Kristín Bjarnadóttir

Orðin í Markaðri íslenskri málheild og íslenskur orðaforði

Samanburður á MÍM og BÍN

Málvísindakaffi
20. janúar 2012

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Verkefnið: Að bæta orðaforðanum úr Markaðri íslenskri málheild (MÍM) við í Beygingarlýsingu íslensks nútímamáls (BÍN).

MÍM: 25 milljón lesmálsorð; hér er úrtakið 17,5 milljónir
- Textar frá 21. öld, valdir textar, blandað efni

BÍN: 270 þúsund beygingardæmi, úr ýmsum áttum
- Orðalistar: Orðabækur, söfn OH, örnefnalistar, fyrirtækjaskrá ...
- Textar: Íslensk orðtíðnibók, 1 árgangur af Morgunblaðinu
- Ábendingar frá notendum

Meginspurningar:
- Hve stór hluti af orðmyndunum í MÍM á heima í BÍN?
- Hvers konar efni úr MÍM á ekki heima í BÍN?
- Hvað skilar MÍM góðu yfirliti um íslenskan orðaforða?
- Hvað þarf stórt textasafn til að sýna dæmi um orðin í BÍN?

Mörkuð íslensk málheild (MÍM):

Alls 25 milljón lesmálsorð.
Tölur um orðaforðann (lemmurnar) eru ekki tiltækar enn.
Hér er skoðaður hluti MÍM, u.þ.b. 17,5 milljón tókar.

Beygingarlýsing íslensks nútímamáls (BÍN):

U.þ.b. 270 þúsund beygingardæmi, u.þ.b. 5,8 milljón beygingarmyndir með greiningarstreng, u.þ.b. 2,8 milljónir orðmynda.

Samanburðartölur:

MÍM:	16.245.429 lesmálsorð*
	737.856 orðmyndir
MÍM og BÍN:	425.238 orðmyndir
MÍM en ekki BÍN:	312.618 orðmyndir

*Greinarmerki o.þ.h. fjarlægð.

Hugtök

... þó að mig minni að það sé minna minni í minni vél ...

tókar: lesmálsorð og aðrir strengir = 14

lesmálsorð: orð í texta = 12

orðmynd: stafastrengur, óháður uppflettiorði = 9

beygingarmynd: lemmuð orðmynd með greiningarstreng = 11

lemma: beygingardæmi, uppflettiorð, fletta (orð) = 11

Gögnin í samkeyrsluna

MÍM (tíðni, orðmynd, mark, lemma)
1326|OM:þar|MARK:nken-s|LEMMA:þar
1302|OM:þó|MARK:nken-s|LEMMA:þó
1270|OM:ávallt|MARK:aa|LEMMA:ávallt
1264|OM:þarna|MARK:nven-s|LEMMA:þarna
1262|OM:vísindavefnum|MARK:nkeþgs|LEMMA:vísindavefur

BÍN (orðmynd) Svona listi var notaður í verkið
vísindaútgáfurnar
vísindaverk

BÍN (beygingarmynd)
vísindaútgáfa;469574;kvk;alm;vísindaútgáfurnar;ÞFFTgr;
vísindaverk;70842;hk;alm;vísindaverk;NFET;

Aðferðin I

1. Orðmyndalistar úr MÍM og BÍN keyrðir saman (**JFD**)
2. Orðmyndir úr MÍM sem ekki eru í BÍN keyrðar út, með tíðni, marki (greiningarstreng) og lemmu
3. Þetta eru 312.618 / 347.198 línur (2 útkeyrslur)
4. Efni sem ekki er í BÍN er yfirfarið, leiðrétt og flokkað

MÍM, orðmyndir sem ekki eru í BÍN (2):

1|OM:hópmeðferð|MARK:nveþ|LEMMA:hópmeðferð
1|OM:mafíós|MARK:nkee|LEMMA:mafíóur
1|OM:sjálfsvígsflugmenn|MARK:nkfn|LEMMA:sjálfsvígsflugmaður
1|OM:diplodocus|MARK:nken-s|LEMMA:diplodocus
1|OM:vaxtasprotana|MARK:nkfog|LEMMA:vaxtasprotur
1|OM:aðildarskilyrðunum|MARK:nhfþg|LEMMA:aðildarskilyrða
1|OM:margþjóða|MARK:nvfe|LEMMA:margþjóð
1|OM:galíleó-geimfarsins|MARK:nheegs|LEMMA:galíleó-geimfarsins
1|OM:laugarvatnsskóla|MARK:nkeþ-s|LEMMA:laugarvatnsskóli
1|OM:farsímarisans|MARK:nkeeg|LEMMA:farsímarisi

