

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

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-

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

²Verkefnisstjóri 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 málfræðiatríða, oftast orðflokks og beygingarmyndar, og tengt við ákveðið uppfletti-orð (lemmu eða les). Með því móti er gerður greinarmunur á tví- eða

³Eyrún 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 birt-ingarmynd 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.

Pegar 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 opin og hin eru misaðgengileg til nota í málrannsóknunum. 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

⁴Benda 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 Íslendingasagna⁶ 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 meg 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

⁵Sumarið 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 réttihöfum textanna er ætlunin að opna aðgang að safninu í því formi á vefsíðu stofnunarinnar. Verkefnið var styrkt af Nýsköpunarsjóði námsmanna.

⁶Hér 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% nytjatekstar (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% nytjatekstar. Þá eru ævisögur og endurminningar flokkaðar með nytjatekstum 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 efni-viður í hagnýtum og fræðilegum verkefnum. Þar má í fyrsta lagi nefna tungutækni-verkefni 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 *vúdu-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 byggji 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 skörum 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 mállysingar 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 notkunarviðs. 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-verkefnið⁷árið 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 Alþingi⁹frá 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 2006¹⁰(u.þ.b. 4 klst.)

⁷Markmið ÍSTAL var að safna efni í íslenskan talmálsbanka til notkunar við málsrannsóknir og tungutæknaverkefni 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 tækniþjóði 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).

⁸MIN 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/>.

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

¹⁰Sigrún 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 rannsóknna. 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 tiltekkin 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 rétttri 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. *óní* (‘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 fram-
burði. Á hinn bóginn þótti venjuleg stafsetning tryggja best þá stöðl-
un 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, framí-
grip, 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: maður ýtir bara á er það ekki já

B: jú bara

B: þú bara sérð og

A: halló mamma

D: <A>æðislegt

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 for-
rit 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: maður 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 heyrir 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 textaskráum 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 textaskrár¹² 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 (*svo-leiðis, rosalega, sko*). Allar koma þær miklum mun sjaldnar fyrir í ritmálsefninu en í talmálgögnunum og í sumum tilvikum er munurinn sláandi (*bara*). Sem kunnugt er byggist *Íslensk orðtíðnibók* (1991) ein-

¹²Umritunarskrám 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 leitar síðunni <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 orðræðu-agnir 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. upphróp-anir (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 talmáls-efni 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 boðhátarmynd 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 sígar-
ettu?* / **heyrðu mig (mér)** hlustaðu nú á mig sbr. *heyra*

Í talmálsefninu í heild eru rúmlega 240 dæmi um orðmyndina *heyrðu*, allt boðhátarmyndir. Sé rýnt í dæmin úr ÍSTAL-samtölunum virðist orðmyndin þó ekki notuð í eiginlegri boðhátarmerkingu, 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 boðhátarmyndin 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:	heyrdur <A1>hvernig hvernig</A1> stendur á helgi þá sko getum við
2 allt safnast saman hjá þér	heyrdur já átjándi og þú átt að fara að já nú sé ég af hverju þú varst svona
3 B: ert þú ekki tilvalin í það A:	heyrdur hún spurði einmitt bara (lang-) langar þig ekki ((dynkir)) að
4 og sagði hvenær og og þá já	heyrdur þá er best að ég geri það líka heldurðu að ég geti fengið þær eftir
5 síðan í morgun maður var= A: =	heyrdur hvað= B: =ég veit ekki hvernig stendur á því þetta er svona langur
6 skrifað dagsetningu á já	heyrdur Kári var hérna þetta gamall þegar já hann var í þessum fótum ú ég
7 guð hvað á ég að segja núna	heyrdur hvernig er þetta nú særi ég hana og (blabla) A: já B: en hún getur
8 tengdamamma allt í einu svona	heyrdur já Ella er ekki einhver sem þú þekkir sem heitir Jóhannes Freyr
9 hefðirðu getað farið að skoða A:	heyrdur já það minnir mig ég ætla að athuga bækurnar þarna kannski B: já
10 þæja gjörði þið svo vel D: já A:	heyrdur Guðrún Ösp kemst ekki C: (x) þau voru hérna áðan B:
11 C: ja hann fékk sér bara aðeins	heyrdur <ID> rjóminn D: já A: ég ætlaði (x) að fara að segja (x) rjómann
12 alla vega bæði á r D: já B:	heyrdur ég man alltaf eftir sögunni af hérna (.) sem amma þín sagði mér (.)
13 C:	heyrdur þetta er allt <A>þetta er allt þetta er allt þrjónað fast
14 þessa uppskrift ### A:	heyrdur Inga þú mátt eiga þetta því að <B1>## frosna</B1> grænmetið
15 B:	heyrdur hvað er að fréttu af Einari A: nei nei ég kem bara með þér ##

Í Íslenskri orðabók er *heyrdur* annars vegar lýst í orðasambandinu *heyrdur mig* (eða *mér*) ‘komdu og talaðu við mig’ og hins vegar sem „upphróp-un til áherzlu, til að vekja athygli eða til að láta í ljós hrifningu“ eins og það var orðað í ÍO-1963. Orðasambandið *heyrdur mig/mér* kemur alls ekki fyrir í talmálsefninu en síðari skýringin felur greinilega í sér við-leitni 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 *rosa-legur* ‘svakalegur, ofsalegur (m.a. um fólk)’. Í textasafni Orðabókar-innar 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 ritmálstextun-um í safninu að bloggginu frátöldu, enda þótt fjöldi texta úr ritmáli sé margfaldur á við talmálið. Heimilda um merkingu og notkun atviks-orðsins 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
 11
 12 fannst þetta
 13 þetta var
 14 keypti svo
 15 jájá það er
 16 ##### mikið
 17 ##### á alveg

rosalega klár í meira kaffi ## #### B: #### A: já mikið
 rosalega er ég fegin í fermingarveislum hérna <|B1> mér
 rosalega skrítið að kunni ekki við þetta ## hm ##### en
 rosalega flint ## nema B: nei ég held ekki A: nei ## ég
 rosalega flotta 1 1:1 sjö hundruð eða sjöþúsundkall ha A:
 rosalega <C>flint</C> minn ekki svona niður B: já ##
 rosalega var mikið af við fórum út að borða í gærkvöld
 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 þekktu 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 endurspeгла 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 mál-aðstæðna og hlutverks orðsins ekki síður en merkingar. Mörg dæmi-gerð talmálsorð kalla einmitt á slíka lýsingu og við samanburð útgáf-anna hefði óneitanlega verið forvitnilegt að hafa aðgang að gömlu tal-málsefni til samanburðar við það nýja.

Ritstjórar orðabókarinnar hafa tæplega haft aðgang að beinum heimildum um talmálið og hafa því þurft að styðjast við eigin mál-kennd 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 orðabók-inni en Þórunn nefnir einnig notkunarafrbrigði 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 orðabókgreininni.

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 gamal-gróið 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* mann-inn 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 boðhált sagnarinnar (jafnvel þótt uppfletti-orðið *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 talmálgögnum 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 getur þvegið glugga en þess að
2 hafa frontpage sjáfur
3 rosaalega klár því hann var ofsalega montinn
4 dauð C: það er hætt við því D: en ég er
5 bundið A: tímabundið út af því að nú þurfa þau
6 svona fílingur þarna (.) út af því að ég get ekki
7 úr vinnunni A: út af því að hann var langur
8 leitar fyrst í því B: ((tungusmellur)) (meta tæknið)
9 í minnum alveg frá því að ég smakkaði hana
10 eða B: nei A: nokkuð í því B: en í rauninni
11 en það er búið að breyta því núna er bara kennt
12 sem sagt maður hefur ekki neitað því upp á upp á
13 verður sjálfstætt nálegt því eða gæti orðið það
14 barnabætur eða hvað B: af því að hún er orðin
15 ((sigur upp í nefni)) af því að þetta (x) alveg
16 A: það er líka af því að það er röddunin
17 ef að það eru einhverjar líkur á því að það standist ekki
18 hafði verið langt á milli þeirra því að <IAG> (x) saman
19 aftur í hana draslið A: nei ég nenni því ekki (.) alls ekki
20 þurftu auðvitað að vera oft að þvo D: nei náttúrulega líka
21 og svo kannski hekk þvottunin upp á snúru í
22 veist að (.) vilji er fyrir D: =ja þú sjáir það alla vega
23 D: já sko það verður náttúrulega ekki í niunda bekk
24 er eitthvað sem þau horfa fram eru standa frammi fyrir
25 að (.) <C> (.) fannst þetta <C> alveg (fár-) fánlegt
26 ja það er það er spáð leiðindaveði núna á pásunum
27 því að það var herna bara svört stýðin í B: (x) B:
28 svarta B: en svo aftur um (sv-) svona um fimm sexdeytir
29 A: já (.) B: og svona lægningur og og yfir dalinn
30 nei A: nú var hún í órætti sem sagt B: jájá þillinn kemur

sko nota nokkuð efni A: er það LÍ:Si: ur
sko A: já A: já af því að annars getur ekki sko þú getur sett þau inn á síðuna þína B: jájá
sko yfir því að hún hafi komist inn og herna en hann ætlað bara að f
sko þú veist en ég hugsaði með mér ég ætla ekki að vera svo lengi í R
sko B: losa þau þau þurfa að fara þau ætla ekki að vera hér sko þ
sko ég á mitt torg B: m A: og ég veit ekki hvernig ég á að þú veist sé
sko þess vegna komstu svona miklu í verk ((hlær)) B: hann var svo ros
sko já að það var það sem ég var að experimenta með þú veist alv
sko B: ((hlær við)) B: já A: hún gerði tveir sko og það var önnur sem va
sko er hefur þetta hálfa starf ekki verið kennslan hún hefur verið aukavin
sko þú þarft að þekkja mismunandi framburð eftir landshlutum D
sko þetta hefur kitið mann að gera þetta C: já C: en það herna Harpa
sko B: takt þú þetta blað herna (.) ef að hann fer að einhver herna eitthv
sko komin yfir þrúgt A: já B: hún er ómenntuð (.) og með (f-) fjögur
sko bara skari og mjög slæmt ((sigur upp í nefni)) að skíða krakkarnir
sko af því að í þýsku og dönsku að þá er raddað ef það kemur herna n
sko (.) og alls ekki í skammtíma huggun eða einhverju svoléiðis
sko en Einar Jóhann hafði leift sér að ((dýr öpnast, blístur hættir)) að k
sko (.) enda þarf ég á þessu litla sem að er í (.) ((skark)) vélinni að hald
sko því þetta þetta er gamalt fólk þannig að þau þvo ekki eins mikl
sko tölf daga D1: (oj) D2: kominn skíðafylla af honum D3: já D: já end
sko alla vega sé ég það núna að (.) sko við verum að fá niunda bekk til okkar
sko það er ekki veija sér ekki náttúrufræðibraut núna sko A: en þetta er þetta e
sko þegar þau koma í tíunda bekk= D: =koma í tíunda <A> bekk sko <A> A: fyrir </D>
sko = D: m C: þetta er (ó-) D: =já C: en þetta er alveg óhæmulega vitlaust að lá
sko A: nú er það B: já A: já (.) B: <A> gæti orðið einhver (norð-) norðannep
sko eftir matinni í gær <IAG> þá herna eða svona um (nónley-) nónleytið að þá hér
sko þá var orðið finasta veður herna A: já (.) það var eins og í morgun þá var s (istal2)
sko A: já ((geisspar)) (.) jájá B: þannig er nú málið (.) A: heimsækirðu aldrei M (.
sko hann heldur áfram og þillinn kom bara beint A: jájá B: beint á (.) A: já það

4.3 Talmálgögn og orðabókagerð

Af umfjölluninni í kafla 4.2 og dæmunum sem þar eru rakin má ljóst vera að greiður aðgangur að talmálfefni, 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 talmálgögn 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 boðhátarmyndarinnar *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 rosaalega* + *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, merkingar-afbrigðum, orðasamböndum og notkunardæmum. Þær geta líka verið til mikils stuðnings við efnisskipan í orðabókum, ekki síst í inn-skipan 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 upplýsingapáttum 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 til-
liti 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 tal-
máli 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 hlut-
verk í rannsóknum og ýmsum hagnýtum verkum að þau endurspegli
raunverulega málnotkun, t.d. á tilteknum tíma eða á ákveðnu notk-
unarsviði. 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 síð-
ur 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 að-
ferðir við greiningu og nýtingu safnsins (sjá nánar um greiningar- og
leitaradferðir í greinum Sigrúnar Helgadóttur (2007) og Eiríks Rögn-
valdssonar (2007) í þessu hefti).

Annað mikilvægt atriði varðar viðhald og eflingu slíkra safna. Mál-
heild sem er ætlað að endurspeglar 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 skipu-
leggja 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 ým-
is sérhæfð verkefni munu eftir sem áður kalla á sjálfstæða efnisöfl-
un. 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 bær-
ust 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 nýttist 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 Føroya.
- 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. Möður 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 Ingólfssdóttir 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 stytta 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