

Kristín Bjarnadóttir

Hvert á að sækja orðaforðann í orðabók?

1 Inngangur

Einn af grundvallarþáttum við gerð nýrrar orðabókar er að velja og takmarka flettiorðaforðann og vegna þessa má halda því fram að flettulistinn sé kjölfestan í hefðbundinni orðabók. Vegna tækni-breytinga hillir nú undir nýjar aðferðir til að nálgast orðaforðann og bylting hefur þegar orðið í aðgengi að orðmyndum í textum og greiningu þeirra. Í þessari grein er fjallað um mismunandi aðferðir við söfnun efnis í flettulistann. Þá er átt við orðalistann einan, án tillits til merkingargreiningar eða samhengis orða, og án tillits til útgáfuforms orðabókarinnar. Hér er eingöngu fjallað um aðferðir til þess að finna sjálft hráefnið, orð og orðmyndir, í þeim gögnum sem nú eru tiltæk.

Til skamms tíma var nánast aðeins ein leið fær við að afmarka orðaforðann í orðabók um íslenskt nútímamál: Að nýta lista sem fyrirrennararnir höfðu sett saman. Af útgáfusögu *Íslenskrar orðabókar* (Árni Böðvarsson 1963, 1983, Möður Árnason 2000 o.áfr.; hér eftir Ío) er ljóst að þar bar hæst Blöndalsorðabók (sjá Guðrún Kvaran 1997) þar til Ío kom út en eftir það hefur síðara verkið verið meginstoð í íslenskri orðabókagerð. Í lýsingu Marðar Árnasonar ritstjóra 3. útg. Ío (2000) á fyrirhuguðum endurbótum á verkinu kemur skýrt fram það hlutverk orðabókar að draga saman efni úr öðrum áreiðanlegum heimildum og þar er fyrst að telja orðabækur:

... en nú sem fyrr hlýtur slík almenn íslensk orðabók að gegna því hlutverki að draga á handhægan hátt saman fróðleik úr sérhæfðum

orðabókum, orðasöfnum, handbókum og fræðigögnum og vera sem almennast hjálpargagn fyrir sem allra flesta þátttakendur í íslensku málsamfélagi, vera „til gagns flestum þeim, sem þurfa skýringa á íslenskum orðum í almennu lesefni, gömlu eða nýju“ eins og það er orðað í formála fyrstu útgáfu (bls. V).

(Mörður Árnason. 1998:3)

Hér er ekki minnst á að orðabók ætti að gefa hugmynd um daglegt mál á vinnslutíma orðabókarinnar enda voru tæknilegar forsendur til þess ekki fyrir hendi hér á landi árið 1998 þegar Mörður skrifaði grein sína. Þar var enn byggt á þeim grunni sem notaður var við fyrstu útgáfu Ío, þ.e. Blöndalsorðabók og söfn Orðabókar Háskólans (OH) sem sett var á stofn tveimur áratugum fyrr.

Orðabókarmönnum sem vinna að útgáfu orðabóka gefst yfirleitt takmarkað ráðrúm til að leggjast í orðtöku; það er einfaldlega of dýrt þegar horfst er í augu við tiltekinn útgáfudag sem venjulega er alltof skammt undan. Gildir þá einu hvort orðtakan er hefðbundin, þ.e. byggist á lestri texta, eða vélræn, þ.e. byggist á söfnun orðmynda úr rafrænum (en ómörkuðum) textum; báðar aðferðirnar eru seinlegar og matið á því hvað á heima í tiltekinni orðabók er erfitt.

Nýjar aðferðir í orðabókargerð byggjast á málheildum, þ.e. völdum textasöfnum sem ætlað er að gefa mynd af málinu eins og það er á hverjum tíma. Þar er m.a. stuðst við tíðnitölur til að meta hvort orð eiga heima í tilteknu orðabókarverki. Fyrsta enska orðabókin sem studdist við tíðnikannanir og málheild, *Cobuild*, kom út árið 1987.

Þar var markmiðið að birta enska tungu í orðabók, eins og hún væri raunverulega notuð (sjá bókarkápu: „... based on a detailed analysis of how today's English is really used“) en ekki aðeins eins og orðabókarhefðin segði til um:

For the first time, a dictionary has been compiled by the thorough examination of a representative group of English texts, spoken and written, running to many millions of words.

(Sinclair1987:XV)

Á Íslandi hefur slíkt orðabókarverk ekki enn litið dagsins ljós en reyndar er óljóst að nokkur málheild nái því að vera **staðgeng** (e. *representative*) þannig að sönn mynd af tungumálinu birtist í henni en það er forsenda þess að orðabók með þessu háleita markmiði verði gerð (Aðalsteinn Eyþórsson 2005:14). Samt sem áður er **jafnvæg** málheild (e. *balanced corpus*) ný gerð af heimild fyrir orðabókarmenn þar sem komast má nær orðaforða og málnotkun en hingað til hefur verið

hægt. Með tilurð *Markaðrar íslenskrar málheildar* (hér eftir MÍM) opnast því nýir möguleikar í íslenskri orðabókargerð (Sigrún Helgadóttir o.fl. 2012). Sama á reyndar við um önnur rafræn textasöfn, t.d. *Íslenskt textasafn* Stofnunar Árna Magnússonar í íslenskum fræðum¹ og *Íslenskan orðasjóð*² hjá Háskólanum í Leipzig (Erla Hallsteinsdóttir 2007, Quasthoff o.fl. 2012), þó að slík söfn séu ekki eins aðgengileg til úrvinnslu og MÍM.

Í þessari grein er fjallað um þá möguleika sem felast í þessum nýju gagnauppsprettum og gamlar og nýjar aðferðir bornar saman. Niðurstaðan er að á þessu stigi verður að nota allt efni sem tiltækt er til þess að notendur eigi möguleika á að finna sem flest af því sem þeir leita að. Íslenskir orðabókarnotendur eru fremur íhaldssamir í orðabókarnotkun og leita einna helst að merkingu orða og viðeigandi notkun þeirra (Ásta Svavarsdóttir 2008), en orðabókarmenn vita að væntingar notenda til orðabóka eru langt umfram það sem nokkur leið er að verða við. Þar þarf ekki að nefna annað en þá kröfu margra notenda að öll orð sem eiga sér viðurkenndan og fastan sess í málinu finnist í **orðabókinni**, án þess að tekið sé fram til hvaða orðabókar er vísað. Orð sem ekki finnast í „orðabókinni“ eru þá talin óbrúkgleg og óviðurkennd. Þessari kröfu geta orðabókarmenn seint sinnt að fullu. Engin orðabók verður nokkurn tíma tæmandi heimild um málið og hver á að taka sér það alræðisvald að ákveða hvaða orð mega lifa í málinu og hver ekki? Sem betur fer hafa orðabækur heldur ekki slíkt heljarafli í reynd en verkefnið er samt sem áður að sýna orðaforðann í málinu sem best, eins og hann er á hverjum tíma, og þess vegna er vert að skoða ný gögn og nýjar aðferðir.

Efnisskipan er þessi: Í 2. kafla er fjallað um hefðbundið íslenskt orðabókarefni, útgefnar orðabækur og önnur orðfræðileg gögn til orðabókargerðar sem unnin hafa verið með hefðbundnum hætti, þ.e. með hefðbundinni orðtöku. Þar er líka sagt frá Norræna verkefninu svokallaða þar sem gerð var tilraun til að sameina sem mest af þessu efni í rafrænan íslenskan orðabókarstofn. Í 3. kafla er fjallað um nýjar gagnalindir til orðabókargerðar, þ.e. rafræna texta, markaða og ómarkaða. Í 4. kafla er fjallað stuttlega um vélræna orðtöku. Í loka-kaflanum er fjallað um samspil gamalla og nýrra aðferða en niðurstað-

1 Í *Íslensku textasafni* eru gamlir og nýir textar, u.þ.b. 67 milljónir lesmálsorða. Niðurstöður leitar birtast sem orðstöðulykill, sjá http://arnastofnun.is/page/arnastofnun_gagnasafn_textasafn.

2 Í *Íslenskum orðasjóði* er nú u.þ.b. hálfur milljarður lesmálsorða af vefnum; sjá http://wortschatz.uni-leipzig.de/ws_ice/.

an er sú að slíkt verklag sé árangursríkasta leiðin við þær aðstæður sem nú eru.

2 Hefðbundið íslenskt orðabókarefni

Hefðbundið efni í íslenskri orðabókargerð er það samsafn sem orðabókarmenn taka í arf frá forverum sínum. Þar er aðallega byggt á útgefnum orðabókum, seðlasöfnum OH og orðalistum af ýmsu tagi úr ýmsum verkefnum. Orðabækur eru dýr langtímaverkefni og þess vegna verða orðabókarmenn að nýta allt tiltækilegt efni á sem allra ódýrasta máta. Þess er ekki kostur í litlu málsamfélagi að frumvinna flettiorðalista í hverja orðabók og þess vegna gengur hver orðabókarmaður í smíðu hjá öðrum. Fljótlegast er að vinna úr tölvutækum orðalistum en þar með er ekki nema hálf sagan sögð. Flettugildi í orðabók er ekki bundið við uppflettiorðin ein (sbr. Jón Hilmar Jónsson 2013), en hér er umfjöllunarefnið þó aðeins flettiorðalistinn sjálfur og hvar og hvernig best eða auðveldast er að nálgast efni í hann. Það efni sem orðabókarmenn taka í arf frá fyrirrennrum sínum skiptist í meginatriðum í tvennt: Í útgefnar orðabækur og gagnasöfn, upp-runalega seðlasöfn, sem ætluð eru til orðabókargerðar.

2.1 Útgefnar orðabækur

Eins og fyrr segir eru meginstoðir í íslenskri orðabókargerð *Íslensk-dönsk orðabók* Sigfúsar Blöndals (1920–1924) og *Íslensk orðabók* (Ío 1963, 1983 o.áfr.) og byggir síðara verkið að miklu leyti á því fyrra. Heimildir að Blöndalsorðabók eru m.a. fornmálsorðabækur, t.d. orðabók Richards Cleasbys og Guðbrands Vigfússonar (Guðbrandur Vigfússon 1874–1876), og orðabók Björns Halldórssonar (1814) en til hennar var safnað í lok 18. aldar. Orðabókarmenn leita síðan fanga eins víða og nokkur kostur er og hér skal nefna *Íslensk-enska orðabók* eftir Sverri Hólmarsson, Christopher Sanders og John Tucker sem dæmi, en í formála þar eru nefndar fjölmargar tvímálaorðabækur sem leitað var í um orðaforða (Sverrir Hólmarsson o.fl. 1989). Fyrsta heimild sem nefnd er um orðaforðann er samt sem áður Ío: „Orðaforða höfum við meðal annars heyjað okkur úr Orðabók Menningarsjóðs, 2. útg. 1983, . . .“ (s.st.).³ Aðgangur að efninu úr Ío gjörbreyttist eftir að

3 Að auki er nefnt í formálanum að orðaforðinn hafi verið borinn saman við tíðnilista

gagnagrunnurinn Lexa varð til hjá Máli og menningu (hér eftir MM) fyrir tölvuútgáfunna árið 2000 og hefur sá gagnagrunnur verið meginstoðin í orðaforðanum í þeim verkum sem síðan hafa verið unnin hjá MM og síðar Forlaginu/JPV.

Orðalistinn úr Ío er ekki að öllu leyti heppilegur í tvímálaorðabók, bæði vegna þess að orðaforðinn þar er alltof mikill fyrir slíkt verk og einnig vegna þess að kröfurnar í nýju verki eru ekki þær sömu og í verki sem er að stofni til frá 1963 og var frá upphafi ætlað að þjóna sem alhliða orðabók fyrir íslenska notendur. Þarna koma fram í hnotskurn þeir gallar sem eru við að nota orðalista úr eldri orðabókarverkum:

- Skilgreining á markhópi og yfirgrípi á ekki við í nýju verki. Þetta á t.d. við í tvímálaorðabókum þar sem markmálið setur viðfangsmálinu skorður, m.a. vegna þess að þýðing getur verið nauðsynleg markmálsins vegna, þó svo að ekki sé þörf á merkingarlýsingu í viðfangsmálinu.
- Hluta orðaforðans vantaði til skamms tíma í orðabækur. Þetta á sérstaklega við um orð úr daglegu máli sem sjaldnast komast á bækur, t.d. heimilisorð ýmiss konar.
- Tilhneiging er til að safna sjaldgæfum orðum, á kostnað þess sem algengt er. Eitt af einkennum Ritmálssafns OH er t.d. betri yfirsýn um sjaldgæf orð og sérkennileg en algengt mál. Þar eru t.d. fjögur dæmi um orðið *eiginmaður* en níu dæmi um orðið *ektamaður*.
- Orðaforðinn úreldist og því þarf bæði að bæta við og grísa. Ýmis orð um tölvur eru þar nærtæk dæmi, t.d. *gataspjald* og *nálaprentari* sem nú eru úrelt, og *blogg* og *spjaldtölva* sem eru á hvers manns vörum.
- Villur og draugorð geta átt sér langa ævi. Orðið *nösgæs* 'e.k. gæs eða ef til vill önd' (Ío 1963, 1983) lifði t.d. lengi vel góðu lífi í orðabókum sem heiti á önd (eða gæs), meira að segja með latneska heitinu *anser nasutum*, enda þótt orðið sé úr gátu í Heiðreks sögu og merki andardráttur (Aðalsteinn Eyþórsson 1998).

Öll þessi atriði verða til þess að gaumgæfa þarf flettulistann sérstaklega í hverju nýju orðabókarverki en jafnframt þurfa orðabókarmenn að nýta þær heimildir sem til eru. Þá er vert að gera sér grein fyrir kostum og göllum þessara heimilda.

á lokastigi, þ.e. 7000 algengustu orðin í lista frá Eiríki Rögnvaldssyni og Vilhjálmí Sigurjónssyni. Ekki kemur fram við hvaða texta var stuðst í þeim lista.

2.2 Gagnasöfn Orðabókar Háskólans

Seðlasöfn OH eru frumheimildir Ío, ásamt Blöndalsorðabók, og þangað sækja orðabókarmenn enn efni. Þetta á einkum við um Ritmálssafnið sem er að stórum hluta aðgengilegt á vefnum til skoðunar. Sökum stærðar og takmarkana í leit er það þó fremur erfitt viðureignar en flettiorð þar eru yfir 700 þúsund. Talmálssafnið er enn sem komið er ekki aðgengilegt til úrvinnslu nema innan stofnunar. Efni úr báðum þessum söfnum var notað við gerð Ío, sem og einnig við gerð *Íslenskrar orðsifjabókar* (Ásgeir Blöndal Magnússon 1989). Bæði þessi söfn voru gerð með hefðbundnum aðferðum, orðtöku og spurningum til almennings. Til Ritmálssafns var safnað allt frá stofnun OH árið 1944 og rétt fram yfir 1990 en efni í Talmálssafn var einkum safnað í tengslum við þáttinn Íslenskt mál sem fluttur var í Ríkisútvarpinu 1956–2005. Eftir það hefur athyglinni verið beint að söfnun á rafrænum textum í *Íslenskt textasafn* (sjá kafla 3.1).

Hefðbundin orðtaka úr ritmáli felst í lestri á textum en þar birtist aðeins hluti af daglegu máli. Við söfnun efnis til Talmálssafnsins var yfirleitt leitað að tilteknum orðum eða lýsingum á fyrirbærum þannig að orðaforðinn var bundinn við tilteknar aðstæður eða orð, eins og liggur í hlutarins eðli í útvarpsþætti. Í þessar heimildir hafa íslenskir orðabókarmenn sótt sér efni í flettulista um langa hrið en gallinn er sá að efnið er bæði of mikið, þar sem orðaforðinn er gríðarlegur og því seinlegur til úrvinnslu, og of lítið, þar sem orð úr daglegu máli nútímans er ekki endilega að finna þarna. Þá er sá galli á söfnunum báðum að þar er safnað efni sem vekur athygli orðtökumannsins fyrir einhverra hluta sakir. Jaðartilvik, t.d. óvenjuleg málnotkun eða sjaldgæf orð, fá þá e.t.v. of mikið vægi á kostnað þess sem er almennt og algengt, sbr. dæmin *ektamaður* og *eiginmaður* sem nefnd voru hér að framan.

Önnur orðfræðileg gagnasöfn koma minna við sögu við gerð almennra nútímamálsorðabóka, þótt þau séu frábærar heimildir í sögulega orðabók, eins og t.d. seðlasafn um orðabók Jóns Ólafssonar úr Grunnavík,⁴ sem er góð heimild um mál 18. aldar. Íðorðaforðann er síðan að finna í Orðabanka Íslenskrar málstöðvar⁵ og þangað leita orðabókarmenn eftir því sem þörf er á. Ekki verður frekar fjallað um

4 Orðabókin er í handritinu AM 433 fol. Hún er óútgefin en seðlar í uppskrift Jakobs Benediktssonar eru varðveittir á orðfræðisviði Stofnunar Árna Magnússonar í íslenskum fræðum og til er aðgengileg tölvuskra yfir uppflettiorðin (sjá: http://www.arnastofnun.is/page/ordabokarhandrit_jons_olafssonar).

5 Sjá: <http://www.ordabanki.hi.is/wordbank/search>.

Þessi söfn hér enda er sjónum einkum beint að því hvar nálgast má almennan orðaforða. Í Ío 1963 birtist fyrsta tilraun til að vinna úr efni í seðlasöfnum OH en í Norræna verkefninu svokallaða var aftur gerð tilraun til þess að sameina orðaforða úr gögnum OH til nota í orðabók.

2.3 Norræna verkefnið

Á árunum 1994–1996 var unnið að gerð íslensks orðabókarstofns sem ætlaður var sem grunnur að tvímála orðabókum milli íslensku og annarra Norðurlandamála. Verkefnið hlaut vinnuheitið Norræna verkefnið (Kristín Bjarnadóttir 1998). Upprunalega var ætlunin að orðabókarstofninn tæki til almenns orðaforða og að flettiorð í honum yrðu u.þ.b. 50 þúsund, úr lifandi almennu nútímamáli. Efni úr *Orðastað* Jóns Hilmars Jónssonar var notað sem grunnur en frumheimild þar er Ritmálssafnið (Jón Hilmar Jónsson 1994, formáli, bls. xxx). Síðan var bætt við meira efni úr Ritmálssafni og öðrum söfnum OH eftir þörfum. Allur orðaforði úr *Íslenskri orðtíðnibók* (Jörgen Pind o.fl. 1991) var hafður með og loks var stuðst við prentaðar orðabækur, s.s. Ío (1983) sem var höfð til hliðsjónar. *Íslensk orðtíðnibók* er byggð á textasafni með 500 þúsund lesmálsorðum og þaðan koma u.þ.b. 30 þúsund uppflettiorð, en annars er efnið í Norræna verkefninu fengið með hefðbundinni aðferð, þ.e. hér er efni úr eldri orðabókum og söfnum endurnýtt en þar liggur að baki hefðbundin orðtaka.

Til að bæta úr eyðum í orðaforðanum í Norræna verkefninu var nokkuð af textum orðtekið sérstaklega. Þar á meðal voru bæklingar um heilbrigðismál, félagsmál, tryggingar, bankakerfið, sálfræði, lögfræði fyrir almenning, sjávarútveg og tölvur, og námsbækur, t.d. í örverufræði, matvælafræði, vélsmiði, húsgagnasmíði, íþróttum o.fl. Einnig voru gerðar tilraunir með orðtöku úr rafrænum textum, aðallega úr *Morgunblaðinu*.

Sigurborg Hilmarsdóttir stjórnaði orðtökunni og við hana unnu t.d. fjórir starfsmenn sumarið 1995 og söfnuðu u.þ.b. 2.500 orðum. Morgunblaðið reyndist drýgst í leit að orðum úr almennu máli. Í töflu 1 eru nokkur dæmi þaðan.

áherslupenni	grasrótarhreyfing	ljósleiðari
barnaís	heimasíða	margmiðlun
bleksprautuprentari	húsbréfalán	myndlykill
buxnadragt	kapalkerfi	raðgreiðsla
eldislax	krókaleyfi	ríkisvixill
fjárlagahalli	leikskólakennari	talhólf
flísefni	líkamsklukka	

Tafla 1. Dæmi um orðtöku úr Morgunblaðinu.

Orðtakan var mjög tímafrek og var unnin með hefðbundnum aðferðum, þ.e. textarnir voru lesnir og leitað að orðum sem þóttu hafa gildi í orðabók. Virkar samsetningar voru ekki teknar með. Afraksturinn varð nokkur þúsund orð, dýrmæt og dýr.

Alls eru 241.784 flettiorð í Norræna verkefninu og eru þau flokkuð eftir orðgerð, þ.e. í ósamsett, afleidd og samsett orð, og eftir flettugildi, þ.e. eftir því hversu mikið erindi orðin eru talin eiga í orðabók. Þar er gerður greinarmunur á virkum samsetningum og orðgervingum (lexíkalíseruðum orðum) og einnig tekið tillit til tíðni og notkunar orða, eftir því sem gerlegt var á þeim tíma þegar verkið var unnið en þá var ekki um greiðan aðgang að tíðnitölum að ræða.

Í Norræna verkefninu var reynt að sameina allar heimildir sem tiltækar voru á OH á þeim tíma sem verkið var unnið. Þetta var í þann mund sem aðgangur að tölvutækum textum var að opnast en engar aðferðir voru tiltækar til þess að nýta þá til orðtöku umfram það sem gert var í *Íslenskri orðtíðnibók*. Ekki fékkst styrkur til þess að halda vinnu við Norræna verkefnið áfram árið 1996 og það lagðist í dvala. Þar varð samt til mikið orðasafn sem að mestu var tekið saman með hefðbundnum aðferðum og það er þess vegna gott til samanburðar við efni sem aflað verður með nýrri aðferðum. Þórdís Úlfarsdóttir (2013) tekur fram að 86% af orðaforðanum í ISLEX sé úr Norræna verkefninu þannig að efnið nýttist ágætlega í því verki.

3 Nýjar heimildir og nýjar aðferðir við orðtöku

Með tilkomu tölvutækra texta breyttist heimur orðabókarmannsins. Í stað hefðbundinnar orðtöku þar sem textar eru lesnir á pappír, kemur vélræn orðtaka sem byggist á því að láta tölvu finna orð eða orðmyndir í rafrænum textum sem ekki eru í samanburðarefni. Milli-stigið er að orðtaka úr rafrænum texta á skjá, með því að lesa textann,

en á þeirri aðferð og hefðbundinni lestrarorðtöku er bitamunur en ekki fjár. Talsverður vinnusparnaður er samt sem áður fólgin í þessari aðferð þar sem ekki þarf að skrifa efnið upp eða slá það inn. Í eiginlegri vélrænni orðtöku er markmiðið að fá yfirlit um allar orðmyndir í textunum en orðtakan verður markvissari eftir því sem greiningin á textunum er betri þar sem tvíræðni orðmynda er gríðarleg í íslensku.⁶

Sá grundvallarmunur er á vélrænni orðtöku og öðrum orðtöku-aðferðum að í vélrænu orðtökunni er allur orðaforðinn í viðkomandi texta undir; þar er unnið með tæmandi lista af orðum og orðmyndum. Þar þarf því að grisja og velja efnið í hvert orðabókarverk fyrir sig, eftir þeim aðferðum sem henta hverju sinni. Í hefðbundinni orðtöku er orðunum hins vegar safnað eftir tilfinningu orðtökumannsins, venjulega án mikils samanburðar við það efni sem fyrir er. Þar takmarkast umfangið á textunum af því sem er í mannlegum mætti að lesa og þess vegna er að einhverju leyti tilviljanakennt hvaða orð eru orðtekin.

Rafrænir textar eru að sjálfsögðu einnig notaðir til dæmaleitar í orðabókargerð. Með því að ganga út frá fyrirframgefnum orðalista og nota textasöfn eða málheildir til dæmaleitar fæst betra yfirlit yfir raunverulega orðanotkun og setningarlegt umhverfi en áður hefur verið kostur á. Hins vegar skilar slík vinna ekki yfirliti yfir heildar-orðaforðann í textunum nema með notkun orðtökutóls þar sem allar orðmyndir eru skoðaðar. Þarna á milli er verulegt stökk í tæknilegum efnum þar sem dæmaleit er möguleg án þess að fyrir liggi málfræðileg greining á textunum en orðtaka byggist á því að textarnir séu greindir og lemmaðir (þ.e. að gefið sé uppflöttiorð eða lemma fyrir hverja orðmynd).

Helstu gagnalindir orðabókarmannsins nú um stundir eru því rafrænir textar, hverju nafni sem þeir nefnast, og *Mörkuð íslensk málheild* (MÍM), þar sem málfræðileg greining fylgir hverri orðmynd (Sigrún Helgadóttir o.fl. 2012). *Beygingarlýsing íslensks nútímamáls* (BÍN) gegnir hlutverki samanburðarefnis í þessu tilliti en BÍN er safn 270 þúsund beygingardæma, u.þ.b. 5,8 milljónir beygingarmynda með greiningarstrengjum (mörkum) (Kristín Bjarnadóttir 2012). Efni úr almennu máli í BÍN er að mestu leyti úr orðabókum og söfnum OH, þ.m.t. Norræna verkefninu.

6 Í *Beygingarlýsingu íslensks nútímamáls* (BÍN) hafa aðeins 31,5% orðmynda bara einn greiningarstreng (Kristín Bjarnadóttir 2012). Sjá nánar um BÍN hér á eftir.

3.1 Rafrænir textar

Á OH var byrjað að safna rafrænum textum um eða upp úr 1980. Safnið heitir *Íslenskt textasafn* og í því eru nú u.þ.b. 67 milljón lesmálsorð. Þar eru gamlir og nýir textar af ýmsum toga sem skiptast á milli 30 efnisflokka. Safninu fylgir ekki orðalisti og engin leið er að komast að því hver orðaforðinn er nema með málfræðilegri greiningu. Orðabókarmaðurinn þarf því að vita fyrir fram að hverju hann er að leita og getur ekki að óbreyttu notað safnið til orðtöku. Það er hins vegar frábær uppspretta dæma, með þeim annmarka þó að flokkun dæmanna getur verið mjög seinleg, m.a. vegna tvíræðni orðmynda. Í algengum orðum geta dæmi líka orðið óviðráðanlega mörg og þá getur verið djúpt á því sem leitað er að og erfitt að ná yfirsýn yfir efnið.

Önnur heimild sem vænlegt er að leita dæma í er t.d. tímaritasafn Landsbókasafns – Háskólabókasafns, *Tímarit.is*. Þar eru ljóslesnir textar, ásamt myndum af ritunum. Kosturinn við tímaritasafnið er að aldursmerking skilar sér vel með dæmunum en helsti ókosturinn er sá að í ljóslestrinum er talsvert af villum. Þá er vert að nefna *Íslenskan orðasjóð*, sem áður var minnst á (sjá nmgr. 2), sem er risavaxið safn af íslenskum vefsíðum en talsvert er af villum í því safni líka. Þessar heimildir nýtast best til leitar að tiltölulega sjaldgæfum orðum sökum þess hve seinlegt er að leita í þeim.

3.2 MÍM og BÍN

MÍM er textasafn með málfræðilegri greiningu sem unnið hefur verið að á OH og Stofnun Árna Magnússonar í íslenskum fræðum frá árinu 2002. Textarnir eru allir frá 2000–2009 og þeir eru úr eins fjölbreyttu efni og nokkur kostur er. Í MÍM eru u.þ.b. 25 milljón lesmálsorð. Hverju orði fylgir mark, þ.e. greiningarstrengur fyrir orðflokk og beygingarlegar upplýsingar. Markamengið er stórt, yfir 600 mörk (Sigrún Helgadóttir 2012).

Með því að bera saman MÍM og BÍN ætti að vera hægt að finna ný orð í málinu eða a.m.k. þau orð sem ekki hafa ratað til orðabókarmanna. Jafnframt finnast gloppur í BÍN sem þarf að uppfæra jafnóðum. Mikilvægasti kosturinn við MÍM til orðabókargerðar er sá að þar gefst í fyrsta sinn kostur á skilvirkri vinnu við orðalista með tíðnitölum sem t.d. er hægt að nota til flokkunar á uppflettiorðum í orðabók. Með þessu ætti að vera tryggt að algengustu orð vanti ekki,

eins og stundum hefur viljað brenna við þegar stuðst er við tilfinningu orðabókarmannsins.

Mörkun og lemmun á MÍM er ekki lokið og þar leynast enn villur, bæði í uppflettimyndum og orðfloknum. Vegna þess liggja tölur um lemmufjölda og tíðni ekki fyrir enn. Hér á eftir er sagt frá tilraun til að áætla lemmufjölda í hluta af MÍM en hún byggist á samanburði orðmynda úr MÍM og beygingarmynda úr BÍN. Samanburðurinn var gerður haustið 2011 og takmarkið með honum var að bæta orðaforðanum úr MÍM við BÍN. Samkvæmt lauslegri áætlun eru u.þ.b. 125 þúsund lemmur í MÍM sem ekki finnast í BÍN.⁷

Í þeim hluta MÍM sem borinn var saman við BÍN eru u.þ.b. 17,5 milljónir lesmálsorða. Þar af voru einkvæmar eða ótvíræðar orðmyndir (með marki) tæplega 738 þúsund og voru þær bornar saman við 5,8 milljónir beygingarmynda úr BÍN. Niðurstaðan úr samanburðinum er í töflu 2.

Orðmyndir í MÍM (með marki)	737.856
Þar af í BÍN	425.238
Þar af ekki í BÍN	312.618

Tafla 2. Orðmyndaafjöldi í MÍM og BÍN.

Orðmyndir úr MÍM sem finnast í BÍN eru taldar vera tækar íslenskar orðmyndir, þ.e. 425.238 orðmyndir af 737.856. Orðmyndir sem ekki voru í BÍN voru skoðaðar hver fyrir sig á skjá, ríflega 312 þúsund orðmyndir. Af þeim eru u.þ.b. 60% íslenskar orðmyndir en 40% er annað efni, t.d. erlend orð (25%), villur (6%), skammstafanir (1,7%) o.fl. Þetta aðskotaefni eru u.þ.b. 13% af einkvæmum orðmyndum í MÍM en ekki liggur fyrir hve stór hluti af lesmálsorðunum þetta er. Allar orðmyndir sem hafa verið aðlagðar að íslensku málkerfi flokkast sem „íslenskar orðmyndir“, þ.m.t. allar beygðar slettur (*pródúsjónin*) og beygð erlend nöfn, t.d. grísku nöfnin *Parmenídes* og *Erastopenes* (ef. *Parmenídesar*, *Erastopenesar*).

Niðurstaða af samanburðinum á MÍM og BÍN er að í MÍM séu u.þ.b. 125 þúsund lemmur sem ekki er að finna í BÍN. Samanlagður lemmufjöldi í BÍN og MÍM er því sennilega yfir 400 þúsund. Til þess að nýta þetta efni til orðabókargerðar þarf að grísa það, t.d. með því að styðjast við tíðnitölur úr MÍM þar sem orðaforðinn í BÍN er fenginn úr orðalistum sem hvorki styðjast við tíðnikannanir né neins konar

⁷ Samanburðurinn var unninn í samvinnu við Jón Friðrik Daðason og Sigrúnu Helgadóttur og kann höfundur þeim bestu þakkir fyrir.

flokkun eftir flettugildi til orðabókarnota. Grisjunin er allt annað en einföld þar sem tíðnitölurnar einar duga ekki. Orðið *doppóttur* kemur t.d. aðeins þrisvar sinnum fyrir í MÍM og orðið *rykking* alls ekki en samt eru orðin síður en svo sjaldgæf í umræðu um saumaskap. Orðaforði af því sviði tilheyrir reyndar þeim hluta orðaforðans sem til skamms tíma vantaði í orðabækur, þ.e. heimilisorð ýmiss konar eða orð af verksviði kvenna. Orð sem koma sjaldan fyrir í rafrænum textum geta því vissulega átt heima í orðabók og orðin tvö sem hér eru nefnd eru bæði uppflettiorð í Ío og í ISLEX.

Þegar tíðni orðanna í MÍM sem ekki eru í BÍN er skoðuð koma strax í ljós gloppur í hefðbundna orðabókarefninu sem BÍN byggist að mestu leyti á. Dæmi um efni af þessu tagi eru í töflu 3 en þar eru m.a. algeng orð úr MÍM sem hljóta að eiga heima í orðabókum.⁸

	<i>Orðmynd</i>	<i>Mark</i>	<i>Lemma</i>	<i>Ofl.</i>
1262	vísindavefnum	nkeþgs	vísindavefur	kk
282	sonu	nkfo	sonur	kk
258	grunnskólabekk	nkeþ	grunnskólabekkur	kk
231	r-listans	nkeegs	r-listi	kk
177	knattspyrnustjóri	nken	knattspyrnustjóri	kk
158	skjöldu	nveo	skjöldur	kk
151	vísindavefsins	nheegs	vísindavefur	hk>kk
141	ameríkufari	nken-s	ameríkufari	kk
126	íbúðalánasjóðs	nkee-s	íbúðalánasjóður	kk
122	r-listinn	nken-s	r-listinn	kk.mgr.
111	vopnfirðinga	nkfe-s	vopnfirðingur	kk
101	ríkislögreglustjóra	nkee	ríkislögreglustjóri	kk
99	óskarinn	nkeng	óskar	kk
96	maður	nken	maður	kk
95	mogganum	nkeþgs	moggi	kk
93	danakonungur	nken-s	danakonungur	kk
93	danakonungs	nkee-s	danakonungur	kk

Tafla 3. Orðmyndir í MÍM sem ekki eru í BÍN.

Vandi orðabókarmannsins við úrvinnslu á efni úr MÍM er tvenns konar. Í fyrsta lagi er massinn gríðarlegur og seinlegt að vinna úr öllu efniinu, jafnvel þótt öll greining væri rétt. Það erfiði er þó fremur fólgið í vinnu við merkingargreiningu en beinlínis í því sem hér er til

8 Í töflunni eru einnig úreltar beygingarmyndir sem ekki finnast í BÍN þó að þær komi fyrir í formlegu máli (*sonu*, *skjöldu*) og stafsetningarafbrigði sem finnast í tölvumáli þar sem séríslenskum stöfum er sleppt (*maður*). Í samkeyrslunni var ekki gerður greinarmunur á hástaf og lágstaf í samkeyrslunni (*r-listi*, *danakonungur*).

umræðu, þ.e. vinnu við orðalistann sjálfan. Í öðru lagi er orðaforðinn ekki eins mikill og æskilegt væri, eins og sjá má af orðinu *rykking* sem nefnt var hér að framan og á samanburði á orðamyndafjöldanum í MÍM og uppflettiorðafjölda í hefðbundnu orðabókarefni, t.d. Ritmálssafni, en MÍM skilar aðeins broti af þeim orðaforða sem þar er.⁹

Textasafn sem ætti að skila öllum þeim orðaforða sem venjulegir málnotendur hafa að einhverju leyti á valdi sínu þyrfti að vera gríðarlegt, án þess að því sé haldið fram að allur sá orðaforði sé á færi hvers manns. Þetta er einmitt efni sem ætla má að notendur vilji leita að í orðabók.

4 Vélræn orðtaka

Þær hugmyndir hafa verið uppi um allnokkurn tíma að gott væri að orðtaka efni jafnóðum og það birtist, t.d. á vefsíðum dagblaða. Með BÍN fylgir orðtökutól og með því er hægt að finna orðmyndir í textum sem ekki eru í BÍN. Tólið er enn á tilraunastigi og er frumstætt; það lemmar ekki og skilar býsna óaðgengilegum listum af orðmyndum (og öðrum strengjum). Úrvinnslan er því mjög seinleg. Til þess að tólið komi að fullu gagni þarf málfræðilega greiningu á textunum (mörkun), lemmun og síu fyrir rusl, samanber tölur um aðskotaefni í samanburðinum á orðaforðanum í MÍM og BÍN hér að framan. Með sjálfvirkri mötun á textum, t.d. dagblöðum hvers dags, væri þannig hugsanlegt að finna ný orð sem birtast frá degi til dags, eða a.m.k. orð sem ekki hafa ratað á fjörur orðabókarmanna fyrr.

Þegar orðtökutólið var prófað á stuttan texta um orðabókarfræðilegt efni var afraksturinn fjögur uppflettiorð (orðmyndir í sviga) og niðurstaðan er býsna dæmigerð, þ.e. fjögur samsett orð:

gagnaefni (gagnaefnis), *einmála* (einmála), *orðabókareining* (orðabókar-eininganna), *orðabókargögn* (orðabókargögn, orðabókargagna).

Íslensk orðmyndun er afskaplega virk og tilraunir með tólið sýna að samsetningar eru líklegasta afurðin í venjulegum texta. Vandí orðabókarmannsins er að finna hvaða samsetningar eiga erindi í orðabók, þ.e. þarna þarf að sigta úr þær samsetningar sem ekki eru fullkomlega gagnsæjar (Kristín Bjarnadóttir 1995).¹⁰

9 Ritmálssafnið er að vísu efni til sögulegrar orðabókar, allt aftur til 16. aldar.

10 Í tvímálaorðabók þarf einnig að taka tillit til þýðingarorðanna.

Virkni og einkenni íslenskrar orðmyndunar verða einnig til þess að tíðnitölur orða og sjálfstæðra orðmynda skila e.t.v. ekki eins afgerandi niðurstöðum fyrir íslensku og t.d. fyrir ensku, eins og sjá má af þessu dæmi:

kransæðarhjáveitugræðlingur
coronary artery bypass graft

Þarna er eitt orð í íslensku fyrir fjögur í ensku. Íslensku uppflettiorðin sem vert er að skoða eru *kransæð*, *hjáveita*, *græðlingur* og/eða *kransæðar-hjáveita/hjáveitugræðlingur*. Vandi orðtökumannsins er því sá að einnig þarf að skoða orðhlutana til þess að grandgæfa orðaforðann í textunum, samanber þetta samsetta orð. Forrit sem greinir samsett orð þarf því að vera hluti af orðtökubúnaðinum¹¹ og tíðni orðhlutanna þarf sennilega einnig að koma við sögu ef nota á orðtíðni í íslenskri orðabókargerð. Það verk bíður þó síðari tíma þegar orðtökutólið hefur batnað.

5 Lokaorð

Nýjar gagnauppsprettur gera orðabókarmönnum fært að endurskoða orðaforðann í verkum sínum á skilvirkari hátt en áður hefur verið mögulegt þannig að betur sé komið til móts við notendur með því að ná yfir fjölbreytta málnotkun á þeim tíma sem orðabókin er samin. Hins vegar eru ýmsir annmarkar á því að byggja íslenska orðabók alfarið á tíðnitölum, eins og gert var í *Cobuild*-orðabókinni sem nefnd er í upphafi þessarar greinar. *Cobuild* er byggð á málheild með u.þ.b. 26 milljónum lesmálsorða en lesmálsorð í MÍM eru u.þ.b. 25 milljónir í endanlegri gerð. Í *Cobuild* munu vera „over 70,000 references“ eða uppflettimyndir (sjá bókarkápu) en samkvæmt ágiskun eru u.þ.b. 300 þúsund uppflettiorð í MÍM. Ætla mætti að svo mörg uppflettiorð skilðu niðurstöðu sem hægt væri að byggja á til orðabókargerðar en mjög virkar samsetningarreglur verða til þess að tíðnitölur í ensku og íslensku eru ekki sambærilegar, eins og vikið var að hér að framan. Vegna þessa verður tíðnidreifing í íslensku ekki sambærileg við ensku. Til dæmis koma rúmlega 308 þúsund orðmyndir úr þeim hluta MÍM sem borinn var saman við BÍN aðeins fyrir einu sinni.¹² Stakdæmi

11 Jón Friðrik Daðason (2012) notar forrit sem greinir samsett orð í leiðréttingarbúnaði sínum en tólinu hefur ekki verið beitt í orðtöku enn sem komið er.

12 Fjöldi lemma í þessum hluta MÍM er 407.192 en af þeim koma 229.895 aðeins

eru veikur grunnur fyrir uppflettiorð í orðabókum og m.a. þess vegna er nauðsynlegt að skoða tíðni orðhlutanna líka. Tilgáta mín er því að orðaforðinn úr MÍM myndi skila gloppótttri orðabók og að hefðbundinn efniviður íslenskra orðabóka, orðalistar og gagnasöfn, sé einnig nauðsynlegur.

Eftir þetta yfirlit um efni og aðferðir til að velja orðaforða í íslenska orðabók er niðurstaðan þessi:

- Hefðbundnir orðalistar eru enn sem komið er ómissandi. Þá þarf að grisja og uppfæra með nýjum aðferðum.
- Vélræn orðtaka er ekki enn komin í gagnið en mun nýtast til að hafa fingurinn á pulsinum og finna ný orð.
- MÍM og önnur stór rafræn söfn munu verða gríðarlega mikilvæg heimild um orðaforðann. Til þess að þau komi að fullum notum þarf að greina samsett orð í sundur.
- Engar vélrænar aðferðir eru enn til til þess að meta flettugildi orða. Þar þarf orðabókarmaðurinn að meta efnið á hefðbundinn seinlegan hátt, þótt hægt sé að styðjast við tíðnitölur, með þeim annmörkum sem á þeim eru í því efni sem nú er til reiðu. Textasöfnin eru einfaldlega ekki enn nægilega yfirgripsmikil til þess að innihalda allan þann orðaforða sem birtist í heimildum sem safnað er til með gamla laginu.

Heimildir

- Aðalsteinn Eyþórsson. 1998. Síðasti furðufuglinn — eða gæsin sem stóð á öndinni. *Frejas Psalter til brug for Jonna Louis-Jensen*, bls. 5–7. København. (Einnig birt á *Kistunni* 20. maí 2002: www.kistan.is; sótt 12.3.2013.)
- Aðalsteinn Eyþórsson. 2005. Hver er kjarni orðaforðans? *Orð og tunga* 7:9–20.
- Árni Böðvarsson (ritstj.). 1963. *Íslensk orðabók* handa skólum og almenningi. Reykjavík: Bókaútgáfa Menningarsjóðs.
- Árni Böðvarsson (ritstj.). 1983. *Íslensk orðabók*. 2. útg. Reykjavík: Bókaútgáfa Menningarsjóðs.
- Ásgeir Blöndal Magnússon. 1989. *Íslensk orðsifjabók*. Reykjavík: Orðabók Háskólans.
- Ásta Svavarsdóttir. 2008. Ordbogsbrug og ordbogskultur i Island. Under-søgelse af ordbogsbrug blandt islandske modernmålslerere. *LexicoNordica* 15:115–134.

einu sinni fyrir. Lemmunin er óleiðrétt, eins og áður kom fram, og tölurnar munu breytast í endanlegri gerð.

- Beygingarlýsing íslensks nútímamáls*. Ritstjóri Kristín Bjarnadóttir. <http://bin.arnastofnun.is>.
- Björn Halldórsson. 1814. *Lexicon Islandico-Latino-Danicum Björnönis Haldor-sionii*, I–II. Havniae.
- Blöndalsorðabók = Sigfús Blöndal. 1920–1924. *Íslensk-dönsk orðabók*.
- Cobuild = Sinclair, John (ritstj.). 1987. *Collins Cobuild English Language Dictionary*.
- Erla Hallsteinsdóttir. 2007. Íslenskur orðasjóður. *Orð og tunga* 7:109–124.
- Guðbrandur Vigfússon. 1874–1876. *An Icelandic-English Dictionary based on the MS. Collections of the late Richard Cleasby*. Oxford.
- Guðrún Kvaran. 1997. Rætur og heimildir. *Orð og tunga* 3:9–14.
- Ío = *Íslensk orðabók* (sjá Árni Böðvarsson 1963, 1983; Mördur Árnason 2000.)
- Íslenskur orðasjóður*. http://wortschatz.uni-leipzig.de/ws_ice/.
- Íslenskt textasafn*. http://arnastofnun.is/page/arnastofnun_gagnasafn_textasafn.
- Jón Friðrik Daðason. 2012. *Post-Correction of Icelandic OCR Text*. MS ritgerð í tölvunarfræði við Háskóla Íslands. (Sjá: <http://hdl.handle.net/1946/12085>.)
- Jón Hilmar Jónsson. 1994. *Orðastaður*. Reykjavík: Mál og menning.
- Jón Hilmar Jónsson. 2013. Að byggja undir íslensk-erlenda orðabók. Forgreining og orðabókarefni. *Orð og tunga* 15:1–22.
- Jörgen Pind, Friðrik Magnússon og Stefán Briem. 1991. *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.
- Kristín Bjarnadóttir. 1995. Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base. Í: Ásta Svavarsdóttir, Guðrún Kvaran, Jón Hilmar Jónsson (ritstj.), *Nordiske studier i leksikografi* 3, bls. 255–263. Reykjavík: NFL, Nordisk språksekretariat og Orðabók Háskólans.
- Kristín Bjarnadóttir. 1998. *Norræna verkefnið*. Skýrsla um íslenskan orðabók-arstofn, 6. mars 1998. <http://www.lexis.hi.is/kristinb/norr.pdf>.
- Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. *Proceedings of "Language Technology for Normalization of Less-Resourced Languages"*, SaLTMiL 8 – AfLaT 2012. Istanbul, Tyrklandi.
- Mördur Árnason. 1998. Endurútgáfa *Íslenskrar orðabókar*: Stefna – staða – horfur. *Orð og tunga* 4:1–8.
- Mördur Árnason (ritstj.). 2000. *Íslensk orðabók*, tölvuútgáfa. Reykjavík: Edda hf.
- Orðabanki Íslenskrar málstöðvar*. Ritstjóri Ágústa Þorbergsdóttir. <http://www.ordabanki.hi.is/wordbank/>.
- Mörkuð íslensk málheild*. Ritstjóri Sigrún Helgadóttir. <http://mim.hi.is>.
- Ritmálssafn Orðabókar Háskólans*. http://www.arnastofnun.is/page/gagnasofn_ritmalssafn.
- Sigfús Blöndal. 1920–1924. *Íslensk-dönsk orðabók*. Reykjavík.
- Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarna-

- dóttir og Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). *Proceedings of "Language Technology for Normalization of Less-Resourced Languages"*, SaLTMiL 8 – AfLaT 2012. Istanbul, Tyrklandi.
- Sinclair, John (ritstj.). 1987. *Collins Cobuild English Language Dictionary*. London and Glasgow: Collins.
- Sverrir Hólmarsson, Christopher Sanders, John Tucker. 1989. *Íslensk-ensk orðabók*. Reykjavík: Iðunn.
- Timarit.is. Landsbókasafn Íslands – Háskólabókasafn: <http://timarit.is>.
- Quasthoff, Uwe, Sabine Fiedler, Erla Hallsteinsdóttir. 2012. *Frequency Dictionary. Icelandic. ISL*. Leipzig: Leipziger Universitätsverlag.
- Þórdís Úlfarsdóttir. 2013. ISLEX – Norræn margmála orðabók. *Orð og tunga* 15:41–71. (Þetta hefti.)

Lykilorð

orðabókarfræði, uppflettiorð, orðaforði, málheild, vélræn orðtaka

Keywords

lexicography, headwords, vocabulary, corpora, automatic excerption

Abstract

The topic of this paper is different methods in assembling a list of headwords for Icelandic dictionaries. Until recently, the classic Icelandic dictionary (*Íslensk orðabók*) originally published in 1963, has been the primary source of data of the vocabulary for Icelandic lexicographers, with the collections at the department of lexicography of The Árni Magnússon Institute for Icelandic Studies as additional sources. With the advent of electronic text collections, notably the tagged Icelandic corpus (*Mörkuð íslensk málheild*, MÍM), Icelandic lexicographers now have access to a huge new source of data. The use of data on frequency from these new sources is of great value to lexicographers in the choice of headwords, but the complete coverage of the words and word forms in the texts made possible by new methods of word extraction also complements older material by filling accidental gaps in material assembled by older methods of dictionary excerption. Partly because of data scarcity problems in a language with a very rich morphology, the conclusion in the paper is that word frequency alone cannot be the base of an Icelandic dictionary at this point in time, as the volume of texts needed for a good coverage of the vocabulary far exceeds anything available now. One of the main reasons for this is a very productive system of compounding, making it necessary to split compounds before attempting to use frequency for the selection of vocabulary.

Kristín Bjarnadóttir

Stofnun Árna Magnússonar í íslenskum fræðum / Háskóli Íslands

kristinb@hi.is

