

# Stafastrengur á milli bila

Málleg gagnasöfn um orðmyndir  
og máltækniþól

Kristín Bjarnadóttir  
Orðfræðisvið SÁ

Hugvísindaping  
26. mars 2011

# Stafastrengur á milli bila?

Til kirkio ligr irakiaholte heima land með ollom landf nýtiom

Til kirkju liggur í Reykjaholti heimaland með öllum landsnytjum

**Hver strengur er mögulegt orð eða orðmynd sem máltækniþolin þurfa að ráða við, hvort sem það er í leiðréttingu, leit, greiningu eða textavinnslu af einhverju tagi.**

Dæmið hér að ofan er úr Reykjaholtsmáldaga

Vinna við önnur málstig en nútímamálið er í frumstigi.  
Gagnasöfn um nútímamál eru lengra komin.

## Efnið:

### Gagnasöfn um orðmyndir:

**Beygingarlýsing íslensks nútímamáls (BÍN):**  
**Eins og nafnið bendir til nær BÍN aðeins til nútímamáls**

**Ritmálssafn Orðabókar Háskólans (Rms):**  
**Dæmasafn, 16. – 20. öld**

**Tilraunaverkefni:**  
**Samtengdar orðmyndir frá ýmsum tímum (1150-1828)**

**Hvernig er hægt er að nýta þessi söfn í máltækni?**

**Dæmi: Leiðrétting á skönnuðum texta**  
**Samræming á stafsetningu úr eldra máli**

# Beygingarlýsing íslensks nútímamáls (23.3.2011)

Beygingarmyndir, með greiningarstrengjum:	5.878.733
Orðmyndir:	2.807.836
Beygingardæmi:	271.321

fara;433568;so;alm;**fara**;GM-NH;  
fara;433568;so;alm;**fer**;GM-FH-NT-1P-ET;  
fara;433568;so;alm;**ferð**;GM-FH-NT-2P-ET;  
fara;433568;so;alm;**fer**;GM-FH-NT-3P-ET;  
fara;433568;so;alm;**förum**;GM-FH-NT-1P-FT;  
fara;433568;so;alm;**farið**;GM-FH-NT-2P-FT;  
fara;433568;so;alm;**fara**;GM-FH-NT-3P-FT;  
fara;433568;so;alm;**fór**;GM-FH-NT-1P-ET;

Öllum opið á [bin.arnastofnun.is/gogn/](http://bin.arnastofnun.is/gogn/)  
Þar eru líka skýringar á gögnunum.

# Ritmálssafn Orðabókar Háskólans: Sögulegt efni

## 708 þúsund uppflettiorð

### Ritmyndir uppflettiorða, dæmi:

gekkst	59	geckst	2	18m-19f
		gekkst	41	17-20ms
		gekst	7	18f-20f
		géckst	3	18s-19f
		gékkst	2	19m
		gékst	2	19m
		gieckst	1	17m
		giekst	1	17m

Ritmyndir af sögninni **ganga** eru 207.

Samsvarandi strengir í nútímamáli eru 37.

Sjá: [arnastofnun.is/page/arnastofnun\\_gagnasafn\\_ritmal](http://arnastofnun.is/page/arnastofnun_gagnasafn_ritmal)

# Ritmyndir uppflettiorða úr Rms og tilsvarandi beygingarmyndir í BÍN

## **BÍN:** Ritmálssafn:

**gangi** gaange gange gänge genge geinge gangi Gangi  
gángi Gángi gänngi gángji

**gangið** ganged gangid gángid gengid Gengid géngid geingid gangið

**gangir** gangir gángir

**gangist** gangest gangist gángist gángjist

**gekk** gegg gech geck gieck Gieck géck geick géckk gekk Gekk gjekk  
gékk

**gekkst** geckst gieckst géckst gekst Gekst giekst gékst gekkst  
Gekkst gékkst Gékkst

**gengi** gengi Gengi géngi geingi geíngi gíngi

**gengið** genged geinged Geinged gejnged gengið Gengið geingið  
geíngið gjeíngið

**genginn** gengengen geingen géingen gengin Gengin geingin geingenn  
gejngenn genginn geinginn Geinginn

...

KB 26.3.2011

# Ritmálssafn Orðabókar Háskólans

Tilraunaverkefni, hluti af úttaki úr gagnagrunninum.  
Dæmasafnið notað sem orðmyndasafn fyrir máltækniól.

Tölvutæk dæmi:	1.721.527
Uppflettiorð (strengir):*	288.262
Ritmyndir uppflettiorða:	818.564
Orðmyndir í dæmum:	1.629.261
16. öld:	70.718
17. öld:	187.862
18. öld:	225.583
19. öld:	804.451
20. öld	1.117.402

ORD:fara|ORDMYND:farið|DÆMI:því hann hefir áður haldið líku fram við stjórnina og enda **farið** líku á flot fyrir annan hrepp.|D-ALDUR:19m

Úr dæmunum er unninn orðmyndalisti með aldursmerkingum og tíðnitölum.

# Önnur gögn um orðmyndir

## Orðasöfn og rafrænar orðabækur:

Talmálssafn

Orðalisti úr orðabók Jóns Ólafssonar úr Grunnavík

Orðabankinn (nú tímamál)

Orðfræðirit fyrri alda

## Textasöfn / hrágögn:

Mörkuð íslensk málheild

Íslenskt textasafn (SÁ)

## Textapör/parallel-textar frá ýmsum málstigum ...

Aðferðir til að búa nýtilega til orða- og orðmyndalista úr textum eru ekki nægilega þróaðar enn.

## Verkefni I: Leiðrétting á ljóslesnum textum

- Hluti af átaki Vinnumálastofnunar: 856 störf, sumarið 2010
- Felst í því að þróa aðferðir og hugbúnað til leiðréttingar á skönnuðum íslenskum textum frá tímabilinu 1875-1925
- Textarnir eru frá Landsbókasafni-Háskólabókasafni, af timarit.is
- Þátttakendur: Jón Friðrik Daðason, Kristján Rúnarsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Ásta Svavarsdóttir

Skannaðir textar eru handleiðréttilir.

Hugbúnaðurinn lærir af leiðréttingunum.

Verkefnið felst m.a. í samanburði við orðmyndalista úr BÍN og Rms.

Verkefnið er enn á tilraunastigi.

Orðalistar eru unnir upp úr leiðréttu textunum eftir því sem þeir verða til.

Sýnishorn →

## Verkefnið Leiðrétting á ljóslesnum textum

Skönnun: **lv**vei'kfjöll en ekki víð Kistufell og segist **liafa** farið **par**

Tólið: **lv**vei'kfjöll en ekki víð Kistufell og segist **hafa** farið **þar**

Handleiðr.: **K**verkfjöll en ekki víð<prentv **við** /> Kistufell og segist **hafa** farið **þar**

Skönnun : **pa** í **lýsmg**ingunni **hven**, héldu menn annarsstaðar að **pað**

Tólið: **þa** í **lýsing**unni **hve**, héldu menn annarsstaðar að **það**

Handleiðr.: **þá** í lýsingingunni<prentv **lýsing**unni /> **hver**i, héldu menn annarsstaðar að **það**

## Villa/leiðrétting/röng leiðrétting

- Textinn er borinn saman við orðalista úr BÍN, Rms og verkefninu sjálfu.
- Giskað er á leiðréttingar á orðmyndum sem ekki finnast í listunum.
- Tólið velur ekki alltaf réttar myndir úr orðasafninu ...
- Við vinnuna verða til orðmyndalistar sem nota má í önnur verkefni.

## Verkefni II:

# Tilraun til vörpunar orðmynda á milli málstiga

- Hugmyndin er að taka stafrétta texta frá ýmsum tímum og breyta þeim til nútímamáls með því að nota lista af orðmyndum (vörpunartöflu).
- Alltaf er unnið með heil orð, **stafastreng milli bila**, eða enn stærri einingar.

- Textarnir:

Riddarasögur: Vilhjálms saga sjóðs, Ektors saga,  
Valdimars saga, Viktors saga, Jarlmanns saga,  
Fimm bræðra saga, Marrons saga

Hallgrímur Pétursson: 4. bindi af útgáfu SÁ

Nikulás Klím, þýðing Jóns Ólafssonar úr Grunnavík

Paradísarmissir, þýðing Jóns Þorlákssonar á Bægisá

Reykjaholtsmáldagi: í vinnslu ...

## Reykjaholtsmáldagi (1150-1208)

Til kirkio ligr irækiaholte heima land með ollom landf nýtiom  
Til kirkju liggur í Reykjaholti heimaland með öllum landsnytjum

## Ectors saga (15. öld)

Efter nidrbrot Troioborgar þaa er Grickir hofdu hana wnith helldr af  
Eftir niðurbrot Trójuborgar þá er Grikkir höfðu hana unnið heldur af

## Hallgrímur Pétursson (Lbs 399 4to II)

Ad morgne þegar eg ärla vppstä/eins ad kvóllde eg hvijlast a  
Að morgni þegar eg árla uppstá/eins að kvöldi eg hvílast á

## Nikulás Klím (1745)

Hjer staulast framm til Yðar húsa karlinn Klíme, styrðr og stumrandi  
Hér staulast fram til yðar húsa karlinn Klími, stirður og stumrandi

## Paradísarmissir (1828)

Aungvum þeim, er nokkut skyn berr á bókmentir Íslendínga, mun  
Engum þeim, er nokkuð skyn ber á bókmenntir Íslendinga, mun

## Vörpunartaflan, sýnishorn

<code>\&lt;geck\&gt;</code>	gekk	Valdimars saga/Vilhjálmss saga sjóðs /HallgP-4/Klím
<code>\&lt;geckst\&gt;</code>	gekkst	Fimm bræðra saga/Marrons saga/Klím
<code>\&lt;geinginn\&gt;</code>	genginn	Fimm bræðra saga/Marrons saga/Klím
<code>\&lt;geingr\&gt;</code>	gengur	Fimm bræðra saga/Marrons saga/Klím
<code>\&lt;geingu\&gt;</code>	gengu	Jarlmanns saga/Klím
<code>\&lt;Geingu \&gt;</code>	Gengu	Fimm bræðra saga/Marrons saga
<code>\&lt;geingum\&gt;</code>	gengum	Klím
<code>\&lt;geingur\&gt;</code>	gengur	Fimm bræðra saga/Marrons saga /HallgP-4/Klím
<code>\&lt;gengid\&gt;</code>	gengið	HallgP-4
<code>\&lt;gengit\&gt;</code>	gengið	Vilhjálmss saga sjóðs/Milton
<code>\&lt;gengr\&gt;</code>	gengur	Valdimars saga/Vilhjálmss saga sjóðs /Milton

8.3.2011: 29.681 færslur í vörpunartöflunni.  
Reykjaholtsmáldagi er þar ekki meðtalinn.

## Fjöldi lesmálsorða í textunum

Reykjaholtsmáldagi:	u.þ.b.	450
Riddarasögur:		109.193
Ektors saga:	22.438	
Fimm bræðra saga:	16.991	
Jarlmanns saga:	12.411	
Marrons saga:	16.727	
Valdimars saga:	5.012	
Viktors saga:	10.461	
Vilhjálmss saga sjóðs:	25.153	
Hallgrímur Pétursson, 4.:		9.270
Nikulás Klím:		85.951
Paradísarmissir Miltons:		107.818

8.3.2011: 29.681 færslur í vörpunartöflunni.  
Reykjaholtsmáldagi er þar ekki meðtalinn.

## Tíðni breytinga í vörpunartöflunni um Klím

1.864	e\$>i	161	at\$>að	79	mm>m
861	lig>leg	148	ann\$>an	74	gjæ>gæ
679	er\$>ir	139	úng>ung	68	e=>i
565	r>ur	135	pt>ft	66	i>í
457	íng>ing	125	o>ó	64	vu>u
361	e>i	125	áng>ang	64	-ig>-ug
334	lld>ld	121	eing>eng	63	ei>ey
296	je>é	108	z>s	62	y>i
282	it\$>ið	108	q>k	57	ia>ía
267	ck>kk	83	isk>ísk	52	aung>öng
207	c>k	80	u>ú	51	e>é

\$: ending

=: morfemaskil

Klím er rúmlega 85 þúsund lesmálsorð.

# Notagildið

Leiðréttingarbúnaður:

Skönnunarleiðréttingarnar: Samhengisóháð leiðrétting

Villuleit og leiðrétting: Samhengisháð leiðrétting

Málgreining:

Auðveldara er að beita tólum sem gerð eru fyrir nútímamál á eldri texta ef aukalagi er bætt við með nútímastafsetningu

Leitarvélar:

Orðmyndatöflurnar geta verið inntak í leitarvélar og gegna þá sama hlutverki og BÍN þegar leitað er að öllum ritmyndum orðs í einu

Breyting á texta til nútímamáls:

Lesendavænar útgáfur ...

**Notagildið vex eftir því sem söfnin verða stærri ...**

## Vilhjálmss saga sjóðs

gengu þeir nú í turn Astrinómiu kongs dóttur og tóku hana með valdi. Voru borgarmenn nú í mikilli sorg. En þó þyrmdu þeir jungfrúinni fyrir hennar bæn. Lofaði hún því þá morgun að henni var eigi mjög hugféllt. Hugsaði hún að **frest væri á illu best**. Reginbaldr var settur í dyflissu og var honum eigi margra hæginda leitað. Síðan leitu þeir kanna valinn og búa um lík höfðingja enn skiptu herfangi með sér. Síðan fóru þeir yfir landið og lögðu það undir sig og þorði engi móti að mæla.

Gengu þeir nú í turn Astrinómíu kóngrs dóttur og tóku hana með valdi. Voru borgarmenn nú í mikilli sorg en þó þyrmdu þeir jungfrúinni fyrir hennar bæn. Lofaði hún því þá mörgu að henni var eigi mjög hugféllt. Hugsaði hún að frest væri á illu best. Reginbaldr var settur í dyflissu og var honum eigi margra hæginda leitað. Síðan leitu þeir kanna valinn og búa um lík höfðingja enn skiptu herfangi með sér. Síðan fóru þeir yfir landið og lögðu það undir sig og þorði engi móti að mæla.

**Takk fyrir áheyrnina!**

**Kristín Bjarnadóttir**

**[kristinb@hi.is](mailto:kristinb@hi.is)**

KB 26.3.2011

STOFNUN ÁRNA MAGNÚSSONAR  
Í ÍSLENSKUM FRÆÐUM