensks orðasjóðs.

2 Textagrunnar: „orðasjóðir“

2.1 Orðasjóðir í Leipzig

Deutscher Wortschatz („Þýskur orðasjóður“) er textagrunnur með innbyggðri orðabók (málheild þýsks ritmáls með leitarmöguleikum) á Veraldarvefnum sem gerð var og viðhaldið er af Uwe Quasthoff og samstarfsfólki hans við Háskólann í Leipzig. Þýski orðasjóðurinn inniheldur nú texta m.a. frá Institut für Deutsche Sprache í Mannheim, stórum þýskum bókaforlögum (t.d. dtv, Reclam, Walter de Gruyter), frá ýmsum stærri blöðum og tímaritum (t.d. *Der Spiegel*, *TAZ*, *Süddeutsche Zeitung*, *Die Zeit*), opinberum aðilum (t.d. þýsku ríkisstjórninni, Frauenhofer-stofnununum). Vefsíður þýska orðasjóðsins (*www.wortschatz.uni-leipzig.de*) eru á þýsku en textagrunna 17 annarra tungumála er að finna á ofangreindu vefsvæði með enskum skýringum. Að auki er hægt að nálgast afmarkaða textagrunna á 15 tungumálum til notkunar í rannsóknum á vefsvæðinu http://corpora.informatik.uni-leipzig.de/download.html. Aðalmarkmiðið með þessum textagrunnum er

að veita frjálsan aðgang að stöðluðum, sambærilegum málgögnum og tölfræðilegum upplýsingum um þessi gögn (sbr. Quasthoff, Richter og Biemann 2006).

Hugmyndin að íslenskum textagrunni kom fram í framhaldi af samstarfi mínu við Uwe Quasthoff um rannsókn á tíðni þýskra orðtaka[1] í textagrunninum Deutscher Wortschatz. Tæknilega hliðin á Íslenskum orðasjóði var prófuð með einum árgangi af *Morgunblaðinu* (1 milljón orða) og næsta skref er að skipta um textasafn og ná tökum á að tengja beygingarform íslensku orðanna við uppflettiorðið. Stefnt er að því að nota til þess beygingarlýsingu fyrir íslensku sem unnin var sem tungutækniverkefni við Orðabók Háskólans.

2.2 Textagrunnur: Vefsíðusafn Landsbókasafns Íslands

Gerður hefur verið samningur við Landsbókasafn Íslands – Háskólabókasafn um að nota safn íslenskra vefsíðna (vefsíður sem enda á.is) sem grunn í Íslenskan orðasjóð. Textagrunnurinn verður byggður upp á sama hátt og Deutscher Wortschatz í Leipzig.

Töluvert magn erlendra texta var í vefsíðusafninu. Það olli þó engum vandræðum við úrvinnslu þess þar sem þróuð hefur verið mjög virk aðferð til að finna og fjarlægja önnur tungumál og „rusl“ eins og leifar af forritunarmálum úr textagrunnum (sbr. Quasthoff og Biemann 2006). Eftir að hafa notað þessa aðferð stendur eftir íslenskur textagrunnur með u.þ.b. 400 milljónum orðmynda. Ótvíræður kostur við vefsíðusafnið er að það inniheldur íslenskt nútímamál í víðum skilningi, bæði texta frá opinberum aðilum og einkaaðilum, texta sem fylgja reglum ritmáls og texta sem telja má vera talmál á rituðu formi.

2.3 Notendahópur og notkunargildi „Íslensks orðasjóðs“

Íslenskur orðasjóður hefur tvenns konar notkunarmöguleika:

[1]Rannsóknin var unnin sem hluti af rannsóknastöðuverkefni RANNÍS við Hugvísindadeild HÍ 2001–2004, einnig styrkt af Launasjóði fræðiritahöfunda 2005, vinnan við íslenska textagrunninn er styrkt af DAAD. Mig langar að koma á framfæri þakklæti til þessara aðila fyrir fjárhagslegan stuðning.

(1) Orðasafn Íslensks orðasjóðs er „öðru vísi“ orðabók sem inniheldur upplýsingar um heildartíðni og hlutfallslega tíðni orðmynda, notkunardæmi, merkingar- og textaumhverfi, þ.e. hægri og vinstri nágranna í textum og orð með marktæka tíðni í sömu setningum og leitarorðið. Orðasafnið er í íslensku notendaumhverfi sem ætlað er almennum íslenskum notendum.

(2) Textagrunnar eru mikilvægt hjálpartæki í tungumálarannsóknum og tungutækniverkefnum. Íslenskur orðasjóður er einn umfangsmesti textagrunnur á íslensku sem er ætlaður til notkunar í rannsóknum á íslensku nútímamáli, en öflugar rannsóknir eru mikilvæg undirstaða varðveislu og eflingar íslenskrar tungu. Sem dæmi um rannsóknir sem eru í undirbúningi má nefna tíðnirannsóknir á íslenskum orðtökum og rannsóknir á nýyrðum, orðmyndunarmöguleikum og úreltum orðum í íslensku.

2.3.1 Almennir notendur

Notendahópur þýska orðasjóðsins er mjög fjölbreyttur. Allir sem eitthvað hafa með málnotkun (skrifa og þýða texta), þýskukennslu eða tungumálarannsóknir að gera virðast nota hann og hafa gagn af honum, bæði notendur með þýsku sem móðurmál og einnig þeir sem eru að læra þýsku. Miðað við notendahóp þýska orðasjóðsins má ætla að Íslenskur orðasjóður muni einnig nýtast mjög fjölbreyttum íslenskum notendahópi og jafnvel einnig málnotendum sem eru að læra íslensku sem erlent tungumál.

2.3.2 Sérhæfðir notendur

Það er viðurkennd staðreynd að textagrunnar eru nauðsynleg hjálpartæki í tungumálarannsóknum (sbr. Quasthoff, Richter og Biemann 2006). Textasafn Orðabókar Háskólans inniheldur samtals um 52 milljónir lesmálsorða úr fjölbreyttum textum (sbr. upplýsingar á heimasíðu Orðabókar Háskólans, http://www.lexis.hi.is/ts_umsafnid.htm) en vegna höfundarréttar er vefaðgangur[2] að safninu einskorðaður við texta án

[2]Starfsmenn Orðabókarinnar hafa reynst mjög hjálpsamir við aðstoð við og aðstöðu fyrir rannsóknarverkefni, þ.e. aðgangur að öllu textasafninu hefur verið mögu-

höfundarréttar. Við Orðabók Háskólans er einnig verið að vinna að markaðri íslenskri málheild (sbr. Sigrún Helgadóttir 2004).

Markmiðið með Íslenskum orðasjóði er að veita aðgang að málnotkun í íslensku eins og hún er í dag í textum (hugsanlegur möguleiki er að leyfa val á milli textategunda eftir uppruna textanna, t.d. úr *Morgunblaðinu*, til að kynna sér málnotkun í þeim). Notkunargildi Íslensks orðasjóðs felst einkum í orðfræðilegum upplýsingum um notkun orða í textum; þessar upplýsingar eru skýrðar nánar í kafla 3. Gagnagrunnurinn sem geymir textana er þannig byggður upp að ekki er hægt að endurgera texta úr honum; þetta er nauðsynleg ráðstöfun til að tryggja að farið sé eftir lögum um höfundarrétt.

Íslenski textagrunnurinn verður hluti af fjölmála textagrunni á vefsvæði þýska orðasjóðsins sem ætlaður er til notkunar í tungumálarannsóknum. Hugsanlegt er að nota þessa textagrunna meðal annars við (sbr. Quasthoff, Richter og Biemann 2006):

- vinnu að einmála orðabókum,
- leit að svörum við málfræðilegum spurningum,
- tölfræðilega unninn samanburð á mismunandi tungumálum,
- gerð mállíkana, t.d. fyrir talgreiningu,
- rannsóknir á orðum sem haga sér tölfræðilega á líkan hátt,
- val á orðum í tilraunir, t.d. í sálfræðilegum málvísindum.

Þetta er ekki tæmandi listi, möguleikarnir eru margvíslegir, m.a. við rannsóknir á tíðni, orðmyndun, merkingu og merkingarlegu umhverfi orða. Dæmi um önnur áhugaverð rannsókna- og tungutækniverkefni sem byggja á gögnum úr textagrunnum má finna í greinum Richter, Quasthoff, Erla Hallsteinsdóttir og Biemann (2006) og Quasthoff, Richter og Biemann (2006) um notkun textagrunna í tungumálarannsóknum.

Eins og áður var nefnt hefur þýski orðasjóðurinn verið notaður sem grunnur í rannsókn á tíðni þýskra orðtaka. Niðurstöðurnar úr þeirri rannsókn hafa þegar verið nýttar á margvíslegan hátt, m.a. við að velja orðtök í þýsk-íslenskan orðtakagagnagrunn (sbr. Erla Hallsteinsdóttir 2005, 2006b), við að velja þýsk orðtök í grunnorðaforða þýsku sem erlends tungumáls (sbr. Erla Hallsteinsdóttir, Sajankova

legur ef unnt er að vinna rannsóknarvinnuna í húsakynnum Orðabókarinnar.

og Quasthoff 2006) sem og við endurskoðun fræðilegra hugmynda um margræðni orðtaka og þróun aðferðafræði við rannsóknir á orðtökum (sbr. Erla Hallsteinsdóttir í prentun). Íslenski textagrunnurinn mun verða notaður við tíðnirannsóknir á íslenskum orðtökum og niðurstöður þeirra rannsókna munu verða nýttar á sama hátt og niðurstöður þýsku rannsóknarinnar. Þrátt fyrir miklar og góðar orðtakarannsóknir í íslensku eru rannsóknir á orðtökum í textagrunni, t.d. á tíðni orðtaka, á frumstigi og það vantar enn skilgreiningu á þeim orðtökum sem tilheyra grunnorðaforða, þ.e. hvaða orðtök eru æskilegur hluti orðaforða við nám í íslensku sem erlendu máli.

3 Íslenskur orðasjóður

Íslenski *Morgunblaðs*-orðasjóðurinn eins og hann er vistaður á vefsíðu Projekt Deutscher Wortschatz í Leipzig í dag býður m.a. upp á eftirfarandi notkunarmöguleika:

(1) leit að orðum/orðmyndum,
(2) leit að notkunarumhverfi,
(3) leit að notkunardæmum,
(4) leit að samsettum og afleiddum orðum með algildistáknum (* og ?),
(5) listar með tíðni orðmynda.

Hér á eftir fylgir stutt lýsing Íslenska orðasjóðnum með dæmum. Að auki mun ég lýsa þeim möguleikum sem eingöngu eru í þýska orðasjóðnum og hægt væri að yfirfæra á orðasafnið í Íslenskum orðasjóði.

3.1 Leit að orðum

Leitarniðurstöður í Íslenskum orðasjóði sýna tíðni, tíðniflokk (miðað við „og“ sem er algengasta orð) og textadæmi með tengli til fleiri dæma, sbr. niðurstöður leitar að orðinu *orðabók* í mynd 2:

Mynd 2: Upplýsingar um tíðni og tengill við fleiri dæmi.

Þar sem ekki er búið að tengja saman beygingarmyndir orða eru eingöngu sýndar upplýsingar um orðið í eintölu í nefnifalli, þolfalli og þágufalli, þ.e. upplýsingar um beygingarmyndina *orðabók*. Leita verður sérstaklega að öðrum myndum orðsins. Einnig er sýnt merkingarlegt umhverfi leitarorðsins, þ.e. gefin eru upp orð sem hafa marktæka tíðni sem nágrannar leitarorðsins í textum. Þessar upplýsingar verða útskýrðar nánar í eftirfarandi köflum.

3.2 Notkunarumhverfi

Með notkunarumhverfi er ekki eingöngu átt við fastar orðastæður (e. collocations) í hefðbundnum skilningi heldur einnig þau orð sem hafa háa tíðni sem nágrannar leitarorðsins í textum. Við þau orð sem hafa marktæka tíðni sem nágrannar er sýnd heildartíðni í gagnagrunninum. Einnig er greint á milli orða sem koma fyrir sem hægri og vinstri nágrannar leitarorðsins og við þau orð er gefin upp sú tíðni sem þau hafa með leitarorðinu, sbr. mynd 3:

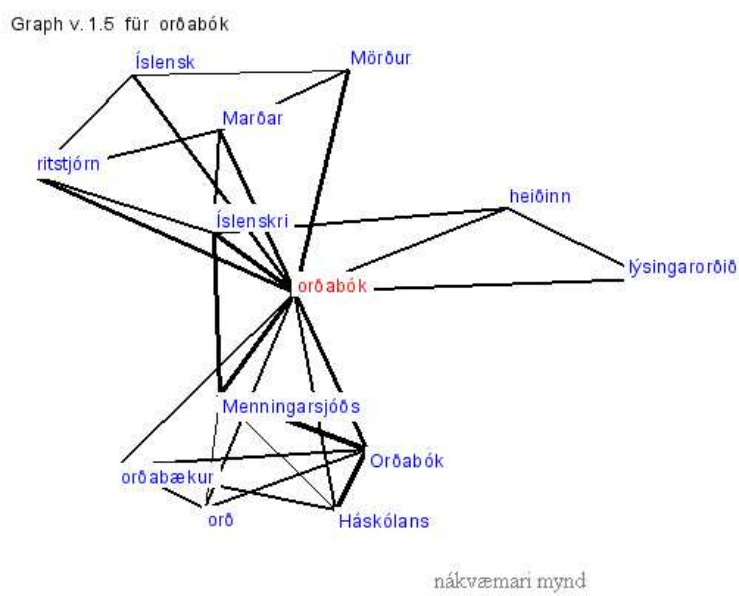
orð sem koma oft fyrir sem nágrannar orðabók:
Menningarsjóðs (85), Íslenskri (64), Íslensk (42), Orðabók (23), Mörður (21), ritstjórn (20), Marðar (20), Í (17), Orðastað (17), lýsingarorðið (15), heiðinn (15), orðabækur (14), orð (14), Háskólans (14), þreyja (13), Árnasonar (13), stórfiskaleikur (13), prentútgáfna (13), lýðveldistímans (13), klyfberi (13), Bókaútgáfu (13), ÍSLENSK (12), uppgjöf (12), delicious (12), útgáfudegi (11), merking (11), lsquo (11), gefast (11), eða (11), Freysteins (11), orðsins (10), orðið (10), orðinu (10), merkir (10), hugum (10), færeyskt (10), fletta (10), ekki (10), dægrastytting (10), Grunnavík (10), Örlygs (9), Íslenska (9), syndrome (9), glöggva (9), forsölu (9), bók (9), Árna (8), viðhorfa (8), slangur (8), samkvæmt (8), samanlögðu (8), ríkjandi (8), nýrri (8), metsölubók (8), heimspekideild (8), fornt (8), endurbætt (8), Blöndals (8), nefni (7), merkingar (7), keyptum (7), er (7), alist (7), Starfaði (7), Orðið (7), Böðvarssonar (7), íslenskri (6), Ö (6), skýringum (6), selst (6), samantekt (6), lektor (6), hinni (6), hin (6), gefin (6), eintök (6), dósent (6), Færeyingar (6), íslensku (5), íslenska (5)

orð með háa tíðni sem vinstri nágrannar orðabók:
Íslenskri (64), Íslensk (34), ÍSLENSK (12), íslenskri (4), Úr (4), Íslenska (4), íslenska (3), samkvæmt (3)

orð með háa tíðni sem hægri nágrannar orðabók:
Menningarsjóðs (50), Freysteins (11), Blöndals (8), ríkjandi (5), Háskólans (5)

Mynd 3: Dæmi um notkunarumhverfi orða: nágrannar í textum.

Upplýsingar um merkingarumhverfi orða eru sýndar á grafísku formi með n.k. merkingarneti, sbr. mynd 4:



Mynd 4: Merkingarnet orðmyndarinnar *orðabók*.

Upplýsingar um orðastæður og textanágranna eru mikilvægar í málnotkun. Þær sýna hvaða orð er hægt að nota saman (sbr. muninn á að

bursta tennurnar í íslensku og „hreinsa tennurnar“ (*die Zähne putzen*) í þýsku). Þessar upplýsingar mynda grunn fyrir tungumálarannsóknir, t.d. merkingarfræði og setningafræði og einnig við orðabókagerð.

3.3 Notkunardæmi

Við hvert orð eru sýndar tvær setningar úr gagnagrunninum sem sýna notkun orðsins. Að auki er tengill við síðu með fleiri notkunardæmum, sbr. eftirfarandi notkunardæmi fyrir leitarorðið *orðabók* (öll dæmi úr *Morgunblaðinu*):

Dæmi:

- Í samantekt Minjasafnsins á Akureyri sem byggir m.a. á bók Hallgerðar Gísladóttur, Íslensk matarhefð, kemur fram að elstu rituðu heimildir um hátíðarbrauð Íslendinga, laufabrauðið, séu í **orðabók** Jóns Ólafssonar frá Grunnavík frá árinu 1736.
- Það er greinilegt að blaðamaður hefur ekki ómakað sig við að fletta upp í nýútkominni **orðabók** til að glöggva sig á málinu, því það er ljóst að hefði hann gert það, hefðu hinar gnarrísku fullyrðingar aldrei komist á prent.
- Í nýrri og prýðilegri **orðabók** Eddu - miðlunar er merking orðsins „skipuleg samtök til að berjast fyrir ákveðinni stefnu og markmiðum í stjórnmálum“.
- Vol og væl var ekki til í hennar **orðabók** og ekki minnist ég þess að hafa nokkurn tíma hitt hana í slæmu skapi.
- Sjálfsvorkunn var hreinlega ekki til í hans **orðabók**.
- Í Íslenskri **orðabók** stendur um lýsingarorðið heiðinn: 1) sem er heiðingi, ókristinn; guðlaus; heiðinn siður Ásatrú; heiðinna manna heilsa fornmannaheilsa, góð heilsa. 2) ófermdur, illa upplýstur um trúmál. 3) sem vantar á: heiðinn klyfberi gjarðalaus klyfberi; verlaus (um sæng): sofa í heiðnu rúmi; bryddingalaus; auður, óskrifaður: heiðin blaðsíða; sviplaus, eyðilegur: þetta er svo heiðið.
- Ný íslensk **orðabók** var kærkomin sending inn í það ógnargímald, en betur má ef duga skal.
- Því er vitnað í þessa bók nú, að Íslensk **orðabók** hefur nú verið gefin út, mikið aukin og endurbætt.

- En eigum við að fljóta sofandi að feigðarósi og bíða þess að lýsingarorðið „delicious“ rati inn í íslenska **orðabók** og sé þar sjálfsögð fletta og eðlilegur hluti talaðs máls á Íslandi?
- Enska lýsingarorðið „delicious“ er hinsvegar ekki að finna í hinni nýju íslensku **orðabók**, þótt það hafi ratað inn í auglýsingu kartöflubænda.
- Margar slettur hafa skamma viðdvöl á vörum fólks, – sem betur fer og raunar þurfa kannski ýmsir á því að halda að slettunum sé hægt að fletta upp í **orðabók**.
- HIN nýja Íslenzka **orðabók** hefur komið af stað miklum umræðum, ekki sízt um það hvort ýmis orð (t.d. sjitt eins og frægt er orðið) megi vera í bókinni eða hvort úthýsa beri „röngu“ máli úr jafnvirðulegri heimild um íslenzkt mál.
- Þau tímamót sem mörkuð eru með nýrri **orðabók** eru hvatning til þess að almenningur taki afstöðu til þess málfars sem nú tíðkast og leiði jafnframt hugann að því það hvernig hann telur ákjósanlegt að málið þróist
- Þegar orðabókinni var fylgt úr hlaði í síðustu viku kom það fram í máli Marðar Árnasonar að vonir stæðu til „að hér eftir liðu ekki nema 5–10 ár á milli prentútgáfna af Íslenskri **orðabók**“.

Greining á þessum dæmum leiðir m.a. í ljós að orðið *orðabók* er oft hluti af sérheitinu *Íslensk orðabók*, það kemur fyrir sem hluti af orðasambandinu „eitthvað er ekki til í hans/hennar orðabók“ og í merkingunni ‘uppsláttarrit um orðaforða’.

3.4 Leit að samsettum og afleiddum orðum

Með því að nota algildistáknin stjörnu eða spurningarmerki (* eða ?) er hægt að leita að orðum sem innihalda ákveðna bókstafi eða orðhluta. Leitarorðið *orðab** gefur t.d. eftirfarandi niðurstöður (tíðni í hornklofa):

orðabanka [4]	orðabókarforrit [1]	orðabókarsmíðinni [1]
orðabankann [1]	orðabókargerð [2]	orðabókarstjóri [3]
orðabankanum [5]	orðabókargerðina [2]	orðabókarstörf [1]
orðabankinn [1]	orðabókargerðinni [1]	orðabókarverk [1]
orðabilum [8]	orðabókarhefð [1]	orðabókarverkefnið [1]
orðablaðra [1]	orðabókarhöfunda [3]	orðabókarvinnslunni [1]
orðablaðran [1]	orðabókarhöfundar [1]	orðabókasmíð [1]
orðabrunnurinn [1]	orðabókarhöfundarnir [1]	orðabókaverkefni [1]
orðabækur [29]	orðabókarhöfundi [1]	orðabókaútgáfa [1]
orðabækurnar [2]	orðabókarinnar [16]	orðabókaútgáfu [2]
orðabók [120]	orðabókarlýsingu [2]	orðabókin [34]
orðabóka [9]	orðabókarmaður [2]	orðabókina [11]
orðabókagerð [2]	orðabókarnotenda [1]	orðabókinni [27]
orðabókagerðina [1]	orðabókarritstjóri [2]	orðabókum [20]
orðabókanotenda [1]	orðabókarskráin [1]	orðabókunum [1]
orðabókar [21]	orðabókarsmiðsins [1]	orðabólgu [1]
orðabókarbreytingar [1]	orðabókarsmíð [1]	

Tafla 1: Orð sem innihalda *orðab.*

Eins og sést í töflu 1 koma öll beygingarform fram sem sérstök orð, án tengsla sín á milli, sbr.:

orðabækur [29], orðabækurnar [2], orðabók [120], orðabóka [9], orðabókar [21], orðabókarinnar [16], orðabókin [34], orðabókina [11], orðabókinni [27], orðabókum [20]. Til þess að öll beygingarform birtist í leitarniðurstöðum verður að samtengja beygingarform og grunnform orða eins og gert hefur verið í þýska orðasjóðnum.

3.5 Orðalistar

3.5.1 Listar með tíðni orðmynda

Upplýsingar um tíðni orða og orðasambanda eru mikilvægar, bæði í hagnýtum tungumálarannsóknum og í málvísindum. Á einfaldan hátt er hægt að gera lista með t.d. algengustu eða sjaldgæfustu orðum í gagnagrunninum. Í töflu 2 eru sýnd dæmi fyrir þýsku og ensku:

þýsk orð		**ensk orð**	
1:	*der*	*of*	:1
2:	*die*	*to*	:2
3:	*und*	*and*	:3
4:	*in*	*a*	:4
5:	*den*	*in*	:5
6:	*von*	*for*	:6
7:	*zu*	*is*	:7
8:	*das*	*the*	:8
9:	*mit*	*that*	:9
10:	*sich*	*on*	:10
	Niðurhal algengustu orða	*Niðurhal algengustu orða*	
	++ 100 ++ 1000 ++ 10000 ++	*++ 100 ++ 1000 ++ 10000 ++*	

Tafla 2: Algengustu orð í þýsku og ensku.

3.5.2 Nýyrði og úrelt orð

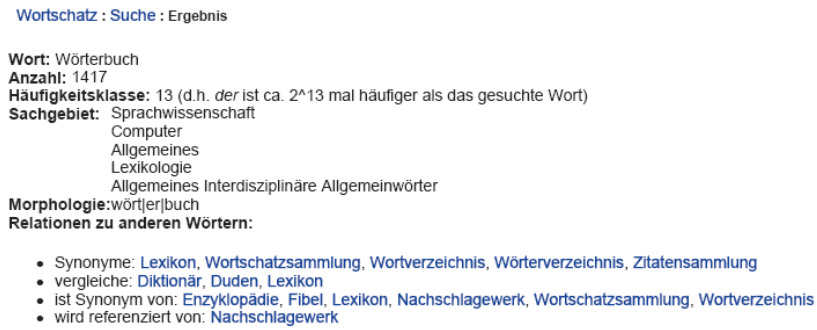
Með því að bera saman orð og beygingarmyndir úr *Beygingarlýsingu íslensks nútímamáls* og orð og beygingarmyndir sem koma fyrir í textagrunninum fást áhugaverðar upplýsingar um nýyrði, orðmyndunarmöguleika og úrelt orð í íslensku. Gera má ráð fyrir að þau orð sem ekki eru í beygingarlýsingunni séu nýyrði eða slangur og að þau orð sem ekki koma fyrir í textagrunninum séu orðin úrelt eða séu beygingarform sem af einhverjum ástæðum eru ekki notuð. Gögn úr þannig samanburði mynda grunn fyrir rannsóknir á þróun og stöðu orðaforðans.

3.6 Deutscher Wortschatz

Eins og eftirfarandi dæmi sýna hefur verið bætt við töluverðu magni af upplýsingum í þýska orðasjóðnum, m.a. um beygingu, merkingu og merkingartengsl (þó ekki merkingarlýsingu), orðmyndun, orðatengsl o.fl. Þessar upplýsingar, sem byggja að hluta til á lokaverkefnum nemenda við Háskólann í Leipzig, eru unnar bæði sjálfvirkt, hálfsjálfvirkt og handvirkt. Notuð er sjálfvirk tenglasetning til að samtengja upplýsingarnar í orðasjóðnum.

Í mynd 5 hafa auk tíðni (Anzahl) og tíðniflokks (Häufigkeitsklasse) verið tilgreindir fagflokkar sem orðið *Wörterbuch* tilheyrir (Sachge-

biet); sýnd er orðhlutagreining orðsins (Morphologie) og gefin eru upp
merkingarleg tengsl (samheiti, svipuð merking o.s.frv.) við önnur orð (Relationen zu anderen Wörtern):


Wortschatz : Suche : Ergebnis

Wort: Wörterbuch
Anzahl: 1417
Häufigkeitsklasse: 13 (d.h. *der* ist ca. 2^13 mal häufiger als das gesuchte Wort)
Sachgebiet: Sprachwissenschaft
Computer
Allgemeines
Lexikologie
Allgemeines Interdisziplinäre Allgemeinwörter
Morphologie:wört|er|buch
Relationen zu anderen Wörtern:

- Synonyme: Lexikon, Wortschatzsammlung, Wortverzeichnis, Wörterverzeichnis, Zitatensammlung
- vergleiche: Diktionär, Duden, Lexikon
- ist Synonym von: Enzyklopädie, Fibel, Lexikon, Nachschlagewerk, Wortschatzsammlung, Wortverzeichnis
- wird referenziert von: Nachschlagewerk

Mynd 5: Upplýsingar við orðið *Wörterbuch* í þýska orðasjóðnum.

Í mynd 6 (Links zu anderen Wörtern) eru sýnd dæmi um fleiri tegundir merkingartengsla og einnig eru gefin upp samsett orð, orðastæður og orðasambönd sem orðið er hluti af og sem hægt er að slá upp sem sjálfstæðum flettum í orðasjóðnum. Að auki koma fram beygingarform og skammstafanir orðsins.


Links zu anderen Wörtern:

- falls positiv bewertet Originalwörterbuch
- Grundform: Wörterbuch
- ist ein(e) Buch, Nachschlagewerk, Wortsammlung
- Teilwort von: im Wörterbuch nachschlagen, Wörterbuch zusammenstellen, kurzgefaßtes Wörterbuch, fremdsprachliches Wörterbuch, ein einsprachiges Wörterbuch, rückläufig sortiertes Wörterbuch, automatisches Wörterbuch, ein gutes Wörterbuch, rückläufiges Wörterbuch, computerisiertes Wörterbuch, in einem Wörterbuch nachschlagen, ein wandelndes Wörterbuch, übersetzende Wörterbuch, ein Wörterbuch kürzen
- Form(en): Wörterbuch, Wörterbücher, Wörterbüchern, Wörterbuchs, Wörterbuches, Wörterbuche
- Abkürzung: WB, Wtb., Wb.

Dornseiff-Bedeutungsgruppen:

- 11.30 Kenntnis: Bibliografie, Buch, Enzyklopädie, Handbuch, Lexikon, Pflichtlektüre, Sekundärliteratur, Standardwerk, Vademekum, Vokabular, Wörterbuch
- 12.16 Bezeichnung, Wort: Duden, Fremdwörterbuch, Glossar, Grundwortschatz, Lexikon, Phrasenkatalog, Sprachschatz, Verzeichnis, Wortschatz, Wörterbuch, Zitatenschatz
- 12.43 Erklärung: Enzyklopädie, Lexikon, Wörterbuch
- 12.55 Schriftliche Überlieferung: Auflistung, Bibliografie, Enzyklopädie, Index, Katalog, Kompendium, Konkordanz, Lexikon, Liste, Register, Tabelle, Verzeichnis, Werkverzeichnis, Wörterbuch

Mynd 6: Upplýsingar við orðið *Wörterbuch* í þýska orðasjóðnum.

Sýnd eru hugtök sem orðið heyrir undir í *Dornseiff*-hugtakaorðabókinni (Dornseiff-Bedeutungsgruppen, sbr. Dornseiff 2004) sem er ein stærsta hugtakaorðabók í þýsku, en Uwe Quasthoff ritstýrði nýjustu útgáfu hennar.

4 Íslensk-þýsk orðabók

Saga þýsk-íslenskra orðabóka hefur verið hálfgerð sorgarsaga (sbr. Erla Hallsteinsdóttir 2004). Fyrir utan skólaorðabók Steinars Matthíassonar (Steinar Matthíasson 2004) hafa ekki verið gefnar út neinar orðabækur með þýsku fyrir íslenska notendur á síðustu árum. Ein af hugmyndum um notkun á Íslenskum orðasjóði í Leipzig er að byrja á grunni fyrir þýsk-íslenska netorðabók á svipuðu formi og þýsk-enska orðabókin sem þegar hefur verið gerð í Leipzig, sjá eftirfarandi dæmi í mynd 7:

Abfrageergebnis für Ihre Anfrage nach »Wörterbuch«

18 Treffer aus dem **deutsch->englisch** Lexikon:		#Belege
Wörterbuch (652)	dictionary (4641)	7
Wörterbuch (652)	vocabulary (1007)	2
Wörterbuch (652)	wordbook	1
Wörterbuch (652)	glossary (689)	1
ein gutes (8783) Wörterbuch (652)	a good dictionary (4641)	1
ein Wörterbuch (652) kürzen (8322)	abridge (18) a dictionary (4641)	1
übersetzende (9) Wörterbuch (652)	translating (1086) dictionary (4641)	1
kurzgefaßtes (2) Wörterbuch (652)	concise (839) dictionary (4641)	1
rückläufiges (38) Wörterbuch (652)	reverse (6828) dictionary (4641)	1
automatisches (152) Wörterbuch (652)	automatic (16704) dictionary (4641)	3
ein wandelndes (59) Wörterbuch (652)	a walking (4486) dictionary (4641)	1
Wörterbuch (652) zusammenstellen (592)	compile (2688) a dictionary (4641)	1
im Wörterbuch (652) nachschlagen (118)	consult (1853) the dictionary (4641)	1
computerisiertes (5) Wörterbuch (652)	computerized (4588) dictionary (4641)	1
ein einsprachiges Wörterbuch (652)	a monolingual (7) dictionary (4641)	1
fremdsprachliches (2) Wörterbuch (652)	foreign-language (218) dictionary (4641)	1
rückläufig (1711) sortiertes (24) Wörterbuch (652)	backward-sorted dictionary (4641)	1
in einem Wörterbuch (652) nachschlagen (118)	consult (1853) a dictionary (4641)	1

Mynd 7: *Wörterbuch* í þýsk-ensku orðabókinni.

Í því sambandi munu möguleikar á sjálfvirkri málgreiningu verða skoðaðir (sbr. t.d. Cysouw, Biemann og Ongyerth 2006 og Rapp 1994) með það í huga að nýta þá möguleika sem til eru á notkun textagrunnsins í sjálfvirkum þýðingum eða orðabókagerð, t.d. með tölfræðilegri greiningu orða, orðastæðna og setninga í textum eða með samanburði á frumtextum og þýðingum þeirra.

5 Samantekt

Íslenskur orðasjóður er verkefni sem unnið er við Háskólann í Leipzig. Orðasjóðurinn, sem byggir á vefsíðusafni Landsbókasafns Íslands – Háskólabókasafns og nýtir tækniþekkingu úr verkefninu Deutscher

Wortschatz, veitir almennum notendum aðgang að upplýsingum um málnotkun í íslensku nútímamáli og fræðimönnum textagrunn sem hægt er að nýta á margvíslegan hátt í rannsóknum á tungumálum. Orðasjóðurinn mun nýtast bæði í hagnýtum rannsóknum eins og orðabókagerð, gerð kennsluefnis og tungumálakennslu og í fræðilegum rannsóknum, t.d. við þróun kenninga og aðferðafræði í tungumálarannsóknum.

Ritaskrá

Beygingarlýsing íslensks nútímamáls. Vefslóð: http://www.lexis.hi.is/beygingarlysing.

Cysouw, Michael, Biemann, Christian og Ongyerth, Matthias. 2006. Using strong's numbers in the bible to test an automatic alignment of parallel texts. Í: Michael Cysouw og Bernhard Wälchli (útg.): Parallel Texts: Using translational equivalents in linguistic typology. *Special issue of Sprachtypologie und Universalienforschung (STUF)*, bls. 66–79.

Dornseiff, Franz. 2004. *Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearb. und mit einem alphabetischen Zugriffsreg. vers. Aufl. von Uwe Quasthoff. Berlin, New York: de Gruyter.

Erla Hallsteinsdóttir. 2004. En kort oversigt over islandsk ↔ tysk leksikografi. Í: *LexicoNordica* (11), bls. 51–65.

Erla Hallsteinsdóttir. 2005. Vom Wörterbuch zum Text zum Lexikon. Í: Ulla Fix, Gotthard Lerchner, Marianne Schröder og Hans Wellmann (útg.): *Zwischen Lexikon und Text – lexikalische, stilistische und textlinguistische Aspekte*, bls. 325–337. Leipzig: Verlag der Sächsischen Akademie der Wissenschaften zu Leipzig.

Erla Hallsteinsdóttir. 2006a. Phraseographie. Í: *HERMES Journal of Language and Communication Studies* (36), bls. 91–128.

Erla Hallsteinsdóttir. 2006b. Konzeption und Erstellung einer computergestützten zweisprachigen Phraseologiesammlung Isländisch – Deutsch. Í: Annelies Häcki Buhofer og Harald Burger (útg.): *Phraseology in Motion*. Proceedings zu Europhras Basel 2004, bls. 211–222. Baltmannsweiler: Schneider Verlag.

Erla Hallsteinsdóttir. Í prentun. Wörtliche, freie und phraseologische Bedeutung. Eine korpusbasierte Untersuchung des Vorkommens von freien und phraseologischen Lesarten bei deutschen Idiomen. Í: Erika KrûiŽnik (útg.): *Phraseologie in der Sprachwissenschaft und anderen Disziplinen*. Akten der Europhras-Tagung in Strunjan/Slowenien, 19.–22. September 2005.

Erla Hallsteinsdóttir, Uwe Quasthoff og Monika Sajankova. 2006. Vorschlag eines phraseologischen Optimums für Deutsch als Fremdsprache auf der Basis von Frequenzuntersuchungen und Geläufigkeitsbestimmungen. Í: *Linguistik online* 27, 2/06: Neue theoretische und methodische Ansätze in der Phraseologieforschung, bls. 117–136. (Vefslóð: *www.linguistik-online.de/27_06/hallsteinsdottir_et_al.pdf*.)

Quasthoff, Uwe og Christian Biemann. 2006. Measuring Monolinguality. Í: *Proceedings of the LREC-06 workshop on Quality assurance and quality measurement for language and speech resources*, Genoa, Italy.

Quasthoff, Uwe og Matthias Richter. 2005. Projekt Deutscher Wortschatz. Í: *Babylonia* 3/2005.
Vefslóð: *http://www.babylonia-ti.ch/BABY305/quaride.htm*.

Quasthoff, Uwe, Matthias Richter og Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. Í: *Proceedings of LREC-06*, Genoa, Italy.

Rapp, Reinhard. 1994. Die maschinelle Generierung von Wörterbüchern aus zweisprachigen Texten. Í: Susanne Beckmann og Sabine Frilling (útg.): *Satz – Text – Diskurs*. Akten des 27. Linguistischen Kolloquiums, Münster, 1993, Bd. I, bls. 203–209. Tübingen: Niemeyer.