Orðmyndir sem ekki eru í BÍN	Ath:
166|OM:þessvegna|MARK:aa|LEMMA:þessvegna	*riháttur?*
166|OM:platon|MARK:nken-s|LEMMA:platon	*íslenska?*
165|OM:dl|MARK:nhfn|LEMMA:dl	*skst*
164|OM:l.|MARK:nken-s|LEMMA:l.	*skst*
163|OM:jack|MARK:nken-s|LEMMA:jack	*erlent*
163|OM:einvörðungu|MARK:aa|LEMMA:einvörðungu	*óbeygt*
162|OM:thad|MARK:nheo|LEMMA:thad	*riháttur?*
162|OM:who|MARK:nken-s|LEMMA:who	*erlent*
162|OM:moskvu|MARK:nveþ-s|LEMMA:moskva	
162|OM:hljópu|MARK:sfg3fþ|LEMMA:hljópa	*villa?*
162|OM:rúv|MARK:nkee-s|LEMMA:rúv	*skst*
161|OM:td|MARK:nkee-s|LEMMA:td	*skst*
159|OM:sl.|MARK:lheosf|LEMMA:síðastliðinn	*skst*
159|OM:sérðu|MARK:sfg2en|LEMMA:sjá	*spurn.*
158|OM:skjöldu|MARK:nveo|LEMMA:skjölda	*fornt/lemma*
158|OM:sh|MARK:nvee-s|LEMMA:sh	*skst*
158|OM:magnúss|MARK:nken-s|LEMMA:magnúss	*?*

Hvað finnst ekki í BÍN?

- Óbeygjanleg orð: *einvörðungu* (163) (upphrópanir, forsetningar, atviksorð ...)
- Beygingarmyndir sem vantar í BÍN (afbrigði eða villur)
- Orðmyndir fornu máli: *skjöldu (176)*
- Stafsetningarafbrigði sem ekki eru notuð í 21. aldar máli:
  sízt (197), *áherzlu* (100), *helzt* (96), *bezt* (89)
- Afbrigðilegt stafasett úr tölvum (merkt sem villur):
  thad 1.211 dæmi/16 mörk, hk/kk/kvk/lo
  thetta 524 dæmi/11 mörk, kvk/so/kk/lo
- Skammstafanir og styttingar
- Útlenska: the 6.005 dæmi/5 mörk, greint sem no-án-kyns
  aabenbaringer, Aachen (stakdæmi)
- Tölustafir, tákn o.þ.h.: *d4* (89), *e4* (93) ... *rf6* (77), *km2* (69)

... og svo beygingarmyndir af orðum sem vantar í BÍN,
þ.e. orðin sem leitað er að ...

Orð úr MÍM sem vantar í BÍN:

félagsþjónusta
frjálslyndisstefna
kveðjustund
sveitarstjórnarkosning
Kárahnjúkavirkjun
heimilisuppbót
erfðagreining
kvenfélagsdeild
vinstrihreyfing
málsgrein
handavinnustofa
leshömlun
efnahagsbrotadeild
dyslexia

Nokkur kvenkynsnafnorð

bloggsíða
stöðutaka
geimvera
bænastund
hljóðkerfisvitund
yfirheyrsla
aukasýning
líknardeild
kóræfing
greiningardeild
tónleikaferð
meðalgreiðsla
tæknigreining
frjálslyndisstefna

Þessi orð eru frekar algeng í MÍM (100)

Aðferðin II: Hvernig er unnið úr orðmyndalistanum?

1. Orðmyndunum í MÍM var skipt í flokka eftir fyrsta hluta af markinu:

Íslensk orð:	Beygjanleg:	hk kk kvk lo so fn gr (to)
	Óbeygjanleg:	ao/fs st uh to
Annað:	Erlend orð, ógreint, nafnorð án kyns	

2. Farið var yfir hvern flokk á skjá og merkingar fyrir þá leiðréttar.
Nýir flokkar voru búnir til eftir þörfum.
Lemmur voru leiðréttar um leið í beygjanlegum orðum:

*mannslunga **nk**eþ*
*mannsl**ungi kk** → mannsl**unga hk***

Aðrar lemmur voru ekki leiðréttar; þar dugar orðmyndin sjálf.
Mörkin sjálf voru ekki leiðrétt.

Aðferðin II: Hvernig finnast viðbætur í BÍN?

1|OM:aðildarskilyrðunum|MARK:nhfþg|LEMMA:aðildarskilyrð**a**
NÝLEMMA:aðildarskilyrð**i hk → hkft**

1|OM:vaxtasprotana|MARK:nkfog|LEMMA:vaxtasprot**ur**
NÝLEMMA:vaxtasprot**i** kk → kk **?**

1|OM:midi.is|MARK:nhee|LEMMA:midi.is
midi.is **hk → veffang**

162|OM:who|MARK:nken-s|LEMMA:who
who **kk → x**

Bætt er við nýjum flokkum, t.d. veffangi, villu o.þ.h., og athugasemdum til frekari flokkunar.

Helstu flokkunaratriði

Orðflokkur, kyn:	hk, kk, kvk, lo, ao, so, to, st, fs, uh ...
Vafi um kyn, orðflokk:	hk/kk ..., kk/kkft ..., so/lo ...
Útlenska:	x
Skammstafanir:	skst
Villur:	v
Vafamál:	?

... og veffang, tákn, tölustafir, orðhlutar, fornmál ...

Farið var yfir tæplega 350 þúsund línur á skjá og efninu síðan skipt í flokka eftir merkingunum á borð við þær hér að ofan.

Niðurstöðurnar? →

Niðurstöður, fyrsta tilraun:

	Lemmur í MÍM:	Rétt:	Leiðrétt:
HK	54.764	21.957	22.378
KK	108.417	28.972	21.800
KVK	58.383	31.270	19.297
LO	16.357	11.716	6.631
AO	1.629	774	1.214
SO	7.640	1.035	2.183

1. dálkur: Tölur um upprunalegan orðflokk í MÍM
2. dálkur: Tölur um réttar lemmur í MÍM
3. dálkur: Tölur um orðflokk eftir leiðréttingu

Niðurstaðan er að lemmunin mætti vera betri.

Sýnishorn af lemmun og leiðréttingu, sagnir:

670|OM:viltu|MARK:sfg2en|LEMMA:vilja|NÝLEMMA:vilja|ORG:so|ATH:so|**sp**
642|OM:skaltu|MARK:sfg2en|LEMMA:skulu|NÝLEMMA:skulu|ORG:so|ATH:so|**sp**
488|OM:e.|MARK:sfg3en|LEMMA:vera|NÝLEMMA:vera|ORG:so|ATH:**skst**
...
63|OM:tja|MARK:sng|LEMMA:tja|NÝLEMMA:tja|ORG:so|ATH:**uh**
62|OM:þanngað|MARK:sþghen|LEMMA:þannga|NÝLEMMA:þannga|ORG:so|ATH:**v**
60|OM:thetta|MARK:sfg3fn|LEMMA:thetta|NÝLEMMA:thetta|ORG:so|ATH:**v**
...
4|OM:baddi|MARK:sfg3eþ|LEMMA:badda|NÝLEMMA:**badd**i|ORG:so|ATH:**kk|gælunafn**
4|OM:ákveddi|MARK:svg3eþ|LEMMA:ákvedu|NÝLEMMA:**ákveða**|ORG:so|ATH:so
4|OM:ákvaðum|MARK:sfg1fþ|LEMMA:ákvaða|NÝLEMMA:ákveða|ORG:so|ATH:**v**
...
3|OM:langneðst|MARK:sþmhen|LEMMA|ORG:so|ATH:lo :langneða|
NÝLEMMA:**langneðstur**|ORG:so|ATH:**lo|est**
3|OM:langerma|MARK:sng|LEMMA:langerma|NÝLEMMA:langerma|ORG:so|ATH:**lo**
3|OM:kynnstist|MARK:sfm3eþ|LEMMA:kynnsja|#|NÝLEMMA:kynnsja|ORG:so|ATH:**v**

Sýnishorn af “aðskotaefni”:

28|OM:frett.html|MARK:nken|LEMMA:frett.html|...|ORG:kk|ATH:veffang
1|OM:e6kepp1.htm|MARK:nkee-s|LEMMA:e6kepp1.htm|...|ORG:kk|ATH:veffang
1|OM:raforka_.htm|MARK:nken|LEMMA:raforka_.htm|...|ORG:kk|ATH:veffang
...
1|OM:ld50|MARK:nkeþ-s|LEMMA:ld50|#|NÝLEMMA:ld50|ORG:kk|ATH:tákn
1|OM:g50|MARK:nken-s|LEMMA:g50|#|NÝLEMMA:g50|ORG:kk|ATH:tákn
1|OM:hs50|MARK:nken-s|LEMMA:hs50|#|NÝLEMMA:hs50|ORG:kk|ATH:tákn
...
4|OM:homogenic|MARK:nken-s|LEMMA:homogenic|...|ORG:kk|ATH:x
2|OM:arsenic|MARK:nken-s|LEMMA:arsenic|...|ORG:kk|ATH:x
1|OM:scénic|MARK:nken-s|LEMMA:scénic|...|ORG:kk|ATH:x
...
1|OM:framlengum|MARK:nheo|LEMMA:framlengum|...|ORG:hk|ATH:v
1|OM:ráðherrra|MARK:nken|LEMMA:ráðherrra|...|ORG:kk|ATH:v
4|OM:árid|MARK:nkeo|LEMMA:áridur|...|ORG:kk|ATH:v
...
1|OM:brd|MARK:nken-s|LEMMA:brd|#|NÝLEMMA:brd|ORG:kk|ATH:skst
1|OM:bsd|MARK:nken-s|LEMMA:bsd|#|NÝLEMMA:bsd|ORG:kk|ATH:skst
1|OM:aðalsteinsd|MARK:nken-s|LEMMA:aðalsteinsd|...|ORG:kk|ATH:skst

Áætlaðar tölur um MÍM

Orðmyndir í MÍM:	**737.856**		
Ekki í BÍN*:	312.618	/	**347.198**
“+Ísl.” alls:			**174.538**
“-Ísl.” alls:			172.650
Þar af til skoðunar:	**95.251**		
Annað, villur í skrám o.þ.h.:			34.580
Í BÍN:	425.238		

Hlutfall “aðskotaefnis” í MÍM, þ.e. efnis sem ekki telst vera íslenskur orðaforði er u.þ.b. 13%. (Gróf ágiskun.)

Áætlaðar viðbætur við BÍN: 125 þúsund beygingardæmi.

Allar tölur eru einungis ætlaðar sem vísbendingar.

Tölur um MÍM

Orðmyndir í MÍM:		**737.856**	
Ekki í BÍN*:		312.618 /	**347.198**
“+Ísl.”	Rétt lemmað:		112.966
	Röng lemma, réttur orðflokkur:		28.334
	(Réttur orðflokkur: 141.300)		
	Röng greining:		61.572
	“+Ísl.” alls:		**202.872**
“-Ísl.”	Útlenska:		69.681
	Skammstafanir:		5.387
	Villur:		20.183
	“-Ísl.” til skoðunar:		**95.251**

Í skránum er að auki dálítið af óskilgreindu drasli.
Í BÍN: 425.238

Íslenskar orðmyndir í MÍM: **yfir 600 þús.**

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Meginspurningum um MÍM og BÍN er ósvarað ...

- Hve stór hluti af orðmyndunum í MÍM á heima í BÍN?
  Þegar villur hafa verið hreinsaðar burt, 85%?

- Hvers konar efni úr MÍM á ekki heima í BÍN?
  Allt sem ekki á heima í íslenska beygingakerfinu, þ.e. það sem ekki telst til íslensks orðaforða ...

- Hvað skilar MÍM góðu yfirliti um íslenskan orðaforða?
  Þessu er ósvarað. Lemmurnar eru ótaldar enn en fjöldinn er ekkert á við flettiorðafjölda í Ritmálssafni ...

→

Meginspurningum um MÍM og BÍN er ósvarað ...

• Hvað þarf stórt textasafn til að sýna dæmi um allar beygingarmyndir í BÍN?

Það er ekki hægt; margar þeirra koma aldrei fyrir. Beygingarmyndir í BÍN eru 5,8 milljónir, orðmyndir í þessum hluta MÍM eru innan við 800 þúsund.

Stærðin á markamenginu veldur m.ö.o. gagnaskorti.

Afraksturinn af verkefninu eru viðbætur í BÍN og ýmsir listar sem nýta má í máltækni, t.d. villulistar, listar um skammstafanir o.s.frv.

Útúrdúr: Hvað er íslensk orðmynd? Dæmi um grísk nöfn.

- Orðmyndir með íslenkum beygingarendingum:
  Parmenídesi, Evripídesar; Dioscoridesar ...

- Orðmyndir þar sem stafsett er upp á íslensku (með afbrigðum):
  Evbúlídes ...

- Aðrar orðmyndir úr sömu/sambærilegum beygingardæmum:
  Parmenídes, Karmídes ...

- ... en *Orfeo* er merkt sem erlent orð ...

Athugið! Samhengið er ekki skoðað. Flokkunin er mjög gróf.
Betra er að taka meira en minna; þetta er fyrsta umferð.

Hvað á að vera í BÍN? Íslenskur orðaforði ...

- BÍN er ætluð til máltækninota og þarf að vera yfirgripsmikil.
- BÍN er ekki orðabók og þar er mikið af virkum samsetningum.
- Í BÍN er reynt að sýna afbrigði eftir því sem fært er.
- BÍN á bara að ná yfir nútímamál.

- BÍN er líka notuð af almenningi og því verður að gæta hófs í beygingarafbrigðum sem þykja “vond”.
- BÍN er ekki leiðbeinandi að því marki að þar séu bara “góð” orð; tökuorð, slettur og stafsetningarafbrigði finnast í BÍN.
- Reynt er að koma til móts við þá sem vilja fá að vita hvað er “gott og rétt” með ábendingum ofan við beygingardæmin.
- Notendum BÍN er ætlað að taka sjálfir afstöðu til þess hvað er við hæfi.

BÍN er í stöðugri þróun og reynt leita að dæmum um vafaatriði eftir því sem hægt er.

Allar ábendingar vel þegnar!
Takk fyrir áheyrnina.

Kristín B: kristinb [hjá] hi.is
BÍN: bin.arnastofnun.is
MÍM: mim.hi.is

Eftirþankar, tölur →

Viðbótartölur um orðmyndir í MÍM sem ekki eru í BÍN

Orðmynd, greiningarstrengur, lemma	347.199
Orðmynd, lemma	288.949
Lemma, orðflokkur, óleiðrétt	270.440
Ný lemma, orðflokkur	235.221
Orðmynd, ný lemma, orðflokkur	280.422

Þessar tölur eru úr fyrstu útkeyrslu JFD.

Orðin í Markaðri íslenskri málheild og íslenskur orðaforði
Kynningartexti

Sagt verður frá samanburði á orðaforðanum í Markaðri íslenskri málheild (MÍM) og Beygingarlýsingu íslensks nútímamáls (BÍN). Hugmyndin er að bæta orðaforðanum úr MÍM við BÍN. Þetta er fyrsta atrenna að samanburði á orðmyndum úr verulega stóru textasafni við hefðbundna lista um íslenskan orðaforða en orðaforðinn í BÍN er að verulegu leyti fenginn úr gagnasöfnum Orðabókar Háskólans og Íslenskri orðabók.

Í MÍM verða 25 milljón lesmálsorð úr 21. aldar máli en hér eru u.þ.b. 17,5 milljón lesmálsorð til skoðunar. Í þessum hluta MÍM eru alls u.þ.b. 738 þúsund orðmyndir, þar af ríflega 312 þúsund sem ekki eru í BÍN. Í BÍN eru u.þ.b. 270 þúsund beygingardæmi, tæplega 2,8 milljón orðmyndir (þ.e. stafastrengir án greiningar). Eftir fyrstu yfirferð yfir orðmyndir úr MÍM er áætlað að um 150 þúsund uppflettiorð (lemmur) bætist við BÍN.

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM