

Kristín Bjarnadóttir

Beygingarlýsing íslensks nútímamáls

Regluverk eða beygingardæmi

1 Inngangur

Beygingarlýsing íslensks nútímamáls (BÍN) er gagnasafn um beygingarkerfið í íslensku nútímamáli. Gagnasafnið hefur nú að geyma ríflega 270 þúsund beygingardæmi með u.þ.b. 5,8 milljón beygingarmyndum, þ.e. orðmyndum með greiningarstrengjum eða mörkum.¹ Beygingarlýsingunni var upprunalega ætlað að þjóna tvíþættum tilgangi, að vera gagnasafn um beygingarmyndir til nota í málþækniverkefnum og til þess að veita almenningi aðgang að beygingardæmunum.

Vinna við BÍN hófst árið 2002 og var hluti af tungutækniátaki menntamálaráðuneytisins (Eiríkur Rögnvaldsson o.fl. 2009). Fyrsta útgáfa BÍN til málþækninota var gefin út á geisladiski í mars 2004 og vefsetrið var opnað síðar sama ár. Gögn úr BÍN hafa verið notuð í ýmis málþækniverkefni, t.d. leitarvélar, mörkun og leiðréttingarhugbúnað, og í kennsluefni og orðabókum. Beygingarlýsingin og *Mörkuð íslensk málheild* (MÍM) (Sigrún Helgadóttir o.fl. 2012) eru systurverkefni sem unnin hafa verið samtímis á Stofnun Árna Magnússonar í íslenskum

¹ Mark er strengur af skammstöfunum fyrir málfræðilegar formdeildir, þ.e. orðflokks- og beygingargreiningu.

fræðum. Upprunaleg gögn sem notuð voru í BÍN voru orðabókar-gögn, aðallega úr rafrænni útgáfu af *Íslenskri orðabók* (Mörður Árnason 2000), en þar voru u.þ.b. 135 þúsund flettiorð, og úr gagnasöfnum Orðabókar Háskólans. Leitað hefur verið í ýmsar aðrar heimildir um orðaforða en næsta stig er að bæta orðaforðanum úr MÍM við.

Frá upphafi var ætlunin að sýna íslenska beygingarkerfið eins og það er í raun, með eins góðri lýsingu á beygingarabrigðum og nokkur kostur er. Vegna þessa var strax ákveðið að búa til safn beygingardæma þar sem hver beygingarmynd er geymd og sýnd á sínum stað í beygingardæminu, í stað þess að setja saman virkt kerfi beygingarreglna þar sem orð eru merkt eftir því hvaða regla á við í hverri formdeild. Meginástæðan fyrir þessu var gagnaskortur en í ljós kom að nauðsynleg gögn til að setja fram slíkt reglukerfi voru einfaldlega ekki til þrátt fyrir langa sögu rannsókna á íslensku máli.

Vandamálið er að reglukerfi sem sett er fram á grunni ónógra gagna um beyginguna er bæði of- og vanvirkt, þ.e. það framleiðir mun fleiri orðmyndir en æskilegt er (t.d. fleirtölumyndina *sykrar* af karlkynsorðinu *sykur*) og sleppir öðrum (t.d. þágufallsmyndinni *hönd* í merkingunni *rithönd*). Til gagnagreiningar og til nota í leitarvélum kemur ofvirkni ekki verulega að sök þar sem tilbúnar orðmyndir koma einfaldlega aldrei fram en vanvirkni verður til þess að tækar orðmyndir verða útundan, t.d. í leit og greiningu, þ.e. þær verða að óþekktum orðum sem eru eitt af helstu vandamálum í textagreiningu. Ofvirkt reglukerfi er líka ótækt til framleiðslu á beygingarmyndum, a.m.k. ef útkoman á að vera rétt beygingardæmi. Þegar slík gögn eru notuð í máltækniþúnaði sem skilar frá sér texta verður niðurstaðan ekki góð, ef tilbúnar orðmyndir koma þar í stað þess sem rétt er. Þetta atriði er mjög mikilvægt, t.d. í ljósi þess að gögn úr BÍN eru notuð til samhengisháðrar leiðréttingar. Þá gera notendur BÍN á vefnum kröfu um að beygingardæmin sem birtast á síðunni séu rétt og skili sér ekki sem afrakstur af ofvirkum reglum.²

2 Flækjustigið í íslenska beygingarkerfinu

Hlutfallið á milli fjölda beygingardæma og beygingarmynda í BÍN

² Notendur BÍN á vefnum eru mjög duglegir við að senda fyrirspurnir um vafaatriði í beygingardæmunum og leiðréttingar, þar sem þess er þörf. Í maí 2013 voru heimsóknir á síðuna 53.907 og flettingar 211.902. Notendur voru 20.301, þar af 79% innanlands.

sýnir hve margar beygingarmyndirnar eru. Beygingarmyndir nafnorðs eru allt að 16, án afbrigða, beygingarmyndir lýsingarorðs eru allt að 120, og sagnorðs allt að 107, ef hvergi eru göt í beygingardæmunum.³ Hverri einstakri beygingarmynd í BÍN fylgir greiningarstrengur eða mark, t.d. NF-ET fyrir nefnifall eintölu af nafnorði og GM-NH fyrir nafnhátt sagnar í germynd. Markamengið fyrir íslensku er stórt og í því eru yfir 700 mörk (Sigrún Helgadóttir 2012), enda eru beygingarmyndirnar margar.

Beygingarkerfið er líka talsvert flókið, þar sem beygingarendingar eru ekki einræðar. Sama beygingarendingin getur verið birtingarvísir fyrir fleiri en eina formdeild og í fleiri en einum orðflokki, og sama formdeildin getur birst sem meira en ein beygingarending. Beygingarendingin *-a* er t.d. birtingarvísir margra beygingarformdeilda í öllum orðflokum (Eiríkur Rögnvaldsson 1990:23) og endingin í eignarfalli eintölu af sterkum karlkynsorðum með nefnifallsendinguna *-ur* getur verið *-s* eða *-ar*, t.d. *hundur*, *hunds*; *köttur*, *kattar*.