Richter, Matthias, Uwe Quasthoff, Erla Hallsteinsdóttir og Christian Biemann. 2006. Exploiting the Leipzig Corpora Collection. Í: *Proceedings of IS-LTC'06*, Ljubljana, Slovenia.

Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild. Í: *Tunga og tækni*, bls. 67–71. (Vefslóð: *www.tungutaekni.is/news/sigrun2.pdf*.)

Steinar Matthíasson. 2004. *Þýsk-íslensk, íslensk-þýsk orðabók*. Aukin og endurbætt útg. Reykjavík: Iðnú.

Abstract

Corpora are important linguistic resources. In this paper I describe the details of an Icelandic corpus, that is a part of a collection of corpora in 17 different languages which can be accessed online from *http://corpora.informatik.uni-leipzig.de*, and, with Icelandic instructions, at *http://wortschatz.uni-leipzig.de/ws_ice/index.php*. I discuss the possible usage of the corpus as a corpus based dictionary for non-linguistic users, and as a research tool for linguistic purposes in both applied and theoretical linguistics.

Keywords:

corpora, dictionary, linguistics, Icelandic

Lykilorð:

textagrunnur, orðabók, málvísindi, íslenska

Erla Hallsteinsdóttir
Vægtens Kvarter 336
DK-5220 Odense
erlahall@yahoo.dk

Veturliði G. Óskarsson

Um þýska forskeytið *an-* og stutta viðdvöl þess í íslensku

1 Inngangur

Norræn mál hafa öldum saman þegið orð og áhrif úr þýsku.[1] Veldi Hansakaupmanna hófst á 12. öld og um fjögurra alda skeið réðu þeir miklum hluta verslunar og viðskipta í Norður-Evrópu. Þeir báru með sér fjölmargar nýjungar, vörur, siði og menningaráhrif, og áhrif tungu þeirra, miðlágþýsku, á tungumál þeirra þjóða sem þeir umgengust voru í mörgum tilfellum afar mikil, einkum á orðaforðann. Löngum bárust einnig orð af latneskum eða öðrum uppruna þessa leið inn í norræn mál. Miðlágþýsk áhrif á norræn mál stóðu allt fram á 16. öld og voru skiljanlega mest á meginlandsmálin norsku, dönsku og sænsku, enda komu kaupmennirnir þýsku sér vel fyrir í þessum löndum og réðu þar svo að segja allri millilandaverslun um nokkurra alda skeið. Á 16. öld fór smám saman að draga úr áhrifum miðlágþýskunnar með dvínandi völdum Hansakaupmannanna og við tóku háþýsk áhrif.[2]

[1]Ég þakka Orðabók Háskólans fyrir afnot af gögnum og skrám og starfsfólki fyrir ýmsa aðstoð.

[2]Um veldi Hansakaupmanna má lesa mjög víða; gott yfirlit er að finna hjá Dollinger (1981) og yfirlit um lágþýsku t.d. hjá Krogmann (1970). Um áhrif miðlágþýsku á norræn mál má t.d. lesa í ýmsum greinum hjá Jahr (2000); um háþýsk áhrif sem beint framhald hinna miðlágþýsku, sjá Braunmüller (2000).

Orð og tunga 9 (2007), 125–149.

Meðal orða sem tekin voru upp í norræn mál voru fjölmörg forskeytt og viðskeytt orð. Frumnorræna hafði haft yfir að ráða ýmsum aðskeytum til orðmyndunar en með tímanum hurfu mörg þeirra (Seip 1934:29–30; Haugen 1976:159, 221). Meðal annars misstu norræn mál forskeytin **be-* og **ga-* (leifar hins síðara eru varðveittar í fáeinum orðum eins og *glíkr* og *granni*). Þau héldust hins vegar betur í vesturgermönsku. Forskeyti voru því heldur færri í norrænu en í vesturgermönskum málum og virkni þeirra nokkuð takmörkuð (Seip s.st.; Haugen 1976:381). Hið sama á að nokkru leyti við um norræn viðskeyti. Smám saman urðu mörg aðskeytin, sem tökuorðunum fylgdu, virk í orðmyndun í viðtökumálunum og er víst að aðskeytafátækt norrænna mála auðveldaði hinum þýsku leið sína inn í viðtökumálin. Nokkur helstu miðlágþýsku forskeytin voru *an-*, *be-*, *bi-*/*bī-*, *vor-*, *over-*, *um-* og *unt-* sem í norsku, dönsku og sænsku urðu *an-*, *be-*, *bi-*, *for(e)-*/*för-*, *over-*/*över-*, *om-* og *und-*/*unn-*. Af viðskeytum (eða endingum sem fengu hlutverk viðskeyta í norrænum málum) má nefna *-achtich*, *-ent*, *-bār*, *-heit*/*-hēt*, *-heftich*, *-inne*/*-in*, *-isch*, *-līk*, *-schap*. Sömu eða sambærileg aðskeyti er að finna í háþýsku og orð af þessi tagi héldu áfram að berast inn í norræn mál eftir siðaskipti; er oft erfitt að greina á milli eldri og yngri áhrifanna.[3] Í sumum tilfellum féllu þessi aðskeyti að einhverju leyti saman við norræn aðskeyti sem fyrir voru (*-līk*, sbr. físl. *-leg*, *-lig*, *-lík*; *-schap*, sbr. físl. *-skap*; *vor-*, sbr. físl. *for-*).

Sum þeirra aðskeyta sem eru af þýskum uppruna eru enn virk í viðtökumálunum, svo sem (da., no., sæ.) *-aktig*/*-agtig*, *-bar*, *-hed*/*-het*, *om-*, *over-*/*över-*, en önnur eru ekki lengur virk eða virkni þeirra hefur minnkað mikið, t.d. *an-*, *be-*/*bi-*, *for(e)-*/*för-* og *und-*/*unn-*.

Mun færri þýskættuð tökuorð bárust inn í íslensku í tímans rás en í frændmálin á meginlandinu. Öfugt við frændmálin bárust orðin ekki beint inn í íslensku úr lágþýsku eða háþýsku heldur í gegnum norsku talsvert fram á 15. öld og síðan mestmegnis um dönsku.[4] Og öfugt við hin málin má segja að „miðlágþýsk“ áhrif á íslensku hafi staðið mun

[3]Guðrún Kvaran (2000:175) tekur nokkur dæmi um mun á miðlágþýskum og háþýskum forskeytum með hliðsjón af ritum Westergård-Nielsens (1946) og Jóns Helgasonar (1929).

[4]Þýskættuð tökuorð í íslenskum prentuðum ritum 16. aldar eru reyndar flest úr miðlágþýsku og sum þeirra hafa borist inn úr þýðingum án danskra millilliða; t.d. kann Oddur Gottskálksson að hafa nýtt sér lágþýska þýðingu á biblíu Lúthers, auk hinnar háþýsku útgáfu (og annarra verka), þegar hann sneri Nýja testamentinu á íslensku (Jón Helgason 1929:179–180; Westergård-Nielsen 1946:lxxiv, lxxvii).

lengur yfir því að gömul orð af miðlágþýskum rótum héldu áfram að berast inn um dönsku í íslensku allt fram á 20. öld.

Í því sem hér fer á eftir verður grennslast fyrir um það hvort, og þá í hversu ríkum mæli og á hvaða tímabili, orð með forskeytinu *an-* bárust inn í íslensku, og hugað að afdrifum þeirra.

2 Forskeytið *an-* og orð með því í íslensku

Forskeytið *an-* í vesturgermönskum málum er að uppruna sagnaforskeyti með margvíslegt hlutverk og hefur sömu orðsifjar og forsetningin *á* í íslensku. Danska, norska og sænska tóku upp fjölmörg orð með þessu forskeyti í aldanna rás. Það telst ekki vera meðal þeirra sem urðu sérlega virk og má í flestum tilfellum telja orð með því vera eiginleg tökuorð fremur en að þau séu mynduð í norrænum málum. Orðin bárust hægt og sígandi inn í dönsku allt frá lokum 14. aldar en fjölgar mjög eftir 1500 (Skautrup 1947:83, 234; 1953:352); í orðabók Kalkars yfir eldri dönsku (1300–1700) eru yfir 100 slík orð, flest úr ritum frá 16.–18. öld. Í orðabók Söderwalls yfir sænsku fram til um 1525 eru ein 15 *an*-orð og eru nokkur dæmi þar frá því fyrir 1500. Í skjali frá síðasta fjórðungi 15. aldar kemur fyrir orðið *anskötning* ‘umsjón, ábyrgð’, myndað af *an-* og sænska orðinu *sköta* (Moberg 1989:214–215).

Í íslenskum ritum fyrir 1500 virðast ekki koma fyrir orð með þessu forskeyti (ONP; Veturliði Óskarsson 2003) né í prentuðum ritum 16. aldar (Jón Helgason 1929, sbr. *http://www.lexis.hi.is/ordlyklar/ntodds/nto.htm;* Westergård-Nielsen 1946). Við upphaf 17. aldar taka orð af þessum toga að berast inn í íslensku og í söfnum Orðabókar Háskólans (OH) er að finna tæplega 60 slík orð eða orðmyndir. Nokkur orðin eru samsetningar og mismunandi orðstofnar eru alls tæplega 40. Sagnir eru um 20, nafnorð 33 og lýsingarorð eru fjögur. Skráð dæmi í Ritmálssafni OH (RM, í gagnagrunninum *http://lexis.hi.is*) eru alls um 250 og fáein að auki er að finna í nokkrum sérsöfnum OH. Í Talmálssafni OH eru dæmi um sex orð frá árunum 1958–1979, öll hin sömu og í RM:[5]

[5]Eitt orð er að auki í Talmálssafni en það er sögnin *anventera* sem kemur fyrir í stöku merktri Eyjólfi Stefánssyni frá Dröngum: „Repentera klerkur kann / kænn með vizkuhóti, / enn ég sjálfur aftur vann / anventera á móti“, og á einum öðrum seðli, með athugasemdinni „bera á móti, mótmæla e-u (sjaldg.)“. Óvíst er um uppruna þessa orðs en hugsanlegur er skyldleiki við lat. *vetō* ‘banna’ (nh. *vetere*) og

(1) *anleiðing* ‘íhugun, athugun’ („taka til anleiðingar“) 3, *anmæla* ‘ákæra’ 1, *anstalt* hk. 1, *anstaltir* kv.ft. ‘læti’ (um krakka) 6, ‘erfiðleikar’ 1, *antaka* ‘taka við, taka gilt’ 1, *antigna* ‘hrósa’ 2, ‘bölva, lasta’ 3

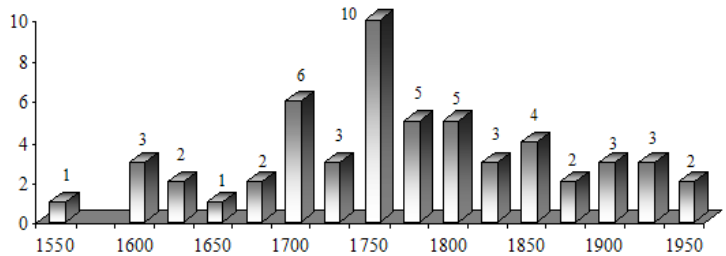
Leit í nokkrum ritum hefur aðeins leitt í ljós eitt orð sem ekki er í söfnum OH, dönsku orðmyndina *anseelse*, en fáein eldri dæmi hafa fundist svo og ein orðmynd sem ekki er hjá OH, *anmerkning* (um 1750) – RM hefur *anmerking*.

Einungis tvö orð af þessum toga er að finna í íslenskum nútímamálsorðabókum: *anstaltir* kv.ft. í Blöndalsorðabók (merkt með spurningarmerki sem vont mál) og *antigna* so. í sömu bók; hið síðara einnig í *Íslenskri orðabók* (1., 2. og 3. útg.) og *Íslenskri orðsifjabók* Ásgeirs Blöndals Magnússonar (1989). Þess má geta til fróðleiks að í nýlegum færeyskum orðabókum eru aðeins talin þrjú *an*-orð: *annám* ‘tak’, *annáma* ‘taka’ (*Föroysk orðabók* 1998, hið síðara merkt sjaldgæft) og *anfall* ‘kast’ (*Donsk-føroysk orðabók* 1995, skýring á da. *anfald*), en í færeysku talmáli eru nokkur í viðbót, a.m.k. þessi: *anbefala*, *anfektilsi*, *anføra*, *anganga*, *angrípa*, *anle(i)dningur*, *anmelda*, *anspora*, *ansøkja* (Simonsen 2002:80, 87).

Til eru í íslensku orð sem hefjast á *á-* og má rekja til baka til þýsks *an-*, t.d. *áklaga* so. (1540, RM) og *áklögun* kv. (1495, ONP) sem eiga vafalítið rætur að rekja til mlþ. *anklagen*, *anklage*, gegnum dönsku eða norsku. Í miðnorsku kemur fyrir orðmyndin *áklagan* (1471, DN 11:199) og í fornsænsku *aklaghan* (1439, Söderwall 1884–1918), lagaðar eftir miðlágþýskum hliðstæðum, en lágþýsk mynd nafnorðsins kemur fyrir í dönsku á fyrsta fjórðungi 16. aldar, *anclage* (1522, Kalkar 1881–1918), sem bendir til þess að um þær mundir hafi orðið verið tekið þar upp á ný, og þá í heild sinni. – Að þessu sinni verður ekki grennslast fyrir um orð af þessu tagi í íslensku.

Yfirlit um *an*-orð í íslensku eftir tímabilum má sjá á mynd 1 og í töflu 1.

að orðið sé myndað með hliðsjón af *repentera* < *repetere* (með *n*-innskoti). (Ég þakka Gunnlaugi Ingólfssyni fyrir þessa ábendingu.)



Mynd 1. Fjöldi nýrra orða með forskeytinu an- eftir tímabilum.

Elst er stakorð um miðja 16. öld (*annáma*) og eru síðan ekki dæmi um ný *an*-orð fyrr en snemma á 17. öld. Upp frá því tekur orðum að fjölga. OH hefur dæmi um átta orð(myndir) sem fyrst koma fyrir á þeirri öld en mest fjölgar orðum á 18. öld, að því er virðist, en frá þeirri öld eru dæmi um 24 ný orð. Hugsanlegt er að munur á fjölda nýrra *an*-orða á 17. öld og þeirri 18. í RM endurspegli að einhverju leyti mun á umfangi varðveittra heimilda frá þessum öldum. Um það verður þó ekki sagt neitt með vissu að sinni. Á 19. öld bætast við 14 orð(myndir) og átta koma fyrst fyrir á 20. öld, tvö yngstu orðin eru úr skáldsögum frá því um miðja öldina (Þórleifur Bjarnason: *Tröllið sagði*, 1958; Guðmundur G. Hagalín: *Konungurinn á Kálfskinni*, 1945).

Í töflu 1 eru tilfærð elstu kunn dæmi um hvert orð ásamt hliðstæðum í dönsku og þýsku, svo og notkunardæmi. Í flestum tilfellum eru hliðstæður teknar upp úr danskri orðsifjabók Nielsens (1989) en þegar hana þrýtur er leitað í aðrar orðabækur eftir þörfum, einkum ODS, Kalkar (1881–1918) og Schiller og Lübben (1875–1881). Látið er nægja að geta um miðlágþýskt dæmi, sé það kunnugt, en oftast er sambærilegt orð að finna í háþýsku. Ártöl eru nálganir og er nánari upplýsingar að finna í RM; örfá eldri dæmi hafa komið í ljós og er aldur þá skráður með hliðsjón af þeim. Heimildaskammstafanir við dæmi úr RM eru hinar sömu og þar eru notaðar. Örfá dæmi eru sótt í aðrar heimildir en í RM.

Tafla 1. Elstu dæmi um orð og orðmyndir með forskeytinu an- í íslensku

Um 1550:

- *annáma* (*annamma*) so. ‘taka við; tileinka sér’ (da. *annamme*; mlþ. *annāmen*), elsta dæmið er formúlan *taka*, *annáma og undfanga* DI XI, 183 (da. *tage*, *annamme og undfange*, mlþ. *annāmen*, *untvan / untvangen*)

Um 1600:

- *anslag* hk., *anslagur* kk. ‘ráðagerð, ráðabrugg’ (da. *anslag*; mlþ. *anslach*): „Þa er þad þo ecki hans [þ.e. djöfulsins] eiginlega rietta Anslag“ SummSp Mm, IIv
- *anmæli* hk. ‘umtal’, *leggja e-ð í anmæli*: „og því er það síðan í annmæli lagt: „að aungvir hafa gjafmildari verið enn Danir““ *Islandske Annaler*, 439; uppruni orðsins er ekki fullljós en tengsl við da. *anmelde* < þý. *anmelden*, m.a. í merkingunni ‘gera kunnugt; umtala’, eru sennileg (fremur en við *anmálsemi*, sjá nmgr. 8)

Um 1625:

- *ansjá* so. 1) ‘telja, álíta’, 2) ‘refsa’ (da. *anse*; mlþ. *ansēn*): 1) „að hans kongleg May^{tt} hafi fyrir gott anséð“ Tyrk, 448 (1638); 2) „vonar það hann (Jón) muni ansjást öðrum til viðvörunar“ Alþb X, 123 (1713)
- *antigna* so. 1) ‘biðja bölbæna, lasta’, 2) ‘lofa ákaflega’ (da. *antegne* ‘geta um e-ð e-m til lofs eða lasts’; mlþ. *antek(en)en*): 1) „vita ecke huørsu jlla þeir skule antigna Christo“ HammKrossg O, VIIr (1618); 2) „hvern Hann […] elskade, og framar flestum antignade“ SvSJJ, 6 (1769)

Um 1650:

- *anhang* hk. ‘hjálparmaður, fylgdarlið’ (í neikv. merk.) (da. *anhang*; mlþ. *anhang*): „Mun ekki Halldóra […] mest hafa haft fyrir því verki með hennar anhangi?“ JMPísl, 160

Um 1675:

- *angefa* so. ‘gefa upp; segja frá’ (da. *angive*; mlþ. *angeven*): „hvert góz í Snæfellssýslu sé angefið og afbetalað“ Alþb VIII, 158 (1687)
- *antasta* so. ‘handtaka’ (da. *antaste*; mlþ. *antasten*): „er óskað, að sagðar persónur séu antastaðar og undir frekara rannsak fluttar til Bessastaða“ Alþb VII, 212 (1671)

Um 1700:

- *angá/anganga* so. ‘varða’ (da. *angå*; mlþ. *angān*): „so vítt som þennann dómsins post angeingur“ ÁMPriv, 555
- *angefning* ‘það að gefa upp; kæra’ (da. *angivning* < *angive*; mlþ. *angeven*): „að veleðla herra amtmaðurinn vildi ei framvegis hérnefnds Tómasar angefningar upp á þeirra dóma meðtaka“ Alþb VIII, 258 (1689)

- *angefningarpóstur* kk. ‘ákæruatriði’: „tóku biskupinn […] og landfógetinn […] próf og rannsak um angefningarpósta þá“ Ann II, 269
- *anpartur* kk. ‘hlutur (í e-u)’ (da. *anpart*; mlþ. *anpart*): „jeg biode honum at leisa inn þeirra anpart“ ÁMTorf, 105
- *anleiðing* kv. ‘tilefni, ástæða’ (da. *anled(n)ing* < *anlede* († ODS); þý. *anleiten*, *Anleitung*): „beidest þo fullkomlega fyrer rettenn frammlegged / þä siälfz herra biskupsenz anleiding sem lóglega umm þujlÿka inviklan hliöde“ Bps AIII 2, 270 (1693)
- *anvenda* so. ‘snúa, venda; nota’ (da. *anvende*; þý. *anwenden*): „ad oss veiter ecke af ad anvenda sierhvøriu augnablike vorra Lijfstunda þvi til fillingar“ BÞorlApp A, IIr

Um 1725:
- *anbefala* so. ‘mæla með’ (da. *anbefale*; þý. *anbefehlen*): „befalaði hans kóngl. majest biskupinum Jóni Árnasyni að troða í þá anbeföluðu Appollonio Svartzkopfs eftirmaalssags comnissivn“ JHBisk I, 396
- *anleggja* so. ‘leggja fram (ákæru)’ (da. *anlægge*; mlþ. *anleggen*): „eftir það hann hafði […] með mikilli frægð sína tíð til stúderinga anlagt“ Jóhann Þórðarson 1720, 44
- *antaka* so. ‘taka við, taka á móti’ (da. *antage*, myndað eftir (m)lþ. *annemen*, þý. *annehmen*): „hitt annad, sem hann vill færa kyrkiunne til skulldar, kann ei ad antakast“ Bps AII 17, 926 (1733)

Um 1750:
- *anbetrúa* so. ‘treysta e-m fyrir e-u’ (da. *anbetro*; þý. *anvertrauen*): „utan sinnar eigenn Anbetrwadrar Syslu“ Alþb 1758, C 2v
- *anfæra* so. ‘færa inn, tilgreina’ (da. *anføre*; þý. *anführen*): „Utgifft kyrkiunar fra 1751 er effter somu Reikningum her ä möt anfærd“ Bps BIII 17, 63 (1757)
- *anordning* kv. ‘tilskipun, fyrirskipun’ (da. *anordning* < *anordne*; þý. *Anordenung*): „effter Anordningu 1747“ Bps AII 19 III, 139 (1752)
- *anskaffelsi* hk. ‘öflun, útvegun’ (da. *anskaffelse* < *anskaffe*; þý. *anchaffen*): „graftoola vidhalld og anskaffelse“ Bps BIII 17, 249 (1760)

- *anstalt* kv., *anstaltir* kv.ft.[6] 1) 'viðbúnaður, umstang', 2) 'læti, gauragangur' (da. *anstalt*; þý. *Anstalt*): 1) „giôrer nu Proprietarius strax þä Anstalt […] ad bok þesse komest i gott stand" Bps BIII 17, 234 (1760); 2) „því ef þessar anstaltir skyldu kenna þér að sjá einhverja ögn að þér" MJLeik, 329
- *ansögning* kv. 'umsókn' (da. *ansøgning* < *ansøge*; mlþ. *ansōken*): „Hvar firer einnig Min underdanigasta ansógnïng Hier innleggst" *Bréf Gunnars Pálssonar* I, 27 (1753)
- *antakanlegur* lo. 'fullnægjandi, viðunandi' (da. *antagelig* < *antage*; sjá *antaka*, um 1725): „hefir […] ecki viljad frambjóda antakanlega borgun" ActYfirr 1750, 11
- *antekning* kv. 'viðtaka' (da. *antagning* († ODS) < *antage*; sjá *antaka*, um 1725): „um hvað Magnúsar Guðmundssonar antekningu til information […] áhrærir" TBókm XI, 193 (1760)
- *anvending* kv. 'notkun' (da. *anvende*, *anvendelse*, sbr. sæ. *användning*; þý. *anwenden*): „Giôfina Medtekur Profasturinn Med Þacklæte […] Enn hennar Anvending Jnnstiller hann til Ædre YferVallda Gott befindende" K VII1 A1, 280 (1770)
- *anvísning* kv. 'tilvísun, ábending; ávísun' (da. *anvisning* < *anvise*; mlþ. *anwīsen*): „effter kyrkiu bokarenar Anvisning" Bps AII 20, 11 (1755)

Um 1775:
- *angefari* kk. 'uppljóstrari' (da. *angiver* < *angive*; mlþ. *angeven*): „sektir […] skulu skiptast jafmt á milli angefarans og fátækra" Lovs IV, 230 (1776)
- *anledning* kv. (sjá *anleiðing*, um 1700): „í anledning af þeim seinustu harðindum" Blanda VIII, 34 (1785)[7]
- *anmerkning* kv. 'athugasemd' (da. *anmærkning* < *anmærke*; þý, *anmerken*): „ad giôra þar [þ.e. í bók] nockrar anmerkningar" *Bréf Gunnars Pálssonar* I, 208 (1768)
- *anmóða* so. 'fara e-s á leit, biðja' (da. *anmode*; mlþ. *anmōden*): „anmodast þvi Proprietarius ad tilhalda Soknar bændum ad láta hann [þ.e. kirkjugarð] Vera fullgiordann og gildann Jnnann 2^{ia} ára" Bps BIII 17, 453 (1781)

[6] RM hefur eitt dæmi um *anstalt* í hk. (IndrIndrDagur, 20, 1946).

[7] Dæmið er úr bréfi frá Jóni Eiríkssyni konferenzráði til Árna Þórarinssonar biskups en Jón hafði þá dvalist í Noregi og Danmörku frá 17 ára aldri og var kominn undir sextugt þegar hann skrifar bréfið sem er mjög dönskuskotið.

- *anskaffa* so. ‘útvega’ (da. *anskaffe*; þý. *anschaffen*): „Kyrkiu bókenn er á Enda, og á ein önnur afftur ad anskaffast“ Bps AII 23, 234 (1781)

Um 1800:
- *anstand* hk. ‘frestun málareksturs’ (da. *anstand*; mlþ. *anstand*): „Løgmadur […] begierir og fær Anstand til morguns“ Act-Yfirr 1792, 8
- *anstendugur* lo. ‘sómasamlegur’ (um kirkjugrip) (da. *anstændig*; þý. *anstendich*): „Nír Sylfur-kaleikur med Patínu af sama rett anstændugur“ Bps BIII 17, 543 (1791)
- *ansækja* so. ‘sækja um’ (da. *ansøge*; mlþ. *ansōken*): „hvar fyrir ég nú auðmjúkast ansæki, að mér fyrir hverja reisu mætti tilstandast 64 sk.“ SöguþLandp II, 21 (1792)
- *anvísa* so. ‘vísa á (þ.e. skrifa ávísun upp á)’ (da. *anvise*; mlþ. *anwīsen*): „Andvirðið anvísaði eg upp á mín tilkomandi laun, en tók sumt til láns“ GVídBr, 205
- *anvísing* kv., sbr. *anvísning*, 1750 (da. *anvisning* < *anvise*; mlþ. *anwīsen*): „effter anvísing Prestsins og diáknans“ Bps BIII 17, 480 (1790)

Um 1825:
- *andtigna* so. (sjá *antigna*, um 1625): 1) „Eins og Gydíngar andtignudu þá óumskornu Heidíngia, sem sauruga, so forsmádu þessir aptur á mót hina, sem þiód miøg hiátrúarfulla“ LeiðNT I, 169; 2) „er hönum þarí mest andtignað fyrir framför i guðfræði“ SvPJEir, 22
- *anklaga* so. ‘kvarta, klaga’ (da. *anklage*; mlþ. *anklagen*): „ég verð að anklaga fyrir yður með heyin, sem burthrifsuð hafa verið úr garði mínum“ Blanda VII, 259 (1829)
- *anmerkja* so. ‘gera athugasemd við’ (da. *anmærke*; þý *anmerken*): „þá anmerkir Gísli Ásmundsson í Nesi, að nefndri jörð tilheyri yzti og neðsti hólminn“ ArnSigEÁsm III, 195 (1818)

Um 1850:
- *anmerking* kv. ‘athugasemd’ (da. *anmærke*; þý, *anmerken*): „viðbætti eg þessum fáu anmerkingum“ Safn IV, 261 (1845)

- *anstæður* lo. ‘hneykslanlegur’ (da. *anstødelig*; myndað af da. *støde* (ísl. *steyta*) og þý. *an-*): „anstödelig; kerling sú hefur leingi anstæð verið“ Sch (seðlasafn)
- *antegnelse* kv. ‘athugasemd, ummæli’ (da. *antegnelse* < *antegne*; þý. *anzeichnen*): „Antegnelser og vanþökk hef eg fengid“ Sonurgull, 203 (1854)
- *antegnelsespóstur* kk. ‘athugasemdagrein’: „ýmislegt á ég eftir, sem mér sýnist minna liggja á svo sem antegnelsespósta besvarelse“ ÁrsrÍsf 1968, 142 (1843)

Um 1875:
- *anbefalning* kv. ‘meðmæli’ (da. *anbefaling*, *anbefalning* († ODS) < *anbefale*; þý. *anbefehlen*): „Eg lét Goos gefa mér anbefaling“ JGuðnSkTh I, 82 (1884)
- *anlegg* hk. ‘verstöðvarbygging(ar), mannvirki í verstöð’ (da. *anlæg*; sbr. þý. *Anlag*, mhþ. *anlāge*; sjá *anleggja*, um 1725): „ómögulegt var að halda þessari reglu nema í aðalkaupstöðum, en ekki á þeim svokölluðu „anleggjum““ Víkv 1874, 70

Um 1900:
- *anleggshús* hk. ‘verstöðvarbygging’ (sjá hér ofar): „Þegar komið var undir miðgóu, var von á að inntökumennirnir kæmu suður á „anleggshúsin“, sem þeir lágu við í ár eftir ár“ Amma, 299
- *anmelda* so. ‘ritdæma’ (da. *anmelde*; þý. *anmelden*): „Þú vilt náttúrlega ekki vinna til að senda mér exemplar af Páli Vídalín til þess að eg anmeldi hann í Stefni, […]?“ Margtsend, 263 (1898)
- *anretningsborð* hk. ‘framreiðsluborð’ (da. *anrette*; mlþ., þý. *anrichten*): „Borð alls konar smá og stór af öllum gerðum, þ. á. m. Anretningsborð“ Ísaf 1904, 88

Um 1925:
- *anleiðning* kv. (sjá *anleiðing*, um 1700, *anleðning*, um 1775): „og [Kristrún gamla] reyndi að taka sér það til anleiðningar“ GHagalKH, 180
- *anretterborð* hk. (sjá *anretningsborð*, um 1900): „anretterborð með kaffistelli“ Skírn 1928

- *anstandsdama* kv. ‘kona sem gætir velsæmis ungrar stúlku’ (da. *andstandsdame* < *anstå* ‘hæfa, sæma’; mlþ. *anstan* ‘hæfa, sæma’): „fólk, sem sæi okkur saman, myndi álykta, […] að ég væri að ganga með kærustunni og Petrea væri anstandsdama, til að hafa umsjón með, að skírlífi væri borgið“ SvbEgFerð II, 823

Um 1950:

- *anstalta* so. ‘sinna verkum, hafa viðbúnað’ (da. *anstalte*; sjá *anstalt*, um 1750): „Hún vill helzt alltaf vera eitthvað að anstalta úti við“ ÞórlBTröllið, 76
- *anstíga* so. „koma anstígandi“ ‘koma askvaðandi’ (da. *komme anstigende*; þý. *ansteigen*): „þessar falssaklausu hænur, sem leggja reyndar á rás, þegar tignin og skartið og mátturinn koma anstígandi“ GHagalKon, 104

Tiltölulega auðvelt er að finna fleiri dæmi um sum orðin, t.d. í bréfasöfnum og annálum síðari alda, en sem fyrr segir er *anseelse* eina nýja orðið fyrir utan þau sem hér voru talin sem fannst í ritum sem skoðuð voru af þessu tilefni.[8]

Þegar skráðum dæmum er skipt lauslega í flokka eftir textategundum sést að rúmur þriðjungur er úr opinberum skjölum af einhverju tagi og annað eins úr ýmsum fræðitextum.[9] Um 10% eru úr bréfum. Einungis örfá dæmi eru úr kveðskap (*anslag* 17. öld, 2 dæmi; *antaka* 1841; *antigna* 17. öld, 2 dæmi) og þjóðsögum (*anstaltir* 1850, *antigna* 1925). Þetta er athyglisvert því að sum *an*-orð voru án efa alþýðumál – og nokkrir rithöfundar á 20. öld notuðu einmitt slík orð

[8]RM hefur eitt dæmi um *ansigt* frá fyrri hluta 20. aldar: „Næst þegar ég hitti hann ætla ég að vera búinn að fá mér Teleskop til að studera á honum ansigtið“ (MagnStefBréf, 104, Örn Arnarson skáld). Orðið er hér trúlega notað í hálfkæringi og gamansemi. – Rétt er að nefna orðin *ansvar* ‘ábyrgð’, *ansvarsmaður*, *ansvarlegur*, *ansvarsfullur*, þar sem *an-* á sér norrænan uppruna, *and-*; en orðmyndir með *an-* eru vafalaust undir dönskum áhrifum og seinni tvö orðin má telja tökuorð úr dönsku. (Sbr. *andsvar*, *andsvarsmaður*, ONP.) – Eitt dæmi er í RM um orðið *anmálsemi* kv. ‘tal, orðsemi’ (e.t.v. ‘ýtni’): „Friðrik hafði margsinnis beðið menn þeirra að líta til festarinnar, og urðu þeir loks fyrtnir af hans anmálsemi“ FrEggFylg II, 126. Beinar samsvaranir í öðrum málum hafa ekki fundist; seinni hluti orðsins er ljóslega af sama toga og *málsemd* ‘þvaður, mælgi’ og fyrri hlutann ber sennilega fremur að tengja *ann-* (*önn*) en þýska forskeytinu þótt ekki sé hægt að útiloka áhrif frá tökuorðum.

[9]Til „fræðitexta“ voru talin sagnfræðileg rit, annálar, rit um ævisögulegt efni, guðfræðirit og fleira í þeim dúr.

í stílfræðilegum tilgangi og lögðu í munn alþýðufólks, eins og síðar verður rætt. Dæmafæð í þjóðsögum má kannski að hluta skýra með því að þær hafi að einhverju leyti verið lagaðar að kröfum um gott mál þegar þær voru skráðar.

Um 20% dæmanna eru svo úr skáldverkum og eru þau langflest frá 20. öld; verður vikið nánar að þeim síðar.

Við fáum þá mynd af gögnum OH að fæst *an*-orðin hafi verið algeng. Tæp 60% orða og orðmynda eru eindæmi (25 orð) eða koma einungis tvívegis fyrir þar (9 orð):

(2) *anbefala* s.hl. 18. aldar og 1885, *anbefaling* 1884, *anbetrúaður* 1758, *anganga* um 1700 og 1758, *angefari* 1776, *angefningarpóstur* um 1700 og um 1725, *anhang* um 1650, *anklaga* 1829, *anleggja* 1770 og um 1800, *anmelda* 1898, *anmerking* 1845, *anmerkja* 1818, *anmæli* um 1600 og um 1800, *anordning* 1752 og 1945, *anpartur* um 1700, *anretningsborð* 1904, *anretterborð* 1928, *anskaffa* 1781, *anskaffelsi* 1760, *anslagur* 1602, *anstalt* hk. 1946, *anstalta* 1950, *anstand* 1792 og 1826, *anstandsdama* um 1925, *anstendugur* 1791, *anstíga* 1945, *anstæður* um 1850, *ansækja* 1792, *antasta* 1671, *antegnelse* 1854, *antegnelsespóstur* 1843, *antekning* 1760, *anvending* 1760 og 1770, *anvísa* um 1800 og 1832, *anvísing* 1790

Sennilega mætti finna fleiri dæmi við nánari athugun en lág tíðni gefur samt til kynna að um fágæt orð sé að ræða.

Rétt er að nefna að sum eindæmin (þau yngstu) eru úr textum sem eru sérstakir á einhvern hátt: eina dæmið um *anmelda* er úr galsafengnu einkabréfi (1898); *anretningsborð* er úr auglýsingu í *Ísafold* 1904; dæmið um orðið *anretterborð* (1928) er úr grein Guðmundar Finnbogasonar „Hreint mál“ og tekið þar sem dæmi um hið „argvítugasta hrognamál“ (Guðmundur Finnbogason 1928:147);[10] sögnin *anstalta* er úr skáldsögu (1958); *anstandsdama* er nefnd í ferðaminningum frá 3. áratug 20. aldar; *anstíga* („koma anstígandi“) er úr skáldsögu (1945); *anstæður* er af seðli úr safni Hallgríms Schevings, án tilvísunar í heimild, og óvíst hvort dæmið er úr talmáli eða ritmáli; *antegnelse* er úr

[10]Guðrún Kvaran (2002) ræðir efni greinar Guðmundar og getur um orðin *anrettuborð* og *anrettuherbergi* (da. *anretterværelse*) sem heimildarmenn Orðabókar Háskólans þekktu.

einkabréfi frá 1854 og er þar í sinni dönsku mynd („Antegnelser og vanþökk hef eg fengid“); og *antegnelsespóstur* (1854) er úr mjög dönskuskotnu bréfi. – Ofangreind dæmi þurfa ekki endilega að vera ótækar heimildir um notkun viðkomandi orða en sérstakt stílgildi sumra textanna og ungur aldur annarra getur bent til þess að ekki sé um dæmigerða orðanotkun að ræða.

Nokkur orð koma fyrir þrisvar til fimm sinnum:

(3) *anlegg* 4, *anleggshús* 3, *anmóða* 3, *annáma* 4, *anslag* 4, *antakanlegur* 5, *anvísning* 4

Og nokkur orð eru tiltölulega algeng:

(4) *an*(*d*)*tigna* 29, *anfæra* 10, *angefa* 17, *angefning* 18, *anle*(*i*)*ð*(*n*)*ing* 12, *ansjá* 12, *anstalt* 22, *ansögning* 8, *antaka* 38, *anvenda* 11

Að minnsta kosti þessi orð hafa verið farin að festa rætur í málinu á seinni öldum og trúlegt má telja að dæmi um ýmis hinna sjaldgæfari orða leynist í textum sem ekki hafa verið gefnir út eða orðteknir. Ólíklegt er að mörg orð hafi farið fram hjá fránum augum orðtökumanna en þó má ekki útiloka slíkt.

3 Athugun á nokkrum textum frá 17.–19. öld

Nokkrir textar frá undanförnum öldum voru athugaðir nánar fyrir þessa rannsókn og leitað dæma um *an*-orð. Slík leit er tímafrek og var látið nægja að lesa nákvæmlega eina bréfabók auk þess sem fanga var leitað í tölvutækum textum og orðaskrám orðabóka.

Þess er fyrst að geta að ekkert orð með forskeytinu *an-* er að finna í Passíusálmum Hallgríms Péturssonar (1614–1674) en þeir komu fyrst út 1666.[11] Sögnin *antigna* ‘lasta’ kemur hins vegar fyrir í einu af veraldlegum kvæðum hans (HPSkv II, 184).

Ekkert dæmi um orð af þessu tagi er í *Lexicon Islandicum*, orðabók Guðmundar Andréssonar (d. 1654) sem út kom 1683.

Eigi heldur er nokkurt *an*-orð að finna í postillu Jóns Vídalíns (1666–1720) sem kom út 1718–1720.[12]

[11] Sjá *http://lexis.hi.is/ordlyklar/salmar/salmar.htm.*

[12] Ég þakka útgefendum *Vídalínspostillu* (1995) fyrir veittan aðgang að tölvutækum texta.

Í *Nucleus Latinitatis*, latneskri orðabók Jóns Árnasonar (1665–1743) frá 1738, er eitt *an*-orð notað í íslenskum skýringum hans, *anleiðing* („**Causa** … Ordsøk, Anleiding, Tilefni, Tilstille“ bls. 28; „**Materia** … Orsøk, Anleiding“ bls. 159). Telja má að einhver hefð hafi verið komin á þetta orð í máli lærdómsmanna úr því að það er tvívegis notað á þennan hátt.

Í orðabók Jóns Ólafssonar úr Grunnavík (1705–1779), í handritinu AM 433 fol., er einungis að finna orðin *anstalt* og *antigna* og ekkert orð af þessu tagi er í riti hans *Hagþenki*.[13] Í hugleiðingum sínum árið 1759 um sótt og dauða íslenskrar tungu gagnrýnir Jón notkun á erlendum orðum í íslensku og segir m.a. að til séu þeir sem ekki viti „hvad i Islendsku skylldi heita **Ordit Ansögning** (*ummbeidni*, edur *Eptirleitni*); og enn sijdur Ordit **Anstalt**, sem ymist er *Undirbwnjngr*, edur *Ræda gjǫrd*, eptir þvi sem Efninu vid-hagar“ (Jón Ólafsson 1998:152) og eru þetta fyrstu merkin, mér kunn, um andóf gegn orðum af þessu tagi.

Í bréfum séra Gunnars Pálssonar (1714–1791) er nokkur *an*-orð að finna: *angefa* (Bréf I:407), *anmerkning* (208, 275), *anordning* (12), *anstalt* (163, 267, 436), *anseelse* ‘álit, virðing’ („med hǫfudanseelse“, 386; da. *anseelse* < *anse*; mlþ. *ansēn*) og *ansögning* (27), en *anordning* er þar eiginlega notað sem stytting á titli reglugerðar, „Anordning om de latinske skoler på Island, 1743“.[14] Séra Gunnar var skólameistari á Hólum og prestur í Hjarðarholti og hefur verið talinn með lærðustu mönnum sinnar tíðar (Bréf II:11 o.áfr.). Varðveitt bréf frá honum eru tæplega 200 talsins, rúmar 500 blaðsíður í útgáfu. Tökuorð eru mörg, bæði af dönskum og latneskum uppruna.[15] Til samanburðar má geta þess að tökuorð sem hefjast á *be-* í bréfum hans eru tíu (*begera*, *benægtelseseiður*, *beskyldning*, *bestilla*, *bestilling*, *besværing*, *betala*, *betient*, *bevísa* og *yfirbevísa*, *bevísing* og *yfirbevísing*) og tökuorð með forskeytinu *for-* eru á milli 30 og 40. – Þess má geta að í stafsetningarkennslubók Gunnars frá árinu 1782, *Lijtid Wngt Støfunar Barn*, eru sárafá ung tökuorð og má væntanlega skýra mikinn fjölda tökuorða og erlendra slettna í bréfun-

[13] Ég þakka útgefanda *Hagþenkis* fyrir veittan aðgang að tölvutækum texta.

[14] Að auki kemur orðið *anseeligt* fyrir í einu bréfa Gunnars: „Mer virdiz ej miǫg anseeligt, hvorki fyrir þá … ne mig“ (Bréf I:291); líklega er rétt að líta á það sem hreina danska glósu.

[15] Orð af erlendum toga eru vel yfir 1000 í bréfum hans (stök orð og samsetningar). Latínuglósur, sem margar eru beygðar að íslenskum hætti, merkir Gunnar sérstaklega (skrifar með latínuletri) en tökuorð sem hann merkir ekki á þennan hátt, og eru yngri en frá um 1500, skipta hundruðum.

um með því að slíkt tilheyrði textategundinni á þessum tíma – a.m.k. í skrifum lærðra manna.

Í riti Magnúsar Ketilssonar um stiftamtmenn og amtmenn á Íslandi 1750 til 1800, skrifað líklega 1802 (Magnús Ketilsson 1948:14), koma fyrir nokkur orð af þessu tagi, svo sem *angefa* 'segja, gefa upp' (44), *angefning* 'kæra' (25), *anleggja* (51) og *ansögning* 'umsókn' (31). Verkið er mjög dönskuskotið og sést það vel með því að tilfæra nokkur helstu orð af erlendum toga sem er að finna á sömu blaðsíðum og ofangreind fjögur orð:

(5) aldeilis, alleina, angefa, angefning, ansögning, behalda, betala, brúka, capellan, compagnie, confessionarius, dempa, direktor, erklæra, exeqvera, fríheit, fullmektugur, höndlun, innheimta, inntekt, mótpartur, ordinantía, óafgjörður, óbetalaður, privat, provision, rescript, restanc, suspendera, tilláta, útvegur

Hugsanlegt er að Magnús sé höfundur greinar sem birtist í *Islandske Maaneds-Tidender* árið 1776 og ber titilinn „Kort Betænkning om et nyt Sprog paa Island", og er ádeila á hugmyndir um að taka upp dönsku í stað íslenskrar tungu í landinu (Bréf II:185). Séra Gunnar Pálsson hefur einnig haft nokkrar áhyggjur af framtíð tungunnar sem er eitt hið „Serligasta raritet og hnoss, Sem heila Evropa hefr, og ein med Stærstu heimsins gersemum teliandi" (Bréf I:435). Í einu bréfa sinna (Bréf I:99–111) gagnrýnir hann kveðskap og að nokkru leyti orðafar í sálmabók 1757 og Hallgrímssálmum 1759; og í bréfi frá 1784 mærir hann tunguna (Bréf II:35). Í ljósi þessa er vert að gefa því gaum að tökuorð og notkun erlendra glósna virðist ekki trufla Gunnar né Magnús. Slíkt kemur að vísu ekki á óvart; merki um málhreinsun koma tæpast fram fyrr en hjá næstu kynslóð á eftir þeim og þeir virðast ekki hafa litið svo á að upptaka tökuorða og notkun framandorða væru lýti í málinu.

Loks var leitað í dagbókum Steingríms Jónssonar (1769–1845), síðar biskups, úr ferðum hans með Hannesi Finnssyni um landið árin 1791–1795, sem varðveittar eru í handritinu Lbs. 95 8°, bl. 1r–52v. Þar koma ekki fyrir orð með forskeytinu *an-* og ung tökuorð eru þar tiltölulega fá.

Í safni Netútgáfunnar (*http://www.snerpa.is/net/*) eru verk frá ýmsum tímum, þar á meðal fáein 19. aldar verk. Engin orð af því tagi sem hér var leitað er að finna í tveimur smásögum Jónasar Hallgrímsson-

ar, *Grasaferð* og *Þegar drottningin á Englandi fór í orlof sitt*, né í *Sögunni af Heljarslóðarorrustu* og *Þórðar sögu Geirmundssonar* eftir Benedikt Gröndal, né í smásögum Gests Pálssonar (*Grímur kaupmaður deyr*, *Hans vöggur*, *Kærleiksheimilið*, *Skjóni*, *Uppreistin á Brekku*, *Vordraumur*) eða í smásögunni *Nýi hatturinn* eftir Stephan G. Stephansson. Engin *an*-orð er heldur að finna í *Pilti og stúlku* eftir Jón Thoroddsen (1850); hann lætur konu nokkra tala afskaplega dönskuskotið mál og notar þá t.d. *befatta*, *bestemt* og *begrípa* en engin *an*-orð. Í *Manni og konu* (1876) kemur fyrir lo. *anstendugur* í texta sem á að vera tilvitnun í gamla biskupsvísitasíugjörð (í þeim texta er einnig sögnin *beþéna*; önnur *be*-orð í skáldsögunni eru *betala* tvisvar, *begríba*, *behalda*, *beþenkja*, *velbeþéntur*, einu sinni hvert) og tvisvar sinnum orðalagið *e-ð kemur an upp á* („Það kemur allt an uppá, hvern veginn maður fer“, „þar kemur an upp á“) sem líklega er dregið af da. *komme an på*. Önnur *an*-orð er þar ekki að finna. Í báðum skáldverkum Jóns er allmikið um dönsk tökuorð í samtölum.

Þessi takmarkaða athugun á nokkrum textum frá 17.–19. öld bendir ekki til þess að *an*-orð hafi verið algeng þá og styrkir, ef eitthvað er, þá mynd sem RM gefur.[16]

4 *an*-orð í 20. aldar máli

Af því sem hér er komið fram má ætla að *an*-orð í íslensku hafi í tímans rás ekki verið umtalsvert fleiri en þau u.þ.b. 60 sem hér hafa verið nefnd til sögunnar. Mörg orðanna voru fágæt en um 15–20% þeirra hafa verið tiltölulega algeng og sum hver voru notuð í málinu um 2ja til 3ja alda skeið. Í upphafi 2. kafla kom fram að einungis tvö orð af þessum toga sé að finna í íslenskum nútímamálsorðabókum: *anstaltir* kv.ft. og sögnina *antigna*.

Lítill vafi leikur á því að þessi orð hafa einkum tínt tölunni í málhreinsun 19. aldar og lágu þau vel við höggi enda auðþekkt. Þau virðast hverfa að mestu úr ritmáli þegar dregur að lokum 19. aldar. Í máli alþýðunnar lifðu nokkur þeirra eitthvað fram á 20. öld, jafnvel fram

[16] Áhugavert hefði verið að skoða fleiri texta til að skerpa myndina enn frekar; til dæmis hefur RM fáein dæmi um *an*-orð úr leikritum Matthíasar Jochumssonar: *annamma* (MJLeik, 448), *anstalt* (MJLeik, 329), *antigna* (MJLeik, 520 – einnig í bréfi Matthíasar til Hannesar Hafsteins, MJBrHH, 66) – og vekur það grunsemdir um að fleira slíkt kynni að finnast í verkum hans ef vel væri að gáð.

yfir miðja öldina eins og dæmin í Talmálssafni OH sýna, en trúlega eru þau dáin út nú.[17]

Þegar *an*-orðin voru svo gott sem horfin úr málinu eignuðust sum þeirra þó svolítið eftirlíf í skáldritum 20. aldar, og er rétt að greina dálítið frá því. Skáldsagnahöfundar lögðu nefnilega stundum eitt og annað fornfálegt í munn gamals fólks eða notuðu slík orð til að færa sögusviðið eilítið aftar í tímann. Þannig frestuðu þeir dauða nokkurra af orðunum um fáeina áratugi þó að það „líf" sem þau eignuðust þannig hafi aldrei verið annað en gálgafrestur.

Árið 1933 kom út bók Guðmundar G. Hagalín, *Kristrún í Hamravík*. Þeir vita sem lesið hafa þessa bók að málfar á henni er mjög sérstakt og sérkennilegt, og hafa margir á það bent. Guðni Jónsson skrifaði fyrsta ritdóminn um bókina árið 1933 og segir:

> Því fer fjarri, að málið sje alstaðar hreint. Margt er um útlend orð að uppruna til, eins og alþýðumálið er líka auðugt af þeim. – Fram hjá slíkum orðum hafa flestir rithöfundar gengið og sett sín orð í staðinn. En Guðmundur hefir látið alt slíkt halda sjer og má því telja þetta með einkennum þessa alþýðustíls.
>
> (Guðni Jónsson 1933:6)

Þess má geta að Matthías Johannessen hefur skrifað greinargóða ritgerð um *Kristrúnu í Hamravík* (Matthías Johannessen 1985). Þar fjallar hann meðal annars um málfar bókarinnar og gerir grein bæði fyrir tökuorðum og sérvestfirskum orðum sem í henni er að finna.

Rithöfundar sem skreyttu verk sín með sjaldgæfum orðum og tökuorðum þekktu án efa þess kyns orð úr eldri ritum eða í munni gamals fólks. Víst er einnig að Hagalín var djarfari en flestir samtímamenn hans í því að brjóta gegn óskráðum reglum málhreinsunarstefnunnar sem ríkt hafði allt frá tímum Sveinbjarnar Egilssonar, að

[17] Ég hef spurst lauslega fyrir um orð af þessu tagi og hafa fæstir kannast við nokkur slík. Guðrún Þórðardóttir (f. 1936) tjáði mér þó í bréfi að hún þekkti orðin *anstaltir* 'umstang, vesen, fyrirhöfn, stúss' og *anstendugur* 'sómasamlegur, siðsamlegur', en ekki úr eigin máli heldur úr máli ömmu sinnar (f. 1875) og móður (f. 1901), og a.m.k. tvær konur (f. 1930 og 1934) og einn karl (f. 1933), sem hún ræddi við, mundu þessi orð vel. Hún telur að þau hafi verið algeng í daglegu máli Reykvíkinga af eldri kynslóðinni þegar hún var að alast upp en minnist þess þó ekki að faðir hennar hafi notað orðin, en hann var úr sveit, háskólamenntaður og líklega málhreinsunarsinni, segir hún.

segja má. Bókmenntagagnrýnandinn Erlendur Jónsson telur, í grein um *Kristrúnu í Hamravík*, að Hagalín hafi orðið „einna fyrstur íslenzkra rithöfunda til að brjótast undan oki málvöndunarstefnunnar“ (Erlendur Jónsson 1966:10). Fróðlegt er að lesa orð Guðmundar G. Hagalíns sjálfs í inngangi að útgáfu *Kristrúnar í Hamravík* árið 1966, en þar segir hann:

> Þá er ég svo fór að skrifa sögur … fann ég ríka hvöt hjá mér til að ganga lengra í því að nota daglegt mál í samtölum og ýmis alþýðleg orð og orðatiltæki heldur en mér eldri íslenzkir rithöfundar, en fylgja aftur á móti hinni almennu málhreinsunarstefnu í öllu, sem ekki væri talað fyrir munn sögupersónanna.
>
> (Guðmundur G. Hagalín 1966:27)

Um þau ‘alþýðlegu orð og orðatiltæki’ sem Hagalín notar í *Kristrúnu í Hamravík* segir hann að mikinn meginhluta þeirra hafi hann heyrt vestra og lætur þess einnig getið að Þórleifur Bjarnason rithöfundur hafi „heyrt þau flestöll í bernsku norður á Hornströndum, og er hann þó maður 10 árum yngri en ég“ (bls. 29). Enn fremur viti hann til þess að margt þeirra hafi verið til víða um land og vitnar um það til dr. Stefáns Einarssonar málfræðings sem þekkti mörg þeirra úr Breiðdal eystra.

Meðal sérkennilegra orða í sögum Hagalíns eru einmitt fáein *an*-orð. Í RM eru dæmi um þessi:

(6) *angefa* (GHagalKH, 14)
anleiðning (GHagalKH, 180, GHagalHam, 75)
anordning (GHagalKon, 155)
anstalt (GHagalKH, 19, 27 og 137, GHagalStV II, 171, GHagalHam, 44)
anstíga (GHagalKon, 104)
antaka (GHagalKon, 346, GHagalMaríum, 93, GHagalHam, 74)
antigna (GHagalMaríum, 49, GHagalHam, 142 og 154)

Nánari leit í sögunni af Kristrúnu í Hamravík leiddi að auki í ljós sögnina *anstalta* (1966:101) og ekki kæmi á óvart að fleira slíkt leyndist í öðrum sögum Hagalíns.

En fleiri skáld en Guðmundur Hagalín lengdu líf *an*-orða og notuðu í ritverkum sínum. Halldór Laxness var líka þekktur fyrir sérstakt málfar og óvenjulegt en í RM eru dæmi úr ritum hans um þessi orð:

(7) *angefa* (HKLSalka, 184, HKLHeimsl I, 34, HKLGuðsg, 76)
anleiðing (HKLSalka, 81, HKLBrekk, 62, HKLKristn, 237)
anstalt (HKLSalka, 17)
antaka (HKLPar, 31)
antigna (HKLSjfólk, 213, 270, HKLHeimsl II, 199, HKLÍsl, 166)

Aðrir 20. aldar rithöfundar voru tæpast jafn-djarftækir og Hagalín og Laxness en nokkrir skreyttu þó verk sín með orðum af þessu tagi þó í minna mæli væri. Í RM eru dæmi úr ritum eftirfarandi rithöfunda:

(8) Björn J. Blöndal: *anstalt* (BjBlöndÖrl, 67)
Guðmundur Daníelsson: *anstalt* (GDanBolafl II, 189, GDanJörð, 228)
Guðmundur Kamban: *anstalt* (GKambSkálh II, 105)
Jóhannes út Kötlum: *antigna* (JóhKötlSigl, 158)
Jón Björnsson: *anstalt*, *annamma* (JBjörnMátt, 323; JBjörnJómf, 212)
Kristmann Guðmundsson: *angefa* (KristmGStutt, 215)
Stefán Júlíusson: *antigna* (StJúlSól, 87)
Torfhildur Þ. Hólm: *anstalt* (THJA II, 189)
Þórleifur Bjarnason: *angefa*, *anstalta* (ÞórlBTröllið, 175; ÞórlBTröllið, 175)

Vafalaust er víðar að finna orð af þessu tagi í skáldsögum en hér verður látið nægja að skoða til viðbótar þær 20. aldar skáldsögur og smásögur eftir Jón Trausta, Torfhildi Hólm og Þorgils gjallanda sem eru í safni Netútgáfunnar. Leitað var í þeim að orðum sem hefjast á *an-* en jafnframt athugað hvort fyrir kæmu orðin *brúka*, *blífa* og *ske* eða einhver samsetning eða afleiðsla með þeim orðstofnum, svo og orð sem hefjast á *be-*; var þetta gert til að fá einhverja hugmynd um það hversu bundnir höfundar hefðu verið af kröfum málhreinsunar en sem vel er kunnugt voru orð af þessum toga meðal helstu skotspóna málhreinsunarstefnunnar. Niðurstaðan varð sú að engin *an*-orð fundust

í 8 skáldsögum og 21 smásögu. Ekki fundust þar heldur dæmi um *be-*orð né sögnina *blífa*, en nokkur dæmi um *brúka* og samsetningar, svo og um *kannski*, *kannske* og *máske*, og örfá dæmi um *ske*.[18] Nefna má að smásagan *Sýður á keipum* eftir Jón Trausta ber undirtitilinn „Saga frá byrjun 17. aldar“ og *Söngva-Borga* undirtitilinn „Saga frá fyrri hluta 16. aldar“ og hafa báðar að geyma talsvert af samtölum en höfundur hefur ekki farið þá leið að leggja sögupersónum sínum orðfæri sögutímans í munn, a.m.k. ekki orð af erlendum uppruna.

Eftirtektarvert er að *an*-orð í skáldsögum 20. aldar rithöfunda koma einna helst fyrir í sögum sem gerast í þeirra eigin samtíma eða skömmu fyrir hann. Höfundarnir virðast hafa litið svo á að þeir væru að sýna málfar gamals fólks á þeirra tíð en leggja orðin síður í munn fólks fyrr á öldum. Þetta segir sitt um stöðu orðanna í ungdæmi höfundanna (snemma á 20. öld).

Vöntun alls kyns ‘alþýðlegra’ orða af erlendum, einkum dönskum, uppruna í ritum margra 20. aldar höfunda segir að sumu leyti meira um afstöðu þeirra til íslenskrar málstefnu á fyrri hluta aldarinnar – og vald stefnunnar yfir penna þeirra – en um raunverulegt málfar sem þeir ólust upp við. Vissulega má sjá dálítið eftir þeim orðaforða sem við það komst aldrei á prent en í ljósi sögunnar er þetta vel skiljanlegt og gildir reyndar að nokkru leyti enn þann dag í dag.

5 Örlög orðanna

Spyrja mætti um tvennt: Hvers vegna bárust ekki fleiri orð af þessu tagi inn í málið, og hins vegar: Hvað veldur því svo að þessi orð hverfa?

Ekki er gott að gefa einhlítt svar við fyrri spurningunni. Hví urðu orðin ekki fleiri en raun ber vitni, og algengari? Tökuorð síðustu fimm hundruð ára eru mörg, fleiri en margir gera sér grein fyrir, og í fljótu bragði hefði mátt ætla að *an*-orð hefðu getað átt álíka auðvelda leið inn í málið og mörg önnur orð. Gott er að hafa eitthvað til samanburðar og þess vegna má geta þess að hátt í 300 tökuorð (nálega 130 orðstofnar, lauslega talið) með forliðum *be-* og *bí-* er að finna í söfnum OH. En

[18] Til samanburðar skal þess getið að engin dæmi um *blífa*, *brúka* og *ske* er að finna í verkum Jónasar Hallgrímssonar, Benedikts Gröndal, Gests Pálssonar og Stephans G. Stephanssonar sem getið var í kaflanum um verk frá 17.–19. öld nema eitt dæmi um *ske* hjá hinum síðastnefnda.

margt bendir til þess að með nokkrum undantekningum hafi *an*-orð í íslensku einkum verið bundin ritmáli embættismanna og lítt farið út fyrir þeirra hóp, jafnvel minna en *be*-orðin sem þó voru varla hvers manns eign. Líklega hefur notkun margra *an*-orðanna meira að segja verið enn þrengri; þau voru í eðli sínu framandorð, fremur glósur eða ívitnanir en eiginleg tökuorð,[19] og skutu upp kollinum í máli embættismanna sem margir hverjir höfðu dvalist í Danmörku, skrifuðu sína texta með hliðsjón af dönskum textum eða voru í samstarfi við danska embættismenn og tóku upp orð eftir þeim.

Þess má geta að einnig í dönsku tilheyrðu þessi orð fyrst og fremst „det stivere skriftsprog" að sögn málfræðingsins Peters Skautrup (Skautrup 1947:234). Skautrup hefur og bent á að í dönsku hafi forskeytið *an-* aldrei haft neitt skilgreint og einsleitt hlutverk og hafi aldrei verið skynjað sem eiginlegt orðmyndunaraðskeyti (s.st.). Í sumum tilfellum voru sjálf hugtökin, sem *an*-orðin tjáðu, tekin upp í heild sinni og þau höfðu þá engin tengsl við danskan orðaforða. Í öðrum var seinni hluti orðanna (aftan forskeytis) fyrir hendi í dönsku sem hluti sameiginlegs, germansks orðaforða og þá gátu tökuorðið og ósamsetta erfðarorðið stundum skipst á án þess að um merkingarmun væri að ræða. Takmarkaður dæmafjöldi *an*-orða í íslensku leyfir sjaldnast að þessi sama ályktun verði dregin um þau en þó eru fáein dæmi til um slíkt, svo sem sagnirnar *anfæra* („Brotastýll sendur til Kaupmhafnar, anfærdur til inntektar í sídasta reikníngi" Klp VIII, 211, 19f) og *antaka* („þessar bækur vill Biskupenn ecke antaka uppi Skulld kyrkiunnar" Bps AII19 III, 113, 1752).

Ekki þarf samt að leita langt að einni ástæðu þess að orðin urðu ekki fleiri en raun ber vitni, og að þau nái t.d. ekki nema um 20% af fjölda *be*-orða sem bárust inn í málið á svipuðu tímabili. Hlutfallslega eru *an*-orð nefnilega miklu færri í dönsku, norsku og sænsku en orð með forskeytinu eða forliðnum *be-* í þessum málum og því er eðlilegt að færri orð með *an-* bærust inn í íslensku en þau fyrrnefndu. En báðir þessir orðahópar skera sig úr hinum innlenda orðaforða og voru því tiltölulega auðveldir viðureignar þegar að hreinsun málsins kom.

[19]Sbr. umfjöllun um hugtökin *tökuorð* og *framandorð* í Veturliði Óskarsson 2003:95 (og tilvísanir þar).

Seinni spurningunni hefur þegar verið svarað; þarna var málhreinsunin að verki. Það kann að vísu að koma á óvart hversu rækileg þessi tiltekt í tungumálinu var en það er ekki innan markmiða þessarar ritgerðar að rýna nánar í þau öfl sem þar voru á ferð; látið verður nægja að slá fram þeirri fullyrðingu að sambærileg hreinsun myndi tæplega takast nú.

6 Lokaorð

Hér var greint frá afmörkuðum hópi tökuorða með forskeytinu *an-*, innreið þeirra í íslenska tungu og afdrifum. Norræn mál voru fremur fátæk að forskeytum og viðskeytum þegar áhrif lágþýsku hófust á 13. öld. Meðal annars þess vegna áttu orð með aðskeytum tiltölulega greiða leið inn í málin, og sum hinna erlendu aðskeyta urðu smám saman virk í innlendri orðmyndun í dönsku, norsku og sænsku. Orð af þessu tagi síast hægt og rólega inn í íslensku í aldanna rás; þau verða að vísu ekki mörg en þó er ein 60 *an*-orð og hátt í 300 orð með forliðum *be-* og *bí-* að finna í söfnum OH.

Með hliðsjón af því hversu tiltölulega mörg dæmi (um 250 talsins) eru skráð í RM um orð með *an-* og hversu algeng nokkur þeirra virðast hafa verið er athyglisvert að þessi hópur orða skyldi hverfa jafnrækilega og raun ber vitni. Við sjáum að vísu *an*-orðin berast inn í málið fram eftir 19. öld og allt fram á 20. öld en í lokin eru þau helst notuð í stílfræðilegum tilgangi.

Ljóst er að brotthvarf þessara orða er afleiðing málhreinsunarbaráttu 19.–20. aldar enda eru orðin tiltölulega auðþekkjanleg og því auðvelt að benda á þau til varnaðar. Hvers vegna orðin urðu ekki fleiri í íslensku í tímans rás er e.t.v. ekki jafn augljóst.

Heimildir

Ásgeir Blöndal Magnússon. 1989. *Íslensk orðsifjabók*. Reykjavík: Orðabók Háskólans.

Braunmüller, Kurt. 2000. Højtysk som 'naturlig' fortsættelse af den nedertyske sprogkontakt i Norden i 1500-tallet? Í: Ernst Håkon Jahr (ritstj.). *Språkkontakt – Innverknaden frå nedertysk på andre nordeuropeiska språk*. Skrift nr. 2 frå prosjektet Språkhistoriske prinsipp for lånord i nordiske språk, bls. 277–288. København: Nordisk ministerråd.

Bréf I = *Bréf Gunnars Pálssonar*. 1984. I. Texti. Gunnar Sveinsson bjó til prentunar. Rit 26. Reykjavík: Stofnun Árna Magnússonar á Íslandi.

Bréf II = *Bréf Gunnars Pálssonar*. 1997. II. Athugasemdir og skýringar. Gunnar Sveinsson bjó til prentunar. Rit 43. Reykjavík: Stofnun Árna Magnússonar á Íslandi.

DN = *Diplomatarium Norvegicum. Oldbreve til Kundskab om Norges indre og ydre Forhold, Sprog, Slægter, Sæder, Lovgivning og Rettergang i Middelalderen* 1–22. 1847–1992. Christiania, Bergen, Oslo.

Dollinger, Philippe. 1981. *Die Hanse*. 3., überarb. Auflage. Stuttgart: Kröner.

Donsk-føroysk orðabók. 1995. Ritstjórar: H.P. Petersen og M. Staksberg. Tórshavn: Føroya Fróðskaparfelag.

Erlendur Jónsson. 1966. Hagalín fyrr og nú. *Morgunblaðið* 16. desember, bls. 10.

Föroysk orðabók. 1998. Ritstjórar: J.H.W. Poulsen, M. Simonsen, J. í L. Jacobsen, A. Johansen og Z.S. Hansen. Tórshavn: Føroya Fróðskaparfelag.

Guðmundur Finnbogason. 1928. Hreint mál. *Skírnir* 102:145–155.

Guðmundur G. Hagalín. 1966. *Kristrún í Hamravík. Sögukorn um þá gömlu góðu konu*. Reykjavík: Almenna bókafélagið.

Guðni Jónsson. 1933. Sigrún í Hamravík. Sögukorn um þá gömlu, góðu konu. *Morgunblaðið* 8. desember, bls. 6.

Guðrún Kvaran. 2000. Hochdeutscher Einfluss auf das Isländische nach der Reformationszeit. Í: Hans-Peter Naumann og Silvia Müller (ritstj.). *Hochdeutsch in Skandinavien. Internationales Symposium, Zürich, 14.–16. Mai 1998*. Beiträge zur Nordischen Philologie 28, bls. 167–181. Tübingen og Basel: A. Franke Verlag.

Guðrún Kvaran. 2002. Auðnæm er ill danska. Fyrirlestur haldinn í málstofu málfræðinga föstudaginn 22. mars 2002. Vefslóð: *http://www.visindavefur.hi.is/malstofa_g-k.html*.

Gunnar Pálsson. [1782] 1982. *Lijtid Wngt Støfunar Barn*. Formáli eftir Gunnar Sveinsson. Íslenzk rit í frumgerð IV. Reykjavík: Iðunn.

Haugen, Einar. 1976. *The Scandinavian Languages. An Introduction to their History*. London: Faber and Faber.

Islandske Annaler indtil 1578. 1888. Udgivne for det norske historiske Kildeskriftfond ved Dr. Gustav Storm. Christiania: Grøndahl & Søns Bogtrykkeri.

Jahr, Ernst Håkon (ritstj.). 2000. *Språkkontakt – Innverknaden frå nedertysk på andre nordeuropeiska språk*. Skrift nr. 2 frå prosjektet Språkhistoriske prinsipp for lånord i nordiske språk. København: Nordisk ministerråd.

Jóhann Þórðarson. [1720] 1920. Brot úr líkræðu yfir Jóni biskupi Vídalín. Með athugasemdum eftir Hannes Þorsteinsson skjalavörð. *Prestafélagsritið – Tímarit fyrir kristindóms- og kirkjumál* 2:43–50.

Jón Helgason. 1929. *Málið á Nýja testamenti Odds Gottskálkssonar*. Safn fræðafjelagsins 7. Kaupmannahöfn: Hið íslenska fræðafjelag.

Jón Ólafsson úr Grunnavík. 1996. *Hagþenkir, JS 83 fol*. Þórunn Sigurðardóttir sá um útgáfuna og ritaði inngang. Reykjavík: Góðvinir Grunnavíkur-Jóns og Hagþenkir, félag höfunda fræðslurita og kennslugagna.

Jón Ólafsson úr Grunnavík. 1998. Animadversiones aliqvot & paulo fusior præsentis materiæ explanato. Hugleiðingar um sótt og dauða íslenskunnar. Birt hafa Gunnlaugur Ingólfsson og Svavar Sigmundsson. *Gripla* 10:137–154.

Jón Þorkelsson Vídalín. [1718] 1995. *Vídalínspostilla. Hússpostilla eður einfaldar predikanir yfir öll hátíða- og sunnudagaguðspjöll árið um kring*. Gunnar Kristjánsson og Mörður Árnason sáu um útgáfuna. Reykjavík: Mál og menning, Bókmenntafræðistofnun Háskóla Íslands.

Kalkar, Otto. 1881–1918. *Ordbog til det ældre danske Sprog (1300–1700)* 1–5. København: Carlsbergfondet.

Krogmann, Willy. 1970. Altsächsisch und Mittelniederdeutsch. Í: Ludwig Erich Schmitt (ritstj.). *Kurzer Grundriß der germanischen Philologie bis 1500*. Band 1, Sprachgeschichte, bls. 211–252. Berlin: De Gruyter.

Lbs. 95 8°. (Dagbækur Steingríms Jónssonar 1790–1795 á vísitasíuferðum með Hannesi biskupi Finnssyni.)

Lexicon Islandicum. [1683] 1999. Orðabók Guðmundar Andréssonar. Ný útgáfa. Gunnlaugur Ingólfsson og Jakob Benediktsson önnuðust útgáfuna. Orðfræðirit fyrri alda 4. Reykjavík: Orðabók Háskólans.

Magnús Ketilsson. [1802] 1948. *Stiftamtmenn og amtmenn á Íslandi 1750 til 1800*. Þorkell Jóhannesson bjó til prentunar. Sögurit 23. Reykjavík: Sögufélag.

Matthías Johannessen. 1985. Stríðið við herrann og höfuðskepnurnar. Um Guðmund Gíslason Hagalín. Í: Matthías Johannessen. *Bókmenntaþættir*, bls. 87–151. Reykjavík: Almenna bókafélagið.

Moberg, Lena. 1989. *Lågtyskt och svenskt i Stockholms medeltida tänkeböcker*. Acta Academiae Regiae Gustavi Adolphi 58. Uppsala.

Netútgáfan. Vefslóð: *http://www.snerpa.is/net/*.

Nielsen, Niels Åge. 1989. *Dansk Etymologisk Ordbog. Ordenes Historie*. (4. útg.). København: Gyldendal.

Nucleus Latinitatis … [1738] 1994. Ný útgáfa. Guðrún Kvaran og Friðrik Magnússon sáu um útgáfuna. Orðfræðirit fyrri alda 3. Reykjavík: Orðabók Háskólans.

ODS = *Ordbog over det danske Sprog* 1–28. 1918–1956. København: Det danske Sprog- og Litteraturselskab.

OH = Orðabók Háskóla Íslands. Söfn.

ONP = *Ordbog over det norrøne prosasprog* 1–. 1995–. København: Den arnamagnæanske Kommission.

Orðabók Jóns Ólafssonar úr Grunnavík. Orðaskrá. Vefslóð: *http://www.lexis.hi.is/JOL_skra.htm*.

Passíusálmar. Orðstöðulykill. Vefslóð: *http://www.lexis.hi.is/ordlyklar/salmar/salmar.htm*.

RM = Ritmálsskrá Orðabókar Háskólans.

Schiller, Karl og August Lübben. 1875–1881. *Mittelniederdeutsches Wörterbuch* 1–6. Münster og Bremen: Kühtmann.

Seip, Didrik Arup. 1934. Om vilkårene for nedertyskens innflytelse på nordisk. Í: Didrik Arup Seip. *Studier i norsk språkhistorie*, bls. 27–31. Oslo: Aschehoug.

Sigfús Blöndal. 1920–1924. *Íslensk-dönsk orðabók*. Reykjavík: Den danske og islandske Statskasse.

Simonsen, Marjun Arge. 2002. Orð við fremmandum atskoytum í føroyskum orðabókum. *Fróðskaparrit* 50:77–91.

Skautrup, Peter. 1947, 1953. *Det danske sprogs historie*. II, III. København: Det danske Sprog- og Litteraturselskab, Gyldendalske Boghandel – Nordisk Forlag.

Söderwall, K.F. 1884–1918. *Ordbok öfver svenska medeltids-språket* 1–2. Lund: Berlingska.

Söderwall, K.F., W. Åkerlund, K.G. Ljunggren og E. Wessén. 1925–1973. *Ordbok över svenska medeltids-språket*. *Supplement*. Lund: Berlingska.

Veturliði Óskarsson. 1997. Sem lágvært bárugjálfur við Íslands strönd. Um tökuorð af miðlágþýskum uppruna í íslensku. Í: Úlfar Bragason (ritstj.). *Íslensk málsaga og textafræði*, bls. 132–143. Rit Stofnunar Sigurðar Nordals 3. Reykjavík.
Veturliði Óskarsson. 2003. *Middelnedertyske låneord i islandsk diplomsprog frem til år 1500*. Bibliotheca Arnamagnæana 43. København: C.A. Reitzels Forlag.
Westergård-Nielsen, Chr. 1946. *Låneordene i det 16. århundredes trykte islandske litteratur*. Bibliotheca Arnamagnæana 6. København: Ejnar Munksgaard.

Abstract

This paper gives an overview of the history of the German prefix *an-* in Icelandic. One isolated example (*annáma*) shows up in a text from the middle of the 16th century, but judged by preserved texts, words of this type were not borrowed until after 1600. Around 60 *an-*words are to be found in the collection of Orðabók Háskólans (Institute of Lexicography) from ca. 1600 to the middle of the 20th century. Many of them are very rare, around 60% appearing only once or twice. Most of the words seem to be borrowed in the 18th and 19th centuries, but very few struck roots in the language, and the prefix has never been productive and has never been used in word formation in Icelandic. As a result of the language purism of the 19th and 20th centuries, most of the *an-*words disappeared together with many other loanwords of Danish-German origin, and literally none exist to-day.

Keywords:

loanwords, German influence, Danish influence, prefixes, language policy

Lykilorð:

tökuorð, þýsk áhrif, dönsk áhrif, forskeyti, málstefna

Veturliði G. Óskarsson
Kennaraháskóla Íslands
v/Stakkahlíð
IS-105 Reykjavík
veturosk@khi.is

Lars S. Vikør

Stóra orðabókin um íslenska málnotkun

Jón Hilmar Jónsson: *Stóra orðabókin um íslenska málnotkun.* Rafræn útgáfa á geisladiski fylgir. JPV útgáfa, Reykjavík 2005. ISBN 9979-791-26-8. xxx + 1562 sider

1 Innleiing

Jón Hilmar Jónsson (JHJ) har utvikla seg til nordisk meister på området fraseologisk leksikografi, og når eg begrensar meg til "nordisk", er det berre fordi eg har for liten detaljkunnskap om alt resten av Europa har å by på; det vil ikkje forundre meg om eg har teke for svakt i. Dei to monumentale ordbøkene *Orðastaður* (OS, 1994, 2. utg. 2001) og *Orðaheimur* (OH, 2002) (begge omtalte av Jóhannes Gísli Jónsson i *Orð og tunga* nr. 7) er i 2005 blitt slått saman til *Stóra orðabókin um íslenska málnotkun* (her forkorta SOÍM).

Det å slå saman to ulike (og så store) ordbøker til éi er i seg sjølv eit stort eksperiment, og det er interessant å sjå korleis det har lykkast. Eg skal i første omgang beskrive sjølve strukturen, eller strukturane, i SOÍM. Så skal eg kort prøve å plassere boka ordbokstypologisk – så vidt eg kan sjå, er dette ein nyskapning innanfor leksikografien. Deretter skal eg prøve nokre konkrete søk for å teste kor brukbar ordboka er i praksis for ein ikkje-islending som kan islandsk godt passivt, men dårleg aktivt. Det eg derimot ikkje kan gjere, er å vurdere den informasjonen ordboka gir. Eg må gå ut frå at vi her får eit hundre prosent kvalitetssikra og korrekt bilete av islandsk fraseologi og orddanning. Det følgjer ein cd med ordboka, men eg avgrensar meg her til å

Orð og tunga 9 (2007), 151–161.

vurdere sjølve boka. Ein presentasjon og ei vurdering av cd-en er gitt av Erla Hallsteinsdóttir (2006: 222-226), med kommentar av JHJ (2006: 234-235).

2 Tre strukturnivå

Eg deler inn strukturbeskrivinga i tre nivå, etter vanleg metaleksikografisk terminologi: megastruktur, makrostruktur og mikrostruktur.

Megastrukturen er sjølve hovudinndelinga av ordboka. Den begynner med forord og innleiing som gir ei grundig beskriving av oppbygging og søkjemoglegheiter, både i sjølve boka og i cd-en (supplert med skjematiske rettleiingar på innsida av permane). Sjølve ordboksdelen er tredelt. Først får vi sjølve *hovuddelen* i ordboka, "orðabókarlýsing": hovudlista over oppslagsord med ordartiklar. Den legg beslag på 842 sider. Deretter får vi *registerdelen*, "orða- og orðasambandaskrá", som omfattar alle ord som er brukte i dei ulike frasane og ordsambanda i ordartiklane der dei sjølve ikkje er oppslagsord, med tilvising til den artikkelen der dei står. Den dekkjer drygt 700 sider. Til slutt får vi så fire lister over "hugtakaheiti", på islandsk, dansk, engelsk og tysk – dei tre siste med islandske ekvivalentar. Dette er ein arv frå OH, som har ei slik liste på engelsk.

Makrostrukturen er glattalfabetisk i heile boka. Alfabetiseringa følgjer det nyare prinsippet som er basert på at kvar bokstav står for seg, altså *á* separat mellom *a* og *b* og tilsvarande ved alle dei aksentuerte vokalane. Tidlegare var det vanleg å integrere *a* og *á*, *e* og *é*, *i* og *í* osv. med kvarandre. Somme vil meine at den nye måten gjer det vanskelegare å finne fram. Det har med vane å gjere, sjølvsagt, men etter mi meining gir den nye måten eit rettare bilete av det islandske skriftspråket – der skiljet mellom aksentuerte og uaksentuerte vokalar er fundamentalt. Ikkje-islendingar blandar lett saman t.d. *i* og *í*, og det er kanskje ikkje så farleg når dei berre skal forstå lesen tekst, men viktig å unngå når dei sjølve skal produsere tekst. At ordbøkene trenar folk opp til å skilje det som faktisk er ulikt, ser eg på som eintydig positivt.

I hovuddelen av ordboka har vi to typar lemma (flettur) som er integrerte i éi alfabetrekkjefølgje: grunnord (leksem, flettiord) frå OS og begrep (hugtök) frå OH. Dei er skilde ved at begrepa står med versalar, store bokstavar, leksema med vanlege små bokstavar, begge

med halvfeit og utrykk. Når lesaren kjenner koden, vil han lett sjå kva som er leksem og kva som er begrep. Registerdelen inneheld naturleg nok berre leksem, også her med halvfeit, små bokstavar og utrykk. Ordboka er slik sett lett å ta seg fram i for å finne det ein er ute etter – på makronivået.

Men *mikrostrukturen* i hovuddelen er ikkje fullt så enkel, for det er svært ulike informasjonstypar som skal ha plass inne i desse artiklane. Eg beskriv dei to typane artiklar (ordartiklar frå OS og begrepsartiklar frå OH) kvar for seg.

Ordartiklane har altså eit grunnord (usamansett leksem) som lemma, av typen innhaldsord (substantiv, verb eller adjektiv). Strukturen er noko ulik for dei ulike ordklassene.

I *substantivartiklane* får vi: 1 reine setningsdøme på bruken av ordet, 2 frasar der formord og andre element kan skifte, mens andre element er faste, og døme på dei skiftande elementa er gitt i plogar (< >), 3 typiske adjektiv som står til substantivet (etter markeringa "ákvæði"), 4 samansetningar der lemmaet er førsteledd (eventuelt åtskilde etter ulike fuger), 5 samansetningar der lemmaet er sisteledd, og 6 eventuelle ordtak der lemmaet står sentralt. I tillegg til grunnord kan ein ha eige oppslag på forledd eller etterledd, nemleg der slike ledd skil seg formelt frå grunnordet. Har ordet fleire tydingar, blir dei skilde med tal i halvfeit og ofte eit synonym eller ei spesifisering.

Adjektivartiklane har mykje av det same: setningsdøme på typisk bruk, og kollokasjonar med eigne setningsdøme, med eventuelle markeringar for ulike tydingar av adjektivet eller frasen.

I *verbartiklane* ligg hovudvekta på å vise kombinasjonar med ulike adverbielle tillegg, ved sida av bruksdøme i setningsform her også. Ulike tydingar blir markerte ved å nemne ulike substantiv som relaterer seg til den aktuelle tydinga, ofte ei substantivering av det aktuelle verbet med markeringa "no". Men ikkje minst sentralt står dei ulike faste verbfrasane, som er førte opp i alfabetisk rekkjefølgje inne i verbartikkelen og deretter blir behandla som oppslag i seg sjølv (underoppslag eller sublemma).

Ingen av desse komponentane i ein ordartikkel er i og for seg obligatorisk, det er heilt avhengig av ordet sjølv kva for nokre som er med og som er omfangsrike eller ikkje. Men det heile går altså ut på å vise i størst mogleg detalj korleis ordet syntagmatisk agerer saman med andre ord – både over og innanfor ordgrenser.

Den andre typen artiklar går altså ut frå *begrep*. Det er 840 av dei – så vidt eg kan sjå direkte overførte frå OH. Begrepet blir sett inn i ulike kontekstar som viser ulike tydingar eller aspekt ved det, ofte ordna i ein hierarkisk struktur, og under kvar tyding blir det gitt ei større eller mindre mengd frasar som uttrykkjer denne tydinga. Frasane treng ikkje innehalde oppslagsordet; her er det det semantiske innhaldet i begrepet som står i fokus, ikkje ordet i seg sjølv. Også her blir fakultative ledd sette inn i plogar, ikkje minst personlege pronomen i terdje person, der kjønna blir eksplisitt jamstilte ved at det står ”hann, hana” eller ”honum, henni” i staden for det tradisjonelle ”e-n” og ”e-m”. Sist i artikkelen står det tilvisingar til andre, beslekta begrep som har eigne artiklar.

Registerdelen er enklare oppsett. Her blir oppslagsordet følgt av alle dei ulike frasane og ordsambanda der det er brukt, alfabetisert på første ordet i frasen, og tilvising blir gitt til den artikkelen eller dei artiklane i hovuddelen (av begge artikkeltypane) der frasen står.

3 Samanlikning: Kva er nytt i SOÍM?

Eg har samanlikna ein kort sekvens i dei tidlegare bøkene og SOÍM for å sjå om tilfanget er auka i særleg grad. Dessverre har eg berre tilgang til førsteutgåva av OS (1994), så eg kan ikkje vite om nokre av dei endringane eg nemner, har skjedd allereie i andreutgåva frå 2001.

Side 100 i OS dekkjer **eftirlit – egg**. Oppslagsorda er dei same i begge bøkene i denne sekvensen, når ein ser bort frå at ”**eftirsóknarvert** lo hvk” og ”**eftirtektarvert** lo hvk”, begge med eitt setningsdøme, er lagde til i SOÍM i tillegg til ”**eftirsóknarverður** lo” og ”**eftirtektarverður** lo”. I OS står begge setningsdøma under **-verður**. JHJ har altså i SOÍM oppretta ein ny lemmakategori, nemleg nøytrumsformer av adjektiv i upersonlege konstruksjonar (”það er —”, ”það þótti ekki —”) – noko som i norske ordbøker neppe ville ha komme på tale, sjølv om vi har same konstruksjonstypen. Til gjengjeld er **eftirmiðdagur** med i OS, men teken vekk i SOÍM, kanskje som ”dönskusletta”, for alt eg veit.

Når eg jamfører ein del ordartiklar på denne sida, ser tilfanget ganske identisk ut, det er (i alle fall her) ikkje lagt til mykje nytt stoff. (Elles i ordboka er det nok komme inn noko nytt materiale under ein del ord.) Ei fornying er det at kasusindikasjonar med t.d. ”e-s” er er-

statta av "hans, hennar", som eg så vidt har vore inne på ovanfor. Det mest synlege er at typografien er endra: Alle frasar, setningar, tydingar osv. kjem på ny line, mens dei i OS er klumpa saman i den løpande teksta etter lemmaet. Det gjer det langt enklare å finne fram i boka, det gir eit tiltalande oversiktleg og pedagogisk inntrykk. Men det betyr også at boka knapt kan lyftast av ein person med dårleg rygg.

Eg ser også på tilskota frå OH i den same sekvensen: **EFTIRLIT**, **EFTIRSJÁ** og **EFTIRTEKT**, og finn at artiklane her er heilt identiske og uforandra. Nokre stikkprøver elles i ordbøkene viser det same. Derimot er registerdelen sjølvsagt kraftig utvida sidan han no dekkjer heile SOÍM og ikkje berre OH-materialet.

4 Kva slags ordbok er SOÍM?

Å klassifisere denne ordboka typologisk, er ikkje så enkelt, og det gjeld først og fremst den delen som stammar frå OH. OH hadde undertittelen "Íslensk hugtakaorðabók med orða- og orðasambandaskrá". I *Nordisk leksikografisk ordbok* blir *begrepsordbok* (hugtakaorðabók) definert slik: "ordbok som grupperer sammen leksikalske enheter som hører til samme semantiske felt, og som i regelen har systematisk makrostruktur". Dette er den klassiske typen eksemplifisert i Rogets Thesaurus. Men OH svarer ikkje til denne beskrivinga. Den grupperer ikkje saman leksikalske einingar, men derimot frasar som kan ha noko å gjere med eit visst allment begrep. JHJ avgrensar seg til abstrakte begrep, som gjeld sinnsstemningar, haldningar, (språk)handlingar, relasjonar mellom menneske, eigenskapar, tilstandar. Den fysiske verda er stort sett fråverande. Og tilgangsstrukturen er alfabetisk, ikkje systematisk. Ein tesaurus av den typen Roget laga, er så vidt eg veit enno ikkje laga for islandsk. JHJ sjølv bruker da heller ikkje termen *tesaurus*, men kallar OH i ein norskspråkleg artikkel (2005: 228) for "en fraseologisk begrepsordbok".

Utgangspunktet for OH er ifølgje JHJ (2005) ei utilfredsheit over fokuseringa på enkeltordet i tradisjonelle ordbøker, og også i OS. Han ønskte ei ordbok der brukaren skulle kunne gå rett på den konteksten enkeltordet skulle fungere i, og ut frå det sjå på frasane som heilskapar og orda som byggjesteinar. Samtidig laga han ein alfabetisk registerdel over enkeltorda, slik eg kort har nemnt ovanfor.

Denne dobbelte strukturen frå OH er altså overført til SOÍM, som

dermed kan beskrivast som ei fraseologisk ordbok med tre inngangar av to typar, alle tre alfabetiske: ein begrepsbasert og to ordbaserte (i hovuddelen og registerdelen).

SOÍM gir ikkje opp tydingar; den berre listar opp frasar, setningar og samansetningar. Den einaste semantiske informasjonen som blir gitt, er korte indikasjonar på deltydingar av oppslagsorda i OS-artiklane, som dei enkelte ordsambanda blir grupperte under. I OH-delen blir frasane også grupperte inn under ulike overordna tydingar. Såleis skil SOÍM seg frå den andre store fraseologiske ordboka i Norden: *Svensk språkbruk. Ordbok över konstruktioner och fraser*, utgitt av salige Svenska språknämnden i 2003, der det blir gitt forklaringar på idiom og andre ordsamband der redaktørane meiner det trengst, og det er ikkje så sjeldan. Det same gjeld stilmerkingar av typen "litterært, alderdommeleg, høgtidleg, kvardagsleg, nedsetjande", som heller ikkje finst i SOÍM.

JHJ har vore ein flittig deltakar i det nordiske leksikografisamarbeidet, og har på ei rekkje nordiske konferansar lagt fram ideane bak dei storverka han har levert (Jón Hilmar Jónsson 1992, 1996, 2003, 2005). Eg viser til dette, og går ikkje vidare inn på intensjonane hans. Eg stiller i staden spørsmålet: Kven skal bruke ordboka, og til kva?

5 Døme på bruk: Nokre søk

Brukarane må i det minste vere folk som kan islandsk godt frå før, og som skal bruke språket aktivt eller forske i det. SOÍM er ei typisk produksjonsordbok – den som skal lese islandsk og treng opplysning om kva ord betyr, må gå til andre ordbøker, anten *Íslensk orðabók* (ÍO) eller ei tospråkleg ordbok. Men i ÍO finn ikkje brukaren så mange ord som JHJ har fått med, når det gjeld samansetningar. Såleis kan den som vil vite om ei gjennomsiktig samansetning som ikkje finst i ÍO, likevel kan brukast på islandsk, slå opp under eitt av ledda i ordet i SOÍM og leite i listene der. Rett nok er samansetningsrekkjene ikkje ordna alfabetisk, men etter tyding, så vidt eg kan sjå, og det stiller store krav til brukaren, for her er ikkje strukturen så godt markert alltid (det kjem eg tilbake til nedanfor).

No skal eg teste SOÍM ved å gjere nokre søk der (i papirordboka). Dei meldingane av OS, OH og SOÍM som eg har sett, er skrivne frå ein islandsk brukarsynsstad (Anna Helga Hannesdóttir 2005, Jóhannes

Gísli Jónsson 2005 og Erla Hallsteinsdóttir 2005 – den islandsk-kompetente nordmannen Erik Simensen (1995) er eit unntak). Eg skal supplere desse ved å stille meg på den islandsk-interesserte utlendings standpunkt. Det er ikkje mange av oss, men listene over "hugtakaheiti" på framande språk bak i boka viser jo at JHJ har tenkt på oss også. Eg vil formulere nokre tilfeldige "utlendingspørsmål" og sjå kor langt SOÍM hjelper meg til å finne svar på dei – jamført med ÍO.

Først har eg late meg inspirere av sjølve orda *orðastaður* og *orðaheimur*. Kva betyr dei eigentleg? JHJ har ikkje forklart dette nokon stad eg har sett, så det er kanskje opplagt for ein islending. Men da må det vel stå i ordbøkene. Eg slår opp.

I ÍO står berre **orðastaður**, forklart som "viðræða, deila", med dømet "**segja e-ð í orðastað e-s** segja e-ð i stað annars, leggja e-m orð i munn". Eit søk på forleddsartikkelen **orða-** i SOÍMs hovuddel hjelper ikkje så mykje. Artikkelen er delt i fem tydingsgrupper, og *orðastaður* står i gruppe 5 saman med *orðaskipti* og ein del andre ord etter tydingsmarkøren *deila*. Men **orðastaður** er òg eige oppslag, forklart som "viðræða" med ordsambandet "eiga orðastað við <hann, hana>", og under ei tyding 2 i "<segja þetta> í orðastað <hans, hennar>". Dessutan finn vi dei same uttrykka i begrepsartiklane SAMTAL/ORÐASKIPTI og UMMÆLI. Dette kan tyde på at ordet er lite brukeleg utanfor desse ganske faste ordsambanda.

Orðaheimur finn eg ingen stad, verken i ÍO eller i SOÍM (etter søk både på sjølve ordet, forleddet *orða-* og etterleddet *-heimur*). Det kan tyde på at JHJ har laga ordet, og er så samvitsfull at han ikkje vil bruke belegg av seg sjølv som kjelde. I innleiinga til OH finn eg heller ingen stad nokon kommentar til sjølve ordet som er namn på ordboka. Er ordet da sjølvforklarande for ein islending trass i at det eigentleg ikkje "finst"? Det verkar slik, for Anna Helga Hannesdóttir (2005) går rett på det i tittelen til meldinga si av OH: "Nya vägar in i ordens värld" (alludert til også av Erla Hallsteinsdóttir (2006)). Dermed er det vel leksikalisert og må komme inn i neste utgåve?

Når *Orðastaður* og *Orðaheimur* er slått saman, er det da *heimstaður orða* vi har framfor oss? Det kan passe, for etter JHJs filosofi er ordas heimstad nettopp frasane, ikkje lemmalistene i ei definisjonsordbok. Eg slår opp i ÍO, og finn ikkje noko ord *heimstaður*. Så prøver eg i SOÍM, både under *heim-*, *heima-* og *heims-*, utan hell. Dette ordet finst openbert ikkje på islandsk. *Heimstad* eller *hjemsted* på norsk betyr, et-

ter Hróbjartur Einarssons *Norsk-islandsk ordbok*, *heimaslóðir* eller *heimastöðvar.* Det første av desse to orda står som lemma i ÍO, men ikkje det siste; det er derimot oppført som synonym under *heimaslóðir*. I SOÍM finn vi eintalsforma *heimaslóð* under *heima-* saman med ein del andre ganske synonyme ord.

Eg prøver eit nytt søk: Ein kjend islandsk oppsong av det meir militante slaget sluttar med lina "*stríðum, vinnum vorri þjóð*". Eg spør meg kva *vinna* tyder her ('sigre' eller 'arbeide'?), og korfor dativen "*vorri þjóð*"? Så ser eg etter i ÍO, der begge tydingane finst, men med 'arbeide' som den primære, og eg finn ingen absolutt ("naken") dativkonstruksjon (berre typen *vinna e-m e-ð*). Så går eg til SOÍM. Her finn eg ulike tydingar, markerte med formelt eller semantisk relaterte substantiv ("no. vinna; no. vinnsla; sigur"). Men den "nakne" dativkonstruksjonen som eg søkjer, finn eg ikkje. Konklusjonen av det må vel vere at denne konstruksjonen ikkje er aktuell og gangbar i dag. Men kanskje han var det i eldre språk. Det fortel ikkje SOÍM noko om.

Men denne historia er litt lengre. Eg hadde berre brukt 1983-utgåva av ÍO. I siste fase av artikkelskrivinga får eg tak i 2002-utgåva og ser etter der, og der finn eg "**vinna e-m** vera í þjónustu e-s". Er det dette som er meint i songen? I SOÍM står det i alle fall ikkje.

Vi kan altså bruke ordbøker til å finne ut at eit ord ikkje finst (som **heimstaður*). Men vi kan altså òg finne ord og konstruksjonar i autentiske islandske tekster som (fleire av) ordbøkene ikkje nemner. SOÍM inneheld openbert det gangbare ord- og frasetilfanget i det moderne islandske allmennspråket, men stilmarkeringar og andre markeringar som viser bruksstatusen orda har i språket, finst ikkje, som eg alt har nemnt. Igjen viser det seg at brukaren må ha ei velutvikla islandsk språkkjensle, dvs. morsmålskjensle eller nesten det, for å få full nytte av ordboka.

Enda eit nytt søk: På islandsk heiter det *fremja sjálfsmorð*, på dansk og norsk bokmål *begå selvmord*, engelsk *commit suicide* og nederlandsk *zelfmoord plegen*. *Begå*, *commit* og *plegen* på dei respektive språka har ein klang av "gjere noko ulovleg eller normstridig", men i alle fall på norsk er *fremme* eit positivt ord som betyr 'arbeide for', 'styrkje' o.l. Er *fremja* på islandsk negativt eller nøytralt? I ÍO finn eg ikkje ut mykje om det, der står berre synonyma *iðka, drýgja* og eit par andre, rett nok med døma *fremja glæp, illvirki*. Eg slår opp i SOÍM og søkjer eit svar der. Og der finn eg det: *fremja* er inndelt i to tydingar med objekttype

angitt: "1 vont verk", med døme frå *lögbrot* til *ranglæti*; også *sjálfsmorð* er med. Den andre tydinga, "2 athöfn", har færre døme. Dermed ser eg at *fremja* på norsk i utgangspunktet bør omsetjast med *begå* heller enn med *fremme* (med atterhald om at konteksten i eit konkret tilfelle kan krevje noko anna).

På denne måten kan SOÍM vere til hjelp når ein vil finne det rette ordet i omsetjing, men ein må ha konsultert ei definisjonsordbok først. Som eg alt har antyda eit par gonger, må ein ha solide bakgrunnskunnskapar om ein skal ha utbytte av SOÍM, for mykje av den informasjonen ordboka gir, er implisitt. Det gjeld ordninga av samansetningar, for eksempel. I motsetning til ÍO, gir SOÍM kunnskap om at forleddet *heim-* går på "rørsle heimover" (*heim-ferð*, *heim-för*, *heim-koma* osv.), *heima-* går på det vi i andre germanske språk mest tenkjer på med *heim, hem, hjem, home*, mens *heims-* knyter seg til substantivet *heimur* 'verd'. Men det blir ikkje sagt, berre implisert.

Samansetningsrekkjene i SOÍM er, som nemnt før, ikkje alfabetisk ordna, men inndelte etter tydingsgruppe (og heller ikkje alfabetiske innanfor tydingsgruppene). Overgangane mellom ulike tydingsnyansar er markerte med semikolon, mens orda elles er åtskilde med komma. Som under **heima-**: Vi har ei hovudinndeling i ordklasser: først substantiv, så adjektiv. Og så ber det laus under "NO": "heimahús, -hagar, -land, -byggð, -sveit" og 24 ord til innan vi får eit semikolon, og så går det vidare: "heima-vinna, -verk, -iðja," atten nye ord, nytt semikolon, så: "heima-stíll, -ritgerð, -lexia" – og så vidare. Eg sparer litt plass her ved å ikkje gjenta forleddet; i ordboka står alle orda fullt utskrivne med ledd-deling markert. Eg ser logikken i systemet, men det er ikkje lett for brukaren å leite etter ei spesiell samansetning i dei lange rekkjene.

Eg gjer òg eit søk på eit begrep, og da vel eg *kjærleik*. Som skandinav slår eg opp i den danske begrepslista på *kærlighed*, og ventar å finne *ást*, for såpass islandsk kan eg. Men eg finn to andre ord: *væntumþykja* og *ástarhugur*. Eg begrensar meg til det sistnemnde.

Under ÁSTARHUGUR finn eg ei lang rekkje synonyme uttrykk med tyding 'bli forelska i, bli glad i', eller i setningsform "dei er / vart glade i kvarandre" – grupperte etter verb, med verbalfrasar og setningsdøme om kvarandre: "fella ástarhug til <hans, hennar>, fella ást til <hans, hennar> … hugir þeirra falla/féllu saman, þau leggja/lögðu hugi saman, leggja ástarhug á <hann, hana>" osv. til "þau unnast hug-

ástum". Deretter får vi nokre døme på tre særtydingar: "með vísun til augnatillits: renna hýru auga til hans, hennar … með áherslu á aðdáun: dá <hann, hana> … með áherslu á þrá, girnd … " – og så vidare.

Dette gir god innsikt i strukturen til dei islandske uttrykksmåtane, og kan òg vere praktisk nyttig for den som skal omsetje til islandsk eller sjølv vil skrive på språket og uttrykkje visse ting. Men sjølvsagt krev også desse oppslaga høg kompetanse hos brukaren når han eller ho skal vurdere ordsambanda og finne fram til det som eignar seg best til det behovet brukaren har akkurat da.

6 Konklusjon

Konklusjonen må bli at SOÍM er eit storverk som det knapt finst maken til, og særleg for eit så lite språksamfunn. Som bruksbok er det først og fremst forskarar og profesjonelle skribentar og omsetjarar som vil ha nytte av ho. Erla Hallsteinsdóttir (2006: 11) meiner at "det er tvivlsomt, om brugere med andre modersmål overhovedet vil kunne bruge den". Eg vil ikkje seie det så sterkt, etter det utvalet av søk eg har presentert ovanfor. Men Jón Hilmar (2006: 230) avfeier likevel dette synspunktet litt for raskt og enkelt i svarreplikken sin. For ein utlending (med avanserte kunnskapar i islandsk) tener SOÍM primært som eit supplement til den meir konvensjonelle definisjonsordboka, for vi har behov for eksplisitte definisjonar og forklaringar og sosialstilistiske bruksmarkeringar som morsmålsbrukarar av islandsk ikkje treng på same måten. Men for dei som driv jamførande studiar mellom islandsk og skandinaviske språk eller andre germanske språk innanfor idiomatikk og fraseologi, er SOÍM ei sann gullgruve.

Kjelder og litteratur

Anna Helga Hannesdóttir. 2005. Nya vägar in i ordens värld. *LexicoNordica 12*: 199–211 Bergenholtz, Henning o.a. 1997. *Nordisk leksikografisk ordbok.* Oslo: Universitetsforlaget.

Clausén, Ulla o.a. 2003. *Svensk språkbruk. Ordbok över konstruktioner och fraser.* Utgitt av Svenska språknämnden. Stockholm: Norstedts.

Erla Hallsteinsdóttir. 2006: I ordenes store verden. *LexicoNordica 13*: 209–228.

Íslensk orðabók handa skólum og almenningi. 1983. (2. utg.) Red. Árni Böðvarsson. Reykjavík: Menningarsjóður.

Íslensk orðabók. 2002. (3. utg.). Red. Mörður Árnason o.a. Reykjavík: Edda.

Jóhannes Gísli Jónsson. 2005. Orðastaður og Orðaheimur. *Orð og tunga* 7:121–129.

Jón Hilmar Jónsson. 1992. Fra en passiv til en aktiv ordbok. Det kombinatoriske aspektet i fokus. I: R. V. Fjeld (red.). *Nordiske studier i leksikografi. Rapport fra Konferanse om leksikografi i Norden 28.–31. mai 1991*, side 88–104. Skrifter utgitt av Nordisk forening for leksikografi 1.

Jón Hilmar Jónsson. 1994. *Orðastaður. Orðabók um íslenska málnotkun*. Reykjavík: Mál og menning.

Jón Hilmar Jónsson. 1996. Nøkler til ordforrådet. Om lemmafunksjon, struktur og informasjonstyper i en ny kombinatorisk ordbok over islandsk. I: Ásta Svavarsdóttir o.a. (red.). *Nordiske studier i leksikografi 3. Rapport fra Konferanse om leksikografi i Norden. Reykjavík 7.–10. juni 1995,* side 245–254. Skrifter utgitt av Nordisk forening for leksikografi 3.

Jón Hilmar Jónsson. 2002. *Orðaheimur. Íslensk hugtakaorðabók með orða- og orðasambandaskrá.* Reykjavík: JPV.

Jón Hilmar Jónsson. 2003. Fraseologien i forgrunnen – fraseologisk register som ledd i ordbokens tilgangsstruktur. I: Zakaris S. Hansen og Anfinnur Johansen (red.). *Nordiske studier i leksikografi 6. Rapport fra Konference om leksikografi i Norden. Tórshavn 21.–25. august 2001*, side 151–167. Skrifter udgivet af Nordisk forening for leksikografi 7.

Jón Hilmar Jónsson. 2005. Orðaheimur – en fraseologisk begrepsordbok. I: R. V. Fjeld og D. Worren (red.). *Nordiske studiar i leksikografi 7. Rapport frå Konferanse om leksikografi i Norden. Volda. 20.–24. mai 2003*, side 228–236. Skrifter utgjevne av Nordisk foreining for leksikografi 8.

Jón Hilmar Jónsson. 2006. Kommentar til anmeldelsen *I ordenes store verden*. *LexicoNordica 13*: 229–236.

Norsk-Islandsk ordbok. Norsk-Íslensk orðabók. 1987. Red. Hróbjartur Einarsson. Oslo: Universitetsforlaget.

Simensen, Erik. 1995. [Melding av] Jón Hilmar Jónsson: *Orðastaður. Orðabók um íslenska málnotkun*. *LexicoNordica 2*: 281–286

Orðabókar- og rannsóknarverkefni

Tungutækniverkefni sem Orðabók Háskólans tekur þátt í

Orðabók Háskólans (nú orðfræðisvið Stofnunar Árna Magnússonar í íslenskum fræðum) er einn þriggja aðila að *Tungutæknisetri*, ásamt Málvísindastofnun Háskóla Íslands og tækni- 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.

Tungutæknisetur gekkst fyrir ráðstefnunni *Íslensk tungutækni* 2006 hinn 26. maí og voru þar fluttir sjö fyrirlestrar um fjölbreytt tungutækniverkefni. Ráðstefnan var haldin í Háskólanum í Reykjavík og sóttu hana um 50 manns. Í desember hófst fyrirlestraröð Tungutækniseturs 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 verkefninu (verkefnisstjóri Bente Maegaard). Vinnan í verkefninu 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 verkefninu 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 forverkefninu *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 verkefninu *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 samkenni 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 sagnaflettur, 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 sagnaflettunum fæst víða glöggt 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 tilteknar formgerðir (án þess að hugsað sé til sérstakrar sagnar).

Samheitatenging nafnorða er komin nokkuð á veg með flokkun orðasambanda sem tengjast lýsingarorða- og sagnaflettum. 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 nafnorðaflétturnar 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 flettiorð auk 25.000 dæmasetninga og orðasambanda. Þá var ákveðið að bæta við 3000 flettiorðum 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 fyrirframgefinni 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

Bókafregnir

Halldóra Jónsdóttir [ritstj.] *Íslensk-dönsk, dönsk-íslensk vasaorðabók.* Önnur útgáfa. Mál og menning, Reykjavík 2006. ISBN 9979-3-2774-X/987997937745. 806 bls.

Í 8. hefti *Orðs og tungu* var sagt frá fyrstu útgáfu þessarar bókar og vísast til þeirrar umsagnar um bókina almennt. Í formála annarrar útgáfu getur Orðabókaritstjórn

Eddu þess að orðaforði íslensk-danska hlutans hafi verið aukinn talsvert og var eins og í fyrri útgáfunni tekið mið af íslensku nútímamáli. Sérstaklega var hugað að fjölgun dæma um málnotkun. Við dansk-íslenska hlutann var bætt á þriðja hundrað uppflettiorðum. Bókinni er ætlað að ná til Íslendinga, sem ferðast til Danmerkur, námsmanna og þeirra sem nota tungumál við dagleg störf.

Ingrid Markan [ritstj.], Jón Skaftason, Pétur Knútsson Ridgewell (hljóðritun). Ensk-íslenska orðabókin. JPV útgáfa, Reykjavík 2006. ISBN 9979-791-87-3. xxxi, 849 bls.

Ensk-íslenska orðabókin er aukin og endurbætt útgáfa á *Ensk-íslenskri skólaorðabók* sem gefin var út fyrst 1986. Bókaforlagið Örn og Örlygur hafði þá útgáfu með hendi. Orðabókin hefur nú verið aukin verulega. Í bókinni eru hátt í 40.000 uppflettiorð. Í formála kemur fram að talsvert á þriðja þúsund nýrra orða og merkinga hefur verið bætt við eldri gerðina auk lagfæringa á eldri flettum. Þar er einkum um að ræða umorðun á skýringartextum og viðbætur notkunardæma og nýyrða sem náð hafa að festast í málinu á liðnum tuttugu árum.

Bókinni fylgja ítarlegar notkunarleiðbeiningar, leiðbeingar um framburð og ýmsar skrár sem létta notkun bókarinnar. Að endurbættri útgáfu unnu Ingrid Markan, sem var í ritstjórn, Jón Skaftason og Pétur Knútsson Ridgewell sem ritstýrði hljóðritun bókarinnar.

Stanisław J. Bartoszek, Paweł Bartoszek, Marta Ewa Bartoszek. *Íslensk-pólsk, pólsk-íslensk skólaorðabók*. sjb, Reykjavík 2006. ISBN 9979-70-219-2. 459 bls.

Nýlega kom á markað Íslensk-pólsk, pólsk-íslensk skólaorðabók sem ætlað er að vera lykill að algengustu orðum í pólsku og íslensku. Í formála kemur fram að fjöldi uppsláttarorða í báðum hlutum er um 25.000 og voru orðin valin með hliðsjón af íslenskum og pólskum orðtíðnibókum. Við val flettiorða var hugað að því að bókin kæmi nemendum, skólum og almennum notanda að sem mestu gagni og var því áhersla aðallega lögð á orð og orðasambönd úr daglegu máli en einnig á orðaforða sem tengist skólastarfi. Aftan við sjálfa orðabókina er leiðbeiningar fyrir Íslendinga um pólskan framburð og sams konar leiðbeiningar eru á pólsku um íslenskan framburð ásamt stuttum leiðbeingum um íslenskar beygingar.

Tuomas Järvelä. *Suomi-islanti-sanakirkja. Finnsk-íslensk orðabók*. Helsinki University Press, Helsinki 2006. ISBN 951-570-651-3. 486 bls.

Þessi nýja finnsk-íslenska orðabók hefur að geyma um 36 þúsund flettur af mörgum sviðum. Markhóparnir eru námsmenn og þýðendur. Finnsku nafnorðin eru samkvæmt notkun annaðhvort sett fram í nefnifalli eintölu eða fleirtölu en við íslensku þýðingarnar er gefið nefnifall og eignarfall eintölu og nefnifall fleirtölu ef um grunnorð er að ræða. Ekki er getið um beygingar samsettra orða. Við sagnir er gefinn nafnháttur og 1. persóna eintala í þátíð framsöguháttar í germynd. Sérstakur kafli er í bókinni um beygingu sterkra og óreglulegra sagna (bls. 480–486) en ekki annars er hvorki fjallað um íslenska málfræði né framburð.