

Merkingarvensl í *Íslenskri orðabók*

Hádegisfyrirlestur Tungutækni­seturs

6.mars 2007

Anna Björk Nikulásdóttir

Um verkefnið

- Sjálfvirk greining merkingarvensla milli nafnorðaflettna *Íslenskrar orðabókar* og orða í skýringartextum þeirra
- Forritið MerkOr þróað í þessum tilgangi
- Verkefnið nýtir málfræðimarkara fyrir íslensku
- Merkingarvenslin eru greind út frá orðflokkamynstrum skýringartextanna

Merkingarvensl

- Merkingarvensl greind úr ÍO:
 - samheiti
 - yfirheiti
 - undirheiti
 - heildheiti
 - hlutheiti
 - tengt lýsingarorð
 - tengt sagnorð
 - eiginleiki
 - ætt

Yfirheiti - undirheiti

Y er yfirheiti X ef:

$$A \text{ er } X \Rightarrow A \text{ er } Y$$

Þetta er *hundur* \Rightarrow Þetta er *dýr*

$$\text{yfirheiti}(X, Y) \Leftrightarrow \text{undirheiti}(Y, X)$$

Samheiti

- Tvö orð eru samheiti ef þau merkja það sama.
- Algjör samheiti eru sjaldgæf.

rúmfatnaður - sængurfatnaður

yfirvaraskegg - hormotta

matarsóði - natríngarbónat

Heildheiti - hlutheiti

X er hlutheiti Y ef:

A er með X og X er hluti A

Bíll er með vél.

Vél er hluti bíls.

$\text{hlutheiti}(Y, X) \Leftrightarrow \text{heildheiti}(X, Y)$

Tengd orð af öðrum orðflokki

- Nafnorð geta verið leidd af sögnum eða lýsingarorðum:

íhugun - íhuga

ofvirkni - ofvirkur

- Einnig getur verið merkingarlegt samhengi milli orða sem ekki eru morfólógískt skyld:

eftirför - elta

Eiginleiki (*attribute*)

- Lýsingarorð sem á við viðkomandi nafnorð:

birkikvistur - lágvaxinn

vinnuþjarkur - vinnusamur

Líffræðileg ætt

- *Líffræðileg ætt* á við um nafnorð úr plöntu- og dýraríkinu:

birkikvistur - rósaætt

hrossafluga - tvívængja

Dæmigerðir skýringartextar

gleraugnaostur súrsaðir hrútsþungar

vaskur virðisaukaskattur

krakki barn, barnungi, krógi

basil einær kryddjurt af varablómaætt, blöðin eru
notuð m.a. í sósur og pastarétti

forritun það að forrita

hræðsla það að vera hræddur, ótti

hryllingur e-ð andstyggilegt

Málfræðileg mörkun

- Vélræn greining orða í texta eftir orðflokkum og beygingu
- Skýringartextar nafnorðaflettna ÍO voru markaðir
- Notaður var TnT-tölfræðimarkari Brants, sem þjálfður hafði verið á íslensku

Niðurstöður markara - dæmi

algórytmi	nhen
röð	nven
af	aþ
reglum	nvfþ
til	ae
að	cn
leysa	sng
ákveðið	lheosf
verkefni	nheo

Orðflokkamynstur

1) nn_a_np_a_c_s_l_no

2) það að_s

Orðflokkamynstur í ÍO

Skýringarhlutar (mynstur) alls: 106.977

<i>Mynstur byrjar á ...</i>	<i>Fjöldi</i>	<i>Hlutfall</i>
nafnorði	73.290	68,51%
lýsingarorði	14.684	13,73%
fornafni	7.362	6,89%
atviksorði/ forsetningu	4.585	4,29%
sagnorði	3.447	3,22%

Reglur - dæmi

`l_n` → `yfirheiti(fletta, n)`
`eiginleiki(fletta, l)`

`n_ og _n` → `hlutheiti(fletta, n)`

`það að _s` → `tengtSo(fletta, s)`

Niðurstöður

<i>Merkingarvensl</i>	<i>Fjöldi</i>
Samheiti / Jafnheiti	51.390
Yfirheiti	43.066
Eiginleiki	12.771
Líffræðileg ætt	2.817
Vísun	2.140
Tengd sögn	1.286
Samheiti	1.201
Hlutheiti	731
Tengt lýsingarorð	662
Heildheiti	382
ALLS	116.446

Niðurstöður

	Alls	Hlutfall greint
Skýringar	77.348	96,45%
Skýringarhlutar	106.972	92,61%

Prófunarsett:

rétt	82,13%
ófullnægjandi	12,64%
röng greining	5,23%

Ástæður rangrar greiningar

- Villur í niðurstöðum markara:
ljós rafsegulkynjuð (nn) ölduhreyfing (nn) sem ...
*yfirheiti(ljós, rafsegulkynjuð)
- Niðurstaða hefur enga merkingu:
bulla svipuð stöng í vél eða dælu
*eiginleiki(bulla, svipuð)
- Reglualgórýpmi

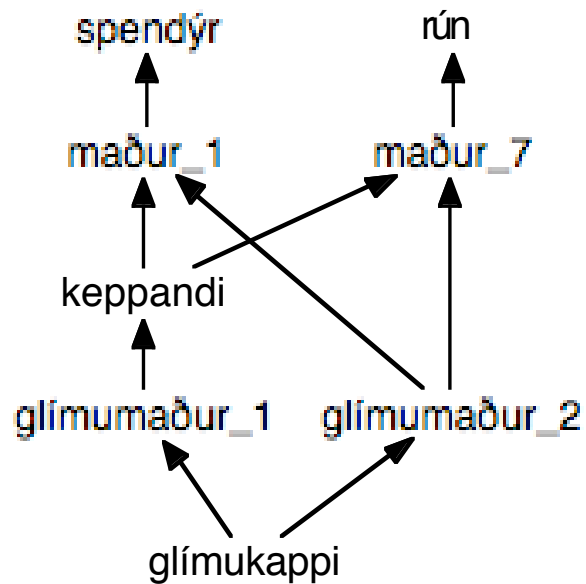
Næstu skref

- Yfirfara niðurstöður með *Beygingarlýsingu íslensks nútímamáls*
- Áframhaldandi þróun reglualgórýpma
- Greina merkingarvensl sagnorða og lýsingarorða
- Athuga með samsvarandi greiningu á öðrum orðabókum

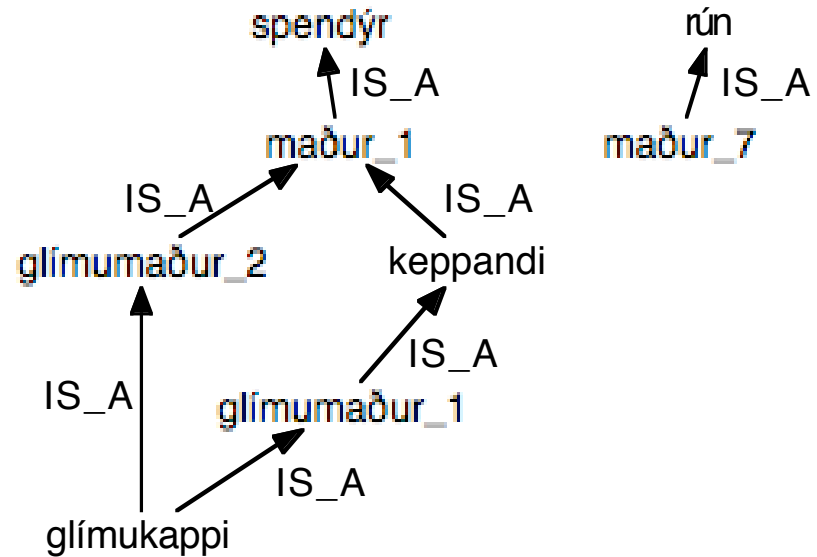
Notkunarmöguleikar

- Notendaviðmót *Íslenskrar orðabókar* á netinu
 - beinar vísanir í merkingarlega tengd orð
 - ýmis konar tengingar, t.d. yfirheitastigveldi og söfn undirheita
- Samræming skýringartexta

Yfirheitastigveldi



a) vélræn greining



b) handgerð greining

Safn undirheita

undirheiti (sund, (baksund (1), björgunarsund (1),
busl (1), bæjarsund (2), húsasund (2), jómfrúrsund
(1), kafsund (1), leið (3), lykkjustund (1),
sprettsund (1), stakkasund (1), áll (3)))