Orðabók Háskólans
Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Mörkun texta

Sigrún Helgadóttir

3. málþing
tímaritsins Orð og tunga

17. febrúar 2006
Orðabók Háskólans,
Reykjavík

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Þarfir tungutækniverkefna

- Tungutækniverkefni þurfa miklar og nákvæmar upplýsingar um tungumálið og notkun þess, t.d. um
  - tíðni orðflokka, orða og beygingar-mynda, orðasambönd, setningargerð og merkingu

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Þessar upplýsingar má fá úr markaðri málheild (e. *tagged corpus*)

- Safn tölvutækra texta af ýmsu tagi svo sem blaðatexta, fræðitexta af ýmsum sviðum, bókmenntatexta og talmáls.
- Hverju orði fylgir mark (e. *tag*) og nefnimynd (*lemma*)
- Hverjum texta fylgja upplýsingar um textann, höfund hans o.fl. (bókfræðilegar upplýsingar)
- Mörkuð málheild er geymd í stöðluðu sniði (XML)

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Unnið er að því að koma upp markaðri íslenskri málheild (mím)

Stefnt skal að:

- 25.000.000 orð
- 900–1.000 textabútar
- allt 40.000 orð hver, 10% sleppt ef texti styttri en 40.000 orð
- 90% nákvæmni í mörkun
- um 1.000.000 orð þar sem mörkun er leiðrétt handvirkt
- stofn málheildar er textasafn Íslenskrar orðtíðnibókar

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Hvað er mörkun?

Með **mörkun** (e. *tagging*) er átt við það að merkja orð í samfelldum texta á kerfisbundinn hátt, t.d. með málfræðilegum upplýsingum, nefnimynd (*lemma*) orðsins, upplýsingum um setningafræðilegt hlutverk og öðru sem nauðsynlegt er svo sem XML-mörkum.

Stundum kallað *markup* (eða *mark-up*), á íslenksu **ívaf**

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Oftast er orðið *mark* notað um málfræðilegt mark (e. PoS tag eða part-of-speech tag)

Málfræðilegt mark er greiningarstrengur sem er tengdur orði í texta og segir til um orðflokk orðsins og önnur málfræðileg atriði, t.d. kyn, tölu og fall fallorða og persónu, tölu og tíð sagna. Jafnframt getur mark sagt til um setningafræðilegt hlutverk orðs.

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Málfræðileg mörkun íslensks texta

- 1985–1991 var búið til textasafn fyrir gerð Íslenskrar orðtíðnibókar. Mörkuð í þremur atrennum. Búið var til forrit (notar beygingarfræðilegar upplýsingar, málfræðilegar reglur og tölfræðilegar aðferðir) sem eftir endurbætur skilaði um 90% greiningarstrengja réttum (Stefán Briem).
- 2002–2004 gerð tilraun með þjálfun gagnamarkara, byggðist á textum Orðtíðnibókar (SH)
- 2004–2005 Hrafn Loftsson býr til málfræðilegan reglumarkara (IceTagger), notar texta Orðtíðnibókar til prófana

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Textasafn Íslenskrar orðtíðnibókar (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991)

100 textar, hver um 5.000 lesmálsorð, gefnir út 1980–1989

5 flokkar: Íslensk skáldverk, ævisögur, þýdd skáldverk, fræðslutextar, barna- og unglingabækur

590.297 lesmálsorð

59.358 orðmyndir

639 greiningarstrengir

að meðtöldum greinarmerkjum

Öll greining var yfirfarin.

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Margræðar orðmyndir í orðtíðnibók

– 15,9% hafa fleiri en eina greiningu
– Margræðasta orðmyndin er: *minni* sem hefur 24 greiningarstrengi í textasafninu, en fleiri eru mögulegir (ég *minni* þig á það; ég geri þetta eftir *minni*; Nonni er *minni* en Siggi; o.s.frv.)

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

orð	**mark**	**skýring**
ég	fp1en	f: fn; p; pfn; 1: 2. pers.; e: et; n: nefnifall
stökk	sfg1eþ	s: so; f: frsh.; g: germ; 1: 1. pers.; e: et; þ: þátíð
á	aa	a: ao; a: stýrir ekki falli
eftir	aþ	a: ao; þ: stýrir þágufalli
stræto	nkeþ	n: no; k: kk; e: et; þ; þgf.
og	c	c: samtenging
veifaði	sfg1eþ	s: so; f: frsh.; g: germ; 1: 1. pers.; e: et; þ: þátíð
,	,	komma
vagnstjórinn	nkeng	n: no; k: kk; e: et; n: nf; g: með greini
sá	sfg3eþ	s: so; f: frsh.; g: germ; 3: 3. pers.; e: et; þ: þátíð
mig	fp1eo	f: fn; p; pfn; 1: 2. pers.; e: et; o: þf
og	c	c: samtenging
stoppaði	sfg3eþ	s: so; f: frsh.; g: germ; 3: 3. pers.; e: et; þ: þátíð
.	.	punktur

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Helstu aðferðir við mörkun
  - Handvirk mörkun
  - Vélrænar aðferðir
    - Gagnaaðferðir (data-driven methods) eða mörkun með námfúsum mörkurum, forrit lærir af fyrir fram greindu textasafni
    - Regluaðferð
      - hverju orði er úthlutað öllum hugsanlegum greiningarstrengjum á grundvelli orðasafns
      - handgerðar málfræðireglur ákveða hvaða greining er rétt (margræðni er eytt)
    - Sambland af fleiri en einni aðferð

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Fjórir gagnamarkarar skoðaðir
  - TnT (Falið Markovslíkan)
  - fnTBL (Transformation –Based Learning)
  - MXPOST (Hámarksóreiðuaðferð)
  - MBT (minnisaðferð, hefur nýlega verið prófuð)
- Sýna nokkrar niðurstöður úr tilrauninni
- Skoða villur sem markararnir gera

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Hvernig eru gagnamarkarar notaðir?

- Textasafni er skipt í þjálfunarsafn (90%) og prófunarsafn (10%)
- Gagnamarkari lærir af þjálfunarsafni, býr til líkan
- Líkan prófað á prófunarsafni
- Niðurstaða borin saman við rétt mörk og nákvæmni fundin
- Endurtekið fyrir 10 pör þjálfunar- og prófunarsafna

Útkoma er skrár (líkan) sem nota má til þess að marka nýjan texta

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

orð	mark	tnt	mxp	fnt
ég	fp1en	fp1en	fp1en	fp1en
stökk	sfg1eþ	sfg1eþ	sfg1eþ	sfg1eþ
á	aa	aa	aa	aa
eftir	aþ	ao	aþ	aa
stræto	nkeþ	nkeo	nkeþ	nkeo
og	c	c	c	c
veifaði	sfg1eþ	sfg3eþ	sfg3eþ	sfg3eþ
,	,	,	,	,
vagnstjórinn	nkeng	nkeng	nkeng	nkeng
sá	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
mig	fp1eo	fp1eo	fp1eo	fp1eo
og	c	c	c	c
stoppaði	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
.	.	.	.	.

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Niðurstaða af þjálfun og mörkun 10 para skráa

	Meðalnákvæmni		
	Öll orð	Þekkt orð	Óþekkt orð
Markari	%	%	%
fnTBL	88,80	91,36	54,02
MXPOST	89,08	91,04	62,51
TnT	90,36	91,74	71,62
MBT	87,00	89,21	56,86

Meðalhlutfall óþekktra orða: 6,84%

Mark er talið rangt þó að aðeins eitt af allt að 6 atriðum sé rangt

TnT gerir 14% færri villur en fnTBL

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Markaskrá Orðtíðnibókar er mjög ítarleg
- Sú greining sem þar er notuð er ekki endilega sú eina rétta
- E.t.v. geta sumar tungutæknilausnir nýtt sér greiningu sem er ekki jafn ítarleg
- Sum tungutækniverkefni gætu þurft mikla nákvæmni í mörkun en ekki mjög ítarlega greiningu

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Niðurstöður mörkunar með TnT-markaranum
Nákvæmni mörkunar þegar markaskrá er einfölduð
Safni skipt í 6 flokka

	Fjöldi	%	Rétt greind	Nákvæmni %	Rangt greind	Fækkun villna	Fækkun villna (%)
Allur greiningarstrengur réttur	533.403	90,36	533.403	90,36	56.894	–	–
Atviksorð ekki greind	6.837	1,16	540.240	91,52	50.057	6.837	12,02
Samtengingar ekki greindar	1.076	0,18	541.316	91,70	48.981	1.076	2,15
Öllum fornöfnum slegið saman	782	0,13	542.098	91,83	48.199	782	1,60
Aðeins orðflokkur réttur	37.197	6,30	579.295	98,14	11.002	37.197	77,17
Rangur orðflokkur	11.002	1,86	–	–	–	–	–
Samtals	590.297	100,00					

Með því að einfalda greiningu atviksorða fækkar villum sem TnT gerir um 6.837 eða 12,02%

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Áhrif mismunandi texta
  - Fyrri niðurstöður eiga eingöngu við textasafn Orðtíðnibókar
  - Bókmenntaleg slagsíða á textasafni Orðtíðnibókar
  - Hvað gerist ef tæknilegir textar eru fjarlægðir?

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Nákvæmni mörkunar TnT fyrir mismunandi texta

	Óþekkt orð	Meðalnákvæmni		
	%	Öll orð	Þekkt orð	Óþekkt orð
Allur textinn	6,84	90,36	91,74	71,62
Efnafr. texti fjarlægður	6,84	90,39	91,77	71,65
Raunv. texti fjarlægður	6,81	90,51	91,91	71,29

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Hvernig má bæta niðurstöður mörkunar?

- Einfalda markaskrá (hefur þegar verið sýnt)
- Bæta mörkun óþekktra orða
  - Bæta aðferðir markaranna
  - Láta í té viðbótarorðasafn og ýmsa lista
- Kjósa á milli markara
- Þjálfa nýjan markara á grundvelli niðurstaðna úr tveimur eða fleiri mörkurum
- Beita málfræðireglum

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Nákvæmni TnT og fnTBL með orðasafni

	Óþekkt orð án orðas. %	Nákvæmni		
		Öll orð	Þekkt orð	Óþekkt orð
TnT (án orðasafns)	6,84	90,36	91,74	71,62
Orðasafn*		**91,54**	91,93	86,31
Hlutf.l. fækkun villna (%)		12,25	2,27	51,75
fnTBL (án orðasafns)	6,84	88,80	91,36	54,02
Orðasafn*		90,06	91,50	70,44
Hlutf.l. fækkun villna (%)		11,20	1,63	35,70

* Orðasafnið er búið til úr helmingi þeirra orða sem álitin eru óþekkt frá sjónarhóli hvers prófunarsafns

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Kjósa um mörk
  - Besta niðurstöðu gaf aðferð þar sem vegið er með heildarnákvæmni hvers markara
  - Þrír markarar: valið það mark sem tveir eða fleiri eru sammála um, ef allir markarar eru ósammála er valið mark þess markara sem hefur hæsta heildarnákvæmni (í þessu tilviki TnT)

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Nákvæmni þriggja markara og nákvæmni sem fæst með kosningu

Aðferð	Öll orð	Þekkt orð	Óþekkt orð
MXPOST	89,08	91,04	62,50
fnTBL	88,80	91,36	54,03
TnT	90,36	91,74	71,60
Kosn. með heildarnákv.	**91,54**	92,99	71,80
Hlutfallsleg fækkun villna frá TnT	12,21	15,11	0,73

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Beita málfræðireglum til þess að velja niðurstöðu eins markara fram yfir niðurstöðu annars
- Skilyrði sem þurfa að vera til staðar
  - Markararnir gera ekki sömu vitleysurnar, þ.e. þeir bæta hver annan upp (*complementarity*)
  - Mismunur er kerfisbundinn en ekki tilviljunarkenndur

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Eftir skoðun var álitið vænlegast að nota útkomu MXPOST fyrir tiltekna samsetningu ef MXPOST gefur rétta greiningu fram yfir kosningu oftar en 5 sinnum.

Reglurnar verða á forminu:

ef útkoma úr kosningu er mark1 og útkoma MXPOST er mark2 þá skal velja mark2

MXPOST greinir réttar fall orða þar sem orðmyndir falla saman

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Nákvæmni þriggja markara
Orðasafn notað
Mörk einfölduð
Málfræðireglum beitt

Mörk	Orða-safn[1]	Aðferð	Óþekkt orð		Þekkt orð		Öll orð	
			Tíðni	%	Tíðni	%	Tíðni	%
Alls			40.392	100,00	549.905	100,00	590.297	100,00
Óbreytt	Nei	MXPOST	25.246	62,50	500.617	91,04	525.863	89,08
Óbreytt	Nei	fnTBL	21.823	54,03	502.378	91,36	524.201	88,80
Óbreytt	Nei	TnT	28.919	71,60	504.484	91,74	533.403	**90,36**
Óbreytt	Nei	MXPOST	25.252	62,50	500.611	91,04	525.863	89,08
Óbreytt	Já	fnTBL	28.461	70,44	503.142	91,50	531.603	90,06
Óbreytt	Já	TnT	34.859	86,28	505.511	91,93	540.370	**91,54**
Einfölduð*	Nei	MXPOST	25.261	62,52	508.747	92,52	534.008	90,46
Einfölduð	Já	fnTBL	28.467	70,46	509.788	92,71	538.255	91,18
Einfölduð	Já	TnT	34.863	86,29	513.797	93,44	548.660	**92,95**
Einf. f. kosn.		Kosn. v. með heildarnákv.	34.336	84,98	517.773	94,16	552.109	**93,53**
		MXPOST fram yfir kosn. m. heildarnkv.	34.013	84,18	518.818	94,35	552.831	**93,65**
Hlutfallsleg fækkun villna frá TnT að lokaniðurstöðu (90,36 – 93,65)								**34,15**

*Einföldun felst í að greina ekki atviksorð og ekki heldur samtengingar
Fornöfn eru sett í einn flokk en að öðru leyti er greining þeirra eftir kyni, tölu og falli látin haldast.

[1] Orðasafn hefur u.þ.b. helming óþekktra orða

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

- Hvers konar villur gera markarar og hversu margar gerðir af villum?
  - TnT gerir 5.373 mismunandi villur
  - fnTBL gerir 5.897 mismunandi villur
  - MXPOST gerir 7.115 mismunandi villur

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Algengustu villur sem TnT og MXPOST gera

MXPOST			TnT		
markari> rétt	Tíðni	%	markari> rétt	Tíðni	%
Samtals	64.434	100,00	Alls	56.894	100,00
aþ>ao	2.218	3,44	aþ>ao	1.734	3,05
ao>aþ	1.514	2,35	ao>aþ	1.489	2,62
aa>ao	616	0,96	ao>aa	1.045	1,84
ao>aa	599	0,93	aþ>aa	911	1,60
nveþ>nveo	586	0,91	nveþ>nveo	887	1,56
nveo>nveþ	547	0,85	nveo>nveþ	865	1,52
sfg3eþ>sfg1eþ	503	0,78	aa>ao	689	1,21
nhen>nheo	489	0,76	ssg>sþghen	671	1,18
sfg3fn>sng	446	0,69	nheo>nhen	659	1,16
c>ct	392	0,61	nhen>nheo	638	1,12
aa>aþ	378	0,59	sng>sfg3fn	599	1,05
nheo>nhen	371	0,58	sfg3eþ>sfg1eþ	584	1,03

Neshagi 16 · 107 Reykjavík · Sími: 525 4430
Bréfasími: 562 7242 · Netfang: ordabok@lexis.hi.is

Algengustu villur sem allir markarar eru sammála um

	Tíðni	%
Samtals	13.055	100,00
markari_rétt		
aþ_ao	499	3,82
sfg3eþ_sfg1eþ	457	3,50
ao_aþ	361	2,77
sng_sfg3fn	235	1,80
nveþ_nveo	214	1,64
nveo_nveþ	212	1,62
sfg3eþ_svg3eþ	203	1,55
ao_aa	190	1,46
lhensf_lheosf	170	1,30
fpkeþ_fpveþ	167	1,28
nhen_nheo	163	1,25
aa_ao	148	1,13
fohen_foheo	144	1,10
ct_c	141	1,08