Orðabók Háskólans

Mörkuð íslensk málheild

Greinargerð um tvö verkefni sem hafa verið styrkt af tungutækniverkefni menntamálaráðuneytisins

- Markari fyrir íslenskan texta (sept. 2002–feb. 2004)
- Mörkuð íslensk málheild (júní 2004–júní 2007)

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

Þessar upplýsingar má fá úr málheild (*corpus*)

- Safn tölvutækra texta af ýmsu tagi svo sem blaðatexta, fræðitexta af ýmsum sviðum, bókmenntatexta og talmáls.
- Nýtist „fyrirtækjum sem hráefni í afurðir“ (skýrsla starfshóps