Í umfjöllun um íslenska beygingarkerfið er almennt litið svo á að orð flakki á milli beygingarflokka þegar tilbrigði eru í beygingarmyndum. Sögnin *fela* er þannig ýmist veik (þt. *faldi*) eða sterk (þt. *fól*) og nafnorðið *grein* er ýmist *greinar* eða *greinir* (1. og 2. flokkur kvenkynsnafnorða, skv. Valtý Guðmundssyni 1922:62). Við gerð BÍN var sú leið farin að setja upp beygingardæmi fyrir hvert orð fyrir sig í einu lagi, þannig að í hverju beygingardæmi birtust öll afbrigði í einu sem hluti af heildarmynstri fyrir orðið. Hvert beygingardæmi er því ein heild en ekki tilvísun í marga beygingarflokka eftir því sem afbrigði gefa tilefni til. Beygingardæmin í BÍN byggjast samt á nefnimyndum (eða uppflettimyndum) þannig að afbrigði í nefnimynd skilja á milli beygingardæma. Orðin *sannleikur* og *sannleiki* eru t.d. tvö beygingardæmi, þó að málnotendur blandi beygingunum saman á mjög áhugaverðan hátt.⁴ Í hverjum beygingarflokki í BÍN eru öll afbrigði sem birtast í orði tekin með og því er beygingarflokkur einungis mengið af þeim beygingarreglum sem þarf til þess að búa til beygingardæmi fyrir orð, eitt eða mörg, eftir því sem verkast vill. Þessi regluþrungi sem standa að baki beygingarflokkunum hafa í sjálfu sér ekki fræðilegt gildi (sjá Ásta Svavarsdóttir 1993) en þau eru tól til að halda utan um flókin gögn.

³ Þarna vantar samt spurnarmyndir sagna, t.d. *ferðu*, *fórstu* o.s.frv.

⁴ Tíðnitölur benda til þess að sterka beygingin sé notuð í nefnifalli en veika beygingin í aukaföllum: *sannleikurinn er sagna bestur*; í *sannleika* sagt.

3 Keyrslan á beygingardæmunum

Upprunalega voru beygingardæmin í BÍN búin til með því að nota tiltölulega einfaldar skriftur þar sem númeraðir orðstofnar í færslum fyrir einstök orð voru keyrðir inn í ramma með beygingarendingum, þ.e. beygingarflokk. Niðurstaðan var safn af XML-skrám sem síðan voru notaðar til að búa til það úttak sem þörf var fyrir, t.d. til birtingar á vefnum eða til nota í máltækniverkefni. Gagnagrunnurinn sem nú hýsir BÍN virkar á svipaðan máta. Færslan um hvert orð geymir öll stofnbrigði og upplýsingar um það hvar í beygingardæminu hvert stofnbrigði á sér stað. Nafnorð eru merkt eftir því hvort þau koma fyrir í eintölu, fleirtölu eða hvort tveggja, sagnir eru merktar eftir því hvort þær koma fram í germynd, miðmynd eða hvort tveggja o.s.frv.⁵ Hver beygingarmynd er síðan geymd í gagnagrunninum með sínu marki, t.d. *akur*<þgf.et.>, sjá Töflu 1.

Án greinis (+)				
	Eintala (+)			Fleirtala (+)
Nf.	1+0	<i>akur</i>	3+ar	<i>akrar</i>
Þf.	1+0	<i>akur</i>	3+a	<i>akra</i>
Þgf.	3+i	<i>akri</i>	2+um	<i>ökrum</i>
Ef.	1+s	<i>akurs</i>	4+a	<i>akra</i>
Með greini (+)				
Nf.	1+inn	<i>akurinn</i>	3+arnir	<i>akrarnir</i>
Þf.	1+inn	<i>akurinn</i>	3+ana	<i>akrana</i>
Þgf.	3+inum	<i>akrinum</i>	2+unum	<i>ökrunum</i>
Ef.	1+sins	<i>akursins</i>	4+anna	<i>akranna</i>

Tafla 1. Ramminn fyrir nafnorðið *akur*.

Í Töflu 1 er ramminn fyrir nafnorðið *akur*, en þar eru engin göt í beygingardæminu og engin afbrigði. Gildin fyrir beygingarformdeildirnar eru því alltaf '+', þ.e. eintala = +et., fleirtala = +ft., og ákveðni = +ángr., +mgr. (Í fleirtöluorði væri gildið fyrir eintöluna -et. og þá verður sá hluti beygingardæmisins tómur.) Stofnbrigðin *akur* (1), *akr* (3 og 4),⁶ *ökr* (2) eru tölusett og eru hluti af inntaki fyrir orðið. Á vefsíðunni birtist beygingardæmið á mjög svipaðan hátt og í Töflu 1 en 2. og 4. dálki er sleppt.

Alengasta úttak á gögnum úr BÍN til máltækninota er einfaldur

⁵ Nákvæmar upplýsingar um rofana sem ráða birtingu beygingarmynda er að finna á vefsíðu BÍN: <http://bin.arnastofnun.is>.

⁶ Sama stofnbrigði fær af tæknilegum ástæðum stundum fleiri en eitt númer.

listi með sex sviðum fyrir hverja beygingarmynd, eins og í úttakinu fyrir nafnorðið *akur* í Töflu 2. Sviðin eru uppfllettiorð, auðkennisnúmer, orðflokkur eða kyn nafnorðs, hluti orðaforðans í BÍN⁷ (hér 'alm', fyrir 'almennt mál'), beygingarmynd og málfræðilegt mark eða greiningarstrengur.

akur;472164;kk;alm;akur;NF.ET.
 akur;472164;kk;alm;akurinn;NF.ET.MGR.
 akur;472164;kk;alm;akur;PF.ET.
 akur;472164;kk;alm;akurinn;PF.ET.MGR.
 akur;472164;kk;alm;akri;PGF.ET.
 akur;472164;kk;alm;akrinum;PGF.ET.MGR.
 akur;472164;kk;alm;akurs;EF.ET.
 akur;472164;kk;alm;akursins;EF.ET.MGR.
 akur;472164;kk;alm;akrar;NF.FT.
 akur;472164;kk;alm;akrarnir;NF.FT.MGR.
 akur;472164;kk;alm;akra;PF.FT.
 akur;472164;kk;alm;akrana;PF.FT.MGR.
 akur;472164;kk;alm;ökrum;PGF.FT.
 akur;472164;kk;alm;ökrunum;PGF.FT.MGR.
 akur;472164;kk;alm;akra;EF.FT.
 akur;472164;kk;alm;akranna;EF.FT.MGR.

Tafla 2. Úttak úr BÍN til máltækninota.

Rammarnir (sbr. Töflu 1) eru búnir til eftir því sem þörf krefur, í hvert skipti sem nýfundid afbrigði passar ekki í þá ramma sem þegar eru til, þ.e. þegar knippi af beygingarreglum dugar ekki fyrir öll afbrigði einhvers orðs. Vegna þessa er BÍN byggð á beygingu einstakra orða, ef í það fer, þ.e. kerfið er búið til neðan frá ('bottom-up'). Dæmi um þetta eru samsett orð sem hafa hvorugkynsorðið *vegur* sem síðari hluta. Ósamsetta orðið *vegur* er oftast *vegar* í eignarfalli eintölu, þó að beygingarmyndinni *vegs* bregði einnig fyrir, t.d. í orðasambandinu *komast til vegs og virðingar*. Í sumum samsetningum af orðinu er eignarfallsendingin *-s* hins *vegar* ráðandi, t.d. í orðunum *farvegur*, *jarðvegur* og *útvegur*. Orðin *vegur* og *akvegur* annars *vegar* og *farvegur*, *jarðvegur* og *útvegur* skiptast því í tvo beygingarflokka, ef. et: *-ar/-s* og *-sl/-ar*. Þarna er ráðandi ending á undan en röðin á beygingarafbrigðum í BÍN ákvarðast yfirleitt af tíðni þó að þar geti margir aðrir þættir komið við sögu.

Beygingarflokkar í BÍN eru nú yfir 630 og í sumum þeirra eru tugþúsundir orða en í öðrum eru stök orð. Þetta stafar stundum af

⁷ Helstu „hlutar BÍN“ eru almennt mál, mannanöfn og örnefni og sérorðaforði ýmiss konar.

Því að gamlar beygingarmyndir koma enn fyrir í nútímamáli, þó svo að beygingarreglurnar sem að baki liggja séu ekki lengur virkur hluti af beygingarkerfinu. Úreltu beygingarmyndirnar eru þá gjarnan samhengisháðar í nútímamáli. Sem dæmi má nefna beygingarmyndina *refarnir* í orðasambandinu *til þess eru refarnir skornir* í stað *refirnir* sem er nefnifall fleirtölu í nútímamáli og beygingarmyndina *skjöldu* í sambandinu *ganga fram fyrir skjöldu*, í stað *skildi* sem er nútímamálsmynd þolfalls fleirtölu. Orðmyndirnar *refar* og *skjöldu* eru strangt til tekið ekki hluti af beygingarkerfi nútímamáls en þær koma vissulega fyrir í textum og eru því bráðnauðsynlegar í BÍN, bæði til máltaekninota og á vefnum. Afbrigði af þessu tagi fjölga beygingarflokkum í BÍN þar sem hvert einasta afbrigði á sér stað í beygingarflokknum og breytir skilgreiningunni á honum.

Uppflettiorð á bilinu	Beygingarflokkar	Uppflettiorð fleiri en	Beygingarflokkar
<10.000	5		
5.000-9.999	7		
1.000-4.999	29	<1000	41
500-999	21	<500	62
100-499	72	<100	134
50-99	47	<50	181
10-49	126	<10	307
2-9	146	<2	453
1	156		Alls: 609

Tafla 3. Stærð og fjöldi beygingarflokka nafnorða, lýsingarorða og sagna.

Beygingarflokkar í BÍN er misjafnlega stórir, sjá Töflu 3, en athyglisvert er að í u.þ.b. fjórðungi beygingarflokka meginorðflokka, nafnorða, lýsingarorða og sagna, er aðeins eitt orð. Í u.þ.b. 42% þessara beygingarflokka er beygingin alveg „regluleg“, þ.e. þar er aðeins ein regla um hverja formdeild. Þar eru engin afbrigði af beygingarmyndum, engin innri beyging (þ.e. beyging fyrri hluta, t.d. í *litliputti*, þf. *litlaputta*) eða afbrigðileg göt í beygingardæmunum, eins og t.d. í kvenkynsorðinu *kröggur* en þar hafa engin dæmi fundist um eignarfall fleirtölu.⁸ Í BÍN eru 49.296 afbrigðapör (t.d. eignarfallið *þröskuld-ar/þröskulds*) og 262 afbrigðaprenndir (t.d. þolfallið *Berglind/Berglindi/Berglindu*), þ.e. afbrigði með sama greiningarstreng innan beygingardæmis (feb. 2012).

⁸ Þeir sem spurðir eru um eignarfallsmyndina neita yfirleitt að svara, jafnt málfræðingar sem aðrir.

4 Gagnaskortur

Upprunalegu gögnin í BÍN voru orðabókarefni, úr rafrænni útgáfu *Íslenskrar orðabókar* (2000), en hún kom fyrst út 1963, og úr Ritmálssafni Orðabókar Háskólans, með milliskrefi í *Norræna verkefninu* svokallaða (Kristín Bjarnadóttir 1998). Önnur meginheimild í upphafi var bókinn *Nöfn Íslendinga* (Guðrún Kvaran og Sigurður Jónsson frá Arnarvatni, 1991), en í henni eru u.þ.b. 4.800 mannanöfn. Í fyrstu útgáfu BÍN (2004) voru u.þ.b. 176 þúsund beygingardæmi og orðaforðinn var aðallega úr þessum þremur heimildum. Viðbótarorðaforði í síðari útgáfum BÍN er úr ýmsum áttum og hann er m.a. afrakstur samvinnu þar sem gögn úr BÍN hafa verið notuð í leitarvélar, t.d. á vef Hins íslenska biblíufélags⁹ og í símaskránni¹⁰, sem er góð uppspretta nafna, örnefna og fyrirtækjaheita.

Upplýsingar um beygingu orða er að finna í *Íslenskri orðabók* og í *Nöfnum Íslendinga* en sú lýsing er ekki alltaf nægileg til þess að setja fram heil beygingardæmi. Í öðrum heimildum eru yfirleitt engar upplýsingar um beygingu. Það var því ljóst frá upphafi að leita þyrfti í málfræðirit til að finna frekari upplýsingar en síðar kom í ljós að meira að segja þar voru gögn brotakennd og ófullnægjandi.

Hefðin í íslenskum orðabókum er að gefa kenniföll nafnorða (eignarfall eintölu og nefnifall fleirtölu) og kennimyndir sagna (þrjár fyrir veikar sagnir og fjórar fyrir sterkar sagnir) (sjá t.d. Eiríkur Rögnvaldsson 1998 og Kristín Bjarnadóttir 2006). Ýmist eru endingarnar einar látnar duga (*bátur*, kk. -s, -ar) eða beygingarmyndir skrifaðar fullum stöfum þar sem stofnbrigði koma fram (*köttur*, kk. *kattar*, *kettir*). Beygingarmyndir aðrar en kenniföll og kennimyndir eru yfirleitt ekki sýndar nema í undantekningartilfellum, þ.e. þegar beyging þykir óregluleg. Samt er beygingarkerfinu þannig háttað að aðrar beygingarmyndir eru ekki endilega fyrirsegjanlegar út frá þessum upplýsingum og nægir þar að nefna þágufall eintölu í sterkum karlkynsorðum (Friðrik Magnússon 1984). Þágufall af *bátur* er t.d. ýmist *báti* eða *bát* en með greini er þágufallið *bátnum* en ekki **bátinum*. Ófyrirsegjanleiki af þessu tagi kemur fram í lýsingarorðum og sögnum líka, í mismiklum mæli eftir stofngerð. Lýsingu á beygingu lýsingarorða er þröngur stakkur skorinn í íslenskum orðabókum og henni er ekki sinnt nema í undantekningartilvikum, þ.e. þar sem

⁹ <http://biblián.is>

¹⁰ <http://ja.is>

óreglan þykir vera veruleg. Niðurstaðan um upplýsingar um beygingu orða í orðabókum er því að hún sé brotakennd en á móti kemur að orðaforðinn er verulegur og að því leyti eru orðabækur góðar heimildir í BÍN.

Í málfræðiritum, sérstaklega í yfirlitsritum, er annar háttur á. Þar er greinargóð lýsing á beygingarkerfinu sem slíku, með beygingardæmum um valin orð. Markmiðið er að lýsa kerfinu sjálfu sem heild og nota til þess viðeigandi dæmi.¹¹ Rannsóknarsagan er orðin býsna löng; fyrsta umfjöllunin er í *Grammaticæ islandicæ rudimenta* eftir Jón Rúgmann sem kom fyrst út í Kaupmannahöfn 1651 (sjá Guðrún Kvaran 2005). Hefðin er líka sterk og segja má að Rask hafi mótað aðferðina í *Veiledningen* 1811. Sett er fram kerfi vel afmarkaðra beygingarflokka og tekin dæmi um þau. Sömu dæmin eru gjarnan notuð aftur og aftur þannig að orðaforðinn er mjög lítill, þó svo að allar heimildir séu lagðar saman.¹²

Málfræðiritin eru ekki góð heimild um beygingu einstakra orða, nema þar sem svo vill til að orðið er tekið sem dæmi. Megináhersla er á erfðaorðaforðann og dæmin eru fá. Vegna þessa verða sumir hlutar orðaforðans út undan, t.d. tökuorð, talmálsorð og slangur, jafnvel þó að þar sé sennilega helst von á nýbreytni í beygingum. Til dæmis má nefna beygingarmyndir í þágufalli eintölu í sterkum hvorugkynsorðum. Þar eru málfræðibækur sammála um að endingin *-i* sé altæk, utan þriggja eða fjögurra orða, þ.e. *fé*, *hné* (*kné*) og *tré*.¹³ Í nýlegri handbók um beygingar- og orðmyndunarfræði (Guðrún Kvaran 2005) koma aðeins fram tvær *i*-lausar myndir. Þágufallið *tré* er sýnt í beygingardæmi ((20–43)d, bls. 242) og þar fylgir athugasemd: „Aðeins *hné* beygist eins og *tré*“. Í yfirliti um beygingu sterkra hvorugkynsorða segir svo: „Þágufall eintölu sterkrar beygingar endar alltaf á *-i*.“ (bls. 244).

Í vinnunni við BÍN kom fljótlega í ljós að undantekningarnar eru mun fleiri þar sem ósamsett fleirkvæð hvorugkynsorð eru ýmist end-

¹¹ Undantekningin er kandiðatsritgerð Ástu Svavarsdóttur (1993) en þar er beygingu nafnorða lýst, eftir tíðnitölum, og stuðst við tiltekið gagnasafn. Annars er viðfangsefnið oftast beygingarkerfið í heild eða tilteknir flokkar innan þess.

¹² Yfirgrípsmesta og gagnlegasta beygingarlýsingin við vinnuna við BÍN reyndist vera bók Valtýs Guðmundssonar *Íslandsk grammatik* (1922), sérstaklega um beygingu sagna, en verkið hefur samt sem áður sömu einkenni og önnur verk af sama toga; þar er gerð grein fyrir meginflokkum og undantekninga getið í viðbótargreinum.

¹³ Hjá Valtý Guðmundssyni er orðið *hlé* haft þarna með. Í nútímamáli virðist það yfirleitt fá endinguna *-i* í þágufalli.

ingarlaus eða fá -i í þágufalli eintölu. E.t.v. ráða málhreinsunarsjónarmið því að slíku efni er sleppt í málfræðiritum en jafnframt verður áherslan á kerfið sem heild til þess að afbrigði og frávik verða út undan. Flækjustigið verður óneitanlega meira þegar í ljós kemur að tökuorð haga sér ekki endilega alveg eins og erfðaorðaforðinn. Þetta gildir líka um rótgróin tökuorð, t.d. orð sem er að finna í textum frá því fyrir tíma Rasks skv. Ritmálssafni Orðabókar Háskólans, t.d. hvorugkynsorðin *alkóhól*, *flannel*, *fennel*, *vitriól* og *indígó* (sjá Kristín Bjarnadóttir 2010). Þágufallsafbrigðin af orðunum *fennel* og *arsenik* eru t.d. tvö í blaðatextum á *Tímarit.is*:

- (1) a Hvönn er svolítið lík fennel<þgf.et.>
b ...ásamt brytjuðu fenneli<þgf.et.>
- (2) a ...byrlað eitur, drepinn með arsenik<þgf.et.>
b Hann var myrtur með arseniki<þgf.et.>

Gagnaskorturinn birtist því með ólíkum hætti í heimildunum. Í orðabókum og orðabókarefni er mikill orðaforði en upplýsingar um beygingu einstakra orða eru ekki tæmandi. Í mállýsingum eru hins vegar ítarlegar upplýsingar um hluta orðaforðans, þ.e. lýsing á kerfinu sem slíku, en eins og dæmið um þágufalls-*i* sýnir er sú lýsing ekki heldur tæmandi. Vegna þessa var óhjákvæmilegt að safna meiri gögnum við gerð BÍN.

5 Rannsóknargildi BÍN

Gagnvart BÍN er einkum þörf á því að leita að ófyrirsegjanlegum beygingarmyndum en talsverð vinna fólst í því að fylla í þau göt sem hér hafa verið nefnd í heimildunum. Til þessa eru öll gögn notuð sem hægt er að komast yfir að skoða og á það bæði við um gögn innan stofnunar og utan. Þar er fyrst að nefna söfn Orðabókar Háskólans, sérstaklega Ritmálssafn, prentaðar orðabækur og rafræn textasöfn (t.d. *Íslenskt textasafn* og safn Landsbókasafns Íslands – Háskólabókasafns af textum úr blöðum og tímaritum).¹⁴ Mat á gögnum styðst við málkennd þess sem vinnur verkið og er hún stundum eina haldreipið þegar engin dæmi finnast. Þá er oft leitað til allra þeirra sem gefa færi á sér og notendur BÍN á vefnum senda líka inn athugasemdir og viðbætur. Í neyð er stuðst við leit á vefnum (með Google), þrátt fyrir alla

¹⁴ *Íslenskt textasafn*: http://www.arnastofnun.is/page/gagnasofn_textasafn; og *Tímarit.is*.

þá annmarka sem því fylgja, bæði við að meta heimildir og einnig vegna þess hve slík leit er tímafrek og ómarkviss. Þrátt fyrir þetta er niðurstaðan sú að sumar beygingarmyndir eru með öllu ófinnanlegar, alveg sama hvaða aðferðum er beitt.

Dæmi um þetta er þágufall eintölu af orðinu *Yggdrasill*. Í fornum textum hefur ekki tekist að finna eitt einasta dæmi um þágufallið sem hlýtur að vera annaðhvort *Yggdrasli* (sbr. reglulega beygingu karl-kynsnafnorða sem hafa endinguna *-ill* í nefnifalli, t.d. *bendill*, *brimill*, *depill*, þgf. *bendli*, *brimli*, *depli*) eða *Yggdrasil*, sem væri þá sambærileg undantekning og endingarleysið í þágufalli í *Huginn* (þgf. *Hugin*, ekki **Hugni*) og *Muninn* (þgf. *Munin*, ekki **Munni* (!)). E.t.v. eru það líkindin við hvorugkynsorðið *drasl* sem verða til þess að fólk virðist veigra sér við að nota þágufallsmyndina *Yggdrasli* og finnst hún jafnvel fyndin, en þágufallsmyndin *Yggdrasil* þykir ekki alveg nógu góð heldur.¹⁵ Flestir virðast því reyna að komast hjá því að nefna verslunina *Yggdrasil* í þágufalli og haga máli sínu á annan veg.

Eins og áður sagði er ekki alltaf hægt að finna dæmi um beygingarmyndir og þá er málkenndin ein til leiðsagnar og m.a. vegna þessa verður BÍN seint fullunnið verk. Þar er giskað á myndir þegar gögnin duga ekki til og þar sem heimildir þykja ekki traustar. Reynt er að leiðbeina notendum vefjarins með því að bæta inn athugasemdum um málnotkun ofan við beygingardæmin þannig að fólk geti sjálft myndað sér skoðun á efniinu. Í athugasemd við orðið *Yggdrasill* er t.d. tekið fram að heimildir um þágufallið séu ekki áreiðanlegar.

Það er m.ö.o. ekki svo að allar orðmyndir sem finnast eigi greiða leið í BÍN, þó að BÍN sé beygingarlýsing en ekki forskrift. Orðmyndir sem enginn íslenskur málnotandi tæki gildar koma fyrir, sérstaklega á vefnum, svo sem fleirtölumyndin *fótarnir* í stað *fætarnir* af karl-kynsnafnorðinu *fótur* og þágufallsmyndin *fótnum* í stað *fætinum*, og þær eiga ekki heima í BÍN. Venjulega er hægt að útiloka slíkar myndir með því að styðjast við tíðnitölur, a.m.k. í algengum orðum, en á endanum er það ákvörðun þess sem vinnur verkið hvaða beygingarmyndir á að sýna.¹⁶ Til málækninota, sérstaklega til greiningar, er kostur að taka sem allra flestar beygingarmyndir með, þar sem fjöldi óþekktra orðmynda minnkar við það. Það er hins vegar jafnvægislist að meta hvaða beygingarmyndir eru afbrigði og hverjar eru villur, t.d. þegar

¹⁵ Á vefnum finnast dæmi um báðar myndir en dæmi um *Yggdrasli* eru helst í skrifum um þágufallið sjálft.

¹⁶ Mörkin á milli afbrigðis af beygingarmynd og villu eru stundum ljós og stundum ekki en í öllum textum leynast villur.

nota á gögnin í leiðréttingarhugbúnað. Lausnin í BÍN er að styðjast við tíðnitölur og málkenndina. Orðmyndum sem koma sárasjaldan fyrir, jafnvel í ótraustum heimildum, og þykja að auki verulega ankannalegar er því sleppt. Þetta er of frjálst fyrir þá sem vilja að verkið sýni málið aðeins eins og það á að vera „að bestu manna vitund“ en of íhaldssamt fyrir þá sem vilja fá allt með, ef það hefur á annað borð sést í málinu. Þar sem BÍN er ætlað að þjóna tveimur herrum, málækni og almennum notendum, verður að fara bil beggja. Sérstök ástæða er til að fara varlega vegna þess að notendur vefsíðunnar virðast oft gera ráð fyrir því að BÍN sé vottunaraðili um það sem talið er rétt mál, þrátt fyrir að í athugasemdunum með beygingardæmunum sé oft tekið fram að tiltekin afbrigði séu ekki í samræmi við núgildandi stafsetningu eða stefnu stífustu málræktarmanna.¹⁷

Lýsingin á beygingarkerfinu í BÍN takmarkast af því að leitað er að einstökum beygingarmyndum eftir því sem ástæða þykir til en ekki viðhöfð allsherjar orðmyndaleit í ómörkuðum texta, enda er slík aðgerð óframkvæmanleg. Ástæðan er m.a. sú að margræðni beygingarmynda í íslensku er gríðarleg, bæði innan beygingardæmis og á milli þeirra. Sem dæmi má nefna orðmyndina *minni* sem kemur fram í fjórum orðum, alls í 36 beygingarmyndum.¹⁸

Margræðnin er augljós af tölum um beygingarmyndir og orðmyndir í BÍN. Þar eru 5,8 milljón beygingarmyndir (þ.e. orðmyndir með marki eða greiningarstreng) en orðmyndir eru 2,8 milljónir. Af þeim eru 1,8 milljón orðmyndir sem bara koma fyrir á einum stað, þ.e. ótvíræðar eða einræðar orðmyndir. Milljón orðmyndir eru því margræðar. Í Töflu 4 eru tölur um tvíræðni beygingarmynda en þar eru tölurnar hærri þar sem beygingarmynd er orðmynd með marki eða greiningarstreng.

¹⁷ Í BÍN er ætlast til þess að notendur myndi sér skoðun sjálfir. M.ö.o. er spurningunni um það hvað er gott og rétt ekki svarað í BÍN. Til þess skortir einfaldlega rannsóknir.

¹⁸ Orðmyndin *minni*: miðstig af lýsingarorðinu *lítill* (20 beygingarmyndir), sögnin *minna* (10 beygingarmyndir), hvorugkynsnafnorðið *minni* (5 beygingarmyndir) og svo þágufall eintölu, kvenkyn af eignarfnafninu *minn*.

Beygingarmyndir í BÍN	5.881.374	
Einræðar beygingarmyndir	1.850.090	31,5%
Margræðar innan beygingardæmis ¹⁹	3.619.482	61,5%
Margræðar milli beygingardæma ²⁰	63.641	1,1%
Margræðar innan og milli beygingardæma ²¹	348.161	5,9%

Tafla 4. Margræðni beygingarmynda í BÍN.

Vegna tvíræðninnar er markaður texti, eins og *Mörkuð íslensk málheild*, einstakt tæki til þess að finna dæmi um einstakar beygingarmyndir.

6 Samanburður við Markaða íslenska málheild

Í *Markaðri íslenskri málheild* (MÍM) eru u.þ.b. 25 milljón lesmálsorð en samanburðurinn á MÍM og BÍN sem hér er lýst byggist á hluta MÍM, u.þ.b. 17,7 milljónum lesmálsorða. Samanburðurinn var gerður með það í huga að endurbæta BÍN og auka.²² Talnastrengjum, táknum o.þ.h. var sleppt úr MÍM, þannig að lesmálsorðin úr MÍM sem borin voru saman við BÍN voru 16.245.429.

Lesmálsorð	16.245.429
Einræðar markaðar orðmyndir	737.856
Þar af í BÍN	425.238
Þar af ekki í BÍN	312.618

Tafla 5. Einræðar markaðar orðmyndir í samanburði MÍM og BÍN.

Þess ber að geta að efnið úr MÍM er ekki yfirfarið og leiðrétt en samanburðurinn fólst í því að keyra saman lista um beygingarmyndir (þ.e. orðmynd með marki) úr MÍM og sambærilegt efni úr BÍN.²³ Beygingarmynd úr MÍM fylgir nefnimynd (þ.e. uppflattmynd eða lemma) en talsvert er um villur í lemmuninni. Síðan var farið yfir niðurstöðuna á skjá, línu fyrir línu. Í samanburðinum var orðflokksgreining úr MÍM

¹⁹ Dæmi: Orðmyndin *penna* er þolfall, þágufall og eignarfall eintölu af nafnorðinu *penni*.

²⁰ Dæmi: Orðmyndin *burstanna* er eignarfall fleirtölu af nafnorðnum *burst* og *bursti*.

²¹ Dæmi: Orðmyndin *minni* er 34 beygingarmyndir af fjórum orðum, lo. *lítill*, so. *minna*, hk. *minni* og fn. *minn*.

²² Sumt af efniinu í þessum hluta MÍM verður í reynd ekki hluti af aðalsafni MÍM en kemur samt að fullum notum við vinnuna við BÍN. Mörkun á MÍM var heldur ekki lokið þannig að ýmislegt af því sem hér er lýst byggir ekki á endanlegri gerð MÍM. Hluti af afrakstri samkeyrslunnar var einmitt hugsaður til þess að koma að gagni við bæði verkin.

²³ Markamengin eru ekki eins en þau voru samræmd fyrir samkeyrsluna.

leiðrétt og lemmunin líka, í því efni sem telst vera íslenskur orðaforði.²⁴ Þess ber að geta að hugbúnaðurinn sem notaður er í lemmunina, svokallað lemmald (Anton Karl Ingason o.fl. 2008), erfir villur í orðflokkgreiningu úr mörkunum í MÍM en textinn er fyrst markaður og síðan lemmaður. Rangur orðflokkur í marki verður þess vegna til þess að sjálf nefnimyndin eða lemman verður óhjákvæmilega röng. Í yfirferðinni var efni sem ekki telst til íslensks orðaforða flokkað sérstaklega og merkt eins og komið verður að hér á eftir.

Alls var farið yfir 312.618 orðmyndir (eða markaða strengi) úr MÍM, þ.e. það efni sem ekki fannst í BÍN. Ríflega þriðjungur (34,5%) þessara orðmynda var rétt orðflokksmerktur og lemmaður í MÍM en hafði einfaldlega ekki ratað inn í BÍN úr öðrum heimildum. Í Töflu 6 er fjöldi beygingarmynda úr meginorðflokkunum þremur úr samanburðarefninu úr MÍM. Í 1. dálki er fjöldi beygingarmynda skv. mörkuninni í MÍM og í 2. dálki er fjöldi rétt markaðra beygingarmynda þar. Mismunurinn á 1. og 2. dálki sýnir ranga greiningu og/eða lemmun í MÍM.²⁵ Í 3. dálki eru tölur um beygingarmyndir í hverjum orðflokki eftir handleiðréttingu á samanburðarefninu. Þar er m.ö.o. niðurstöðutala um beygingarmyndir í hverjum orðflokki í því efni úr MÍM sem ekki finnst í BÍN, eftir leiðréttingu.

	Heildartala í MÍM	Rétt í MÍM	Niðurstæða
hk.	70.770	31.942	54.317
kk.	142.630	42.910	64.709
kvk.	78.908	46.207	64.499
lo.	23.835	11.713	18.345
so.	7.639	1.035	3.210

Tafla 6. Fjöldi beygingarmynda í MÍM, fyrir og eftir leiðréttingu.

Samanburðurinn á MÍM og BÍN var aðallega gerður til þess að finna íslenskar beygingarmyndir sem vantaði í BÍN. U.þ.b. 60% af öllum mörkuðum strengjum úr MÍM sem ekki fundust í BÍN reyndust vera íslenskar beygingarmyndir. Í yfirferðinni var afgangurinn, u.þ.b. 40%, flokkaður lauslega. Þar var erlent efni fyrirferðarmest (24,6%), en einnig má nefna villur (5,8%), skammstafanir (1,6%) og ýmiss konar tölvustrengi, vefföng, slóðir o.þ.h. (0,7%). Þetta efni skiptir ekki bein-

²⁴ Til dæmis má nefna mark og lemmun á beygingarmyndinni *Miklagarði* sem greind er sem þolfall eintölu af kvenkynsnafnorðinu *Miklagarð* í MÍM, í stað þágufalls eintölu af karlkynsnafnorðinu *Mikligarður*.

²⁵ Athugið að villufjöldinn á aðeins við efni sem ekki finnst í BÍN og hlutfallið gefur alls ekki mynd af heildarvillufjölda í greiningu og/eða lemmun í MÍM.

línis máli í vinnunni við BÍN en getur nýst sem ýtarefni í máltækni, t.d. í leiðréttingarbúnað o.þ.h., þar sem eitt af meginvandamálum í máltækni er greining á óþekktum „orðum“ eða stafastrengjum.

Fyrsta niðurstaða úr samanburðinum á MÍM og BÍN er sú að bæta þurfi 125 þúsund beygingardæmum við í BÍN. Meðalfjöldi orðmynda í beygingardæmi í BÍN er rétt yfir 20, þannig að alls gætu viðbótarorðmyndirnar orðið yfir 2,5 milljónir en heildarfjöldi beygingardæma yrði þá hátt í 400 þúsund og beygingarmyndir yfir 8,3 milljónir. Þetta eru bráðabirgðatölur þar sem bera þarf saman uppflettiorð í BÍN og MÍM og gáta lemmunina en því verki verður ekki lokið fyrr en allt efnið er komið inn í BÍN. Í vinnuferlinu verður gengið úr skugga um að allar beygingarmyndir í MÍM séu á sínum stað í BÍN og er það lokastigið í samanburðinum.

Vinna við viðbótarefnið í BÍN er hafin. Beygingarflokkur orða er fundinn með samanburði við efni sem fyrir er í BÍN, m.a. með því að nota greiningartól fyrir samsett orð.²⁶ Síðan er færslan fyrir hvert orð fullunnin, með því að merkja gildi á beygingarformdeildum sem koma eiga fram í hverju beygingardæmi, t.d. +/-et, +/-ft. o.s.frv., sbr. Töflu 1. Beygingarmyndir úr MÍM verða hafðar til hliðsjónar, ásamt efni úr öðrum heimildum eftir þörfum, sbr. það sem sagt er um gagnaoöflun hér að framan. Jafnframt verða beygingardæmi sem þegar eru í BÍN endurskoðuð eftir því sem viðbótarbeygingarmyndir úr MÍM gefa tilefni til.

7 Gagnaskortur

Samanburðarefnið úr MÍM er 17,7 milljónir lesmálsorða, alls 737.856 strengir sem teljast vera orðmyndir. Þar af eru beygingarmyndirnar sem hér eru til athugunar tæplega 623 þúsund (84%). Þetta er lítið brot af beygingarmyndunum sem þegar eru í BÍN, þ.e. 5,8 milljónum.²⁷ Þessar tölur gefa vísbendingu um það hve stór málheild þyrfti að vera til þess að hægt væri að nota hana til þess að setja fram beygingarlýsingu á borð við BÍN eða samsvarandi beygingarreglur um ís-

²⁶ Greiningartólið er verk Jóns Friðriks Daðasonar í samvinnu við Stofnun Árna Magnússonar í íslenskum fræðum og er hluti af meistaraverkefni hans um leiðréttingu á ljóslesnum texta (Jón Friðrik Daðason 2012).

²⁷ Af tæknilegum ástæðum er ekki enn gerlegt að bera saman heildarfjölda uppflettiorða í MÍM og BÍN þar sem gera verður ráð fyrir talsverðum fjölda af villum í lemmun í MÍM á orðmyndum sem einnig eru BÍN. Það efni hefur ekki enn verið athugað.

lensku. Beygingarlýsing sem aðeins byggði á dæmum úr MÍM væri ekki stór enda er beygingarmyndafjöldinn í MÍM innan við tíundi hluti af beygingarmyndafjöldanum í BÍN og orðaforðinn er aðeins lítill hluti af því sem venjulegur Íslendingur hefur á valdi sínu. Þetta vandamál kemur íslenskum orðabókarmönnum ekki á óvart, þar sem sams konar gagnapurrd eða gagnaskortur kemur fram þegar leitað er að beygingarmyndum m.t.t. setningarlegrar stöðu eða samhengis (Jörgen Pind o.fl. 1993).

Eftir fyrirhugaðar viðbætur við BÍN þar sem byggt er á samanburðinum við MÍM verða orðmyndir þar alls 8,3 milljónir. Úr 623 þúsund orðmyndum úr MÍM bætast því við 2,5 milljónir nýrra beygingarmynda í BÍN og beygingardæmin þar verða sennilega yfir 395 þúsund. Til samanburðar má nefna að í Ritmálssafni Orðabókar Háskólans eru yfir 700 þúsund uppflettiorð. Þar er að vísu sögulegt efni, þ.e. orðaforði frá miðri 16. öld til nútímans, en samt sem áður er þetta vísbending um hve gríðarlegur orðaforðinn er í raun og veru. Orðmyndun í íslensku er virk og því þarf að nota allar aðferðir til þess að fá heildarmynd af orðaforðanum og allar heimildir, orðasöfn, málheildir, og hvað eina sem til fellur, auk greiningartóla og reglna.

8 Niðurstaða

Eins og fram kom í inngangi var upprunalegur tilgangur BÍN sá að búa til efni til málækninota og jafnframt að koma efninu á framfæri til almennra notenda. Við vinnuna hefur gildi BÍN við rannsóknir á beygingarkerfinu smám saman komið betur og betur í ljós. Þrátt fyrir aldalinga hefð í lýsingu á íslenska beygingarkerfinu er enn mikið verk óunnið. Íslenska beygingakerfið er flókið, þar er mikill fjöldi beygingarflokka, afbrigði eru algeng og jafnframt er skrið á orðum og orðmyndum á milli beygingarflokka. Vegna þessa var strax frá upphafi ákveðið að í BÍN yrðu sett upp beygingardæmi fyrir einstök orð og þau geymd sem ein heild í stað þess að búa til kerfi af ofvirkum beygingarreglum. Hverju beygingardæmi fylgja tvenns konar athugasemdir. Í vinnuathugasemdum eru niðurstöður af athugunum á einstökum orðum og annað tilfallandi efni um vinnuna en í athugasemdum til notenda á vefnum eru leiðbeiningar um málnotkun, t.d. um notkun afbrigða, eftir merkingu, stíl o.s.frv.

Enn sem komið er byggist BÍN á því að rannsaka einstök orð og geyma beygingardæmin sem heild. Smám saman verður raunhæft í ljósi þess sem áður er rakið að búa til reglukerfi til þess að lýsa íslenska beygingarkerfinu, eftir því sem gögnin í BÍN aukast. Slíkt tól kæmi áreiðanlega að góðum notum í ýmsum málækniverkefnum, sérstaklega ef það væri notað ásamt orðhlutagreini, þar sem BÍN er ekki og verður aldrei tæmandi.

Rétt er að minna á að BÍN er enn verk í vinnslu. Gögn úr BÍN eru aðgengileg án endurgjalds á vefsíðu Stofnunar Árna Magnússonar í íslenskum fræðum, með þeim skilmálum sem þar koma fram.

Heimildir

- Anton Karl Ingason, Hrafn Loftsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). Í: Raante, Arne, og Bengt Nordström (ritstj.): *Advances in Natural Language Processing*, s. 205–216. (Lecture Notes in Computer Science, Vol. 5221.) Berlín: Springer.
- Ásta Svavarsdóttir. 1993. *Beygingakerfi nafnorða í nútímaíslensku*. Reykjavík: Málvísindastofnun Háskóla Íslands.
- Eiríkur Rögnvaldsson. 1990. *Íslensk orðhlutafræði*. Reykjavík: Málvísindastofnun Háskóla Íslands.
- Eiríkur Rögnvaldsson 1998. Málfræði í íslenskri orðabók: Hvernig og til hvers? *Orð og tunga* 4:25–32.
- Eiríkur Rögnvaldsson, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna B. Nikulásdóttir, Matthew Whelpton og Anton Karl Ingason. 2009. Icelandic Language Resources and Technology: Status and Prospects. *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Odense.
- Friðrik Magnússon. 1984. Ein lítil beygingarending. *Mímir* 32:33–43.
- Guðrún Kvaran. 2005. *Orð. Handbók um beygingar- og orðmyndunarfræði. Íslensk tunga 2*. Reykjavík: Almenna bókafélagið.
- Guðrún Kvaran og Sigurður Jónsson frá Arnarvatni. 1991. *Nöfn Íslendinga* (1. útg.). Reykjavík: Heimskringla.
- Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir og Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. Sarasola, Kepa, Francis M. Tyers og Mikel L. Forcada (ritstj.), *SalTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, s. 53–60. Valetta, Möltu. [Rafræn útgáfa 8.7.2013: <https://notendur.hi.is/eirikur/corpus tagging.pdf>.]
- Jón Friðrik Daðason. 2012. *The Post-Correction of Icelandic OCR Text*. Skemman.is. [<http://hdl.handle.net/1946/12085>]

- Jörgen Pind, Friðrik Magnússon og Stefán Briem. 1991. *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.
- Jörgen Pind, Kristín Bjarnadóttir, Jón Hilmar Jónsson, Guðrún Kvaran, Friðrik Magnússon og Ásta Svavarsdóttir. 1993. Using a Computer Corpus to Supplement a Citation Collection for a Historical Dictionary. *International Journal of Lexicography* 6/1:1–18.
- Kristín Bjarnadóttir. 1998. *Norræna verkefnið*. [8.7.2013: <http://lexis.hi.is/kristinb/norr.pdf>.]
- Kristín Bjarnadóttir. 2006. Málfræði í orðabókum. *Orð og tunga* 8:27–43.
- Kristín Bjarnadóttir. 2010. *Algilt -i eða hverfult: Um þágufall eintölu í sterkum hvorugkynsnafnorðum*. 24. Rask-ráðstefnan 30. janúar 2010. [<http://lexis.hi.is/kristinb/KB-Rask2010.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. [Rafræn útgáfa 8.7.2013: <http://www.lexis.hi.is/kristinb/lrec2012-dmii.pdf>.]
- Mörður Árnason (ritstj.). 2000. *Íslensk orðabók*. Rafræn útgáfa. Reykjavík: Edda hf.
- Rask, Rasmus Christian. 1811. *Vejledningen til det Islandske eller gamle nordiske Sprog*. København: Schuboths Forlag.
- Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir 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. [Rafræn útgáfa 8.7.2013: <http://www.lexis.hi.is/kristinb/lrec2012SH-saltmil.pdf>.]
- Valtýr Guðmundsson. 1922. *Íslensk grammatik*. Kaupmannahöfn: Hagerup.

Lykilorð

máltækni, gagnasöfn, beyging, málheild, gagnaskortur

Keywords

language technology, language resources, inflection, corpora, data scarcity

Abstract

The topic of the article is the *Database of Modern Icelandic Inflection* (DMII), a collection of about 270 thousand paradigms with over 5.8 million inflectional forms. The DMII was created as a multipurpose resource, for use in language technology, lexicography, and as a reference tool for the general public online. In spite of a long tradition of morphological research, it was clear from the beginning that the necessary data for a productive rule system for Icelandic inflection was not available, in a manner that would produce all and only the appropriate inflectional forms. The result is therefore

a database containing the full set of inflectional forms, with grammatical tags. The article contains a description of the creation of the DMII and its sources, initially almost completely lexicographic. The creation of the *Tagged Icelandic Corpus* (MÍM) has, for the first time, made an empirical study of Icelandic inflection possible, but comparison of the data from the DMII and MÍM shows that data scarcity makes it imperative to use both kinds of data for a thorough description of the inflectional system, due to the huge tag set in a very rich inflectional system.

Kristín Bjarnadóttir

Stofnun Árna Magnússonar í íslenskum fræðum / Háskóli Íslands

Neshaga 16

107 Reykjavík

kristinb@hi.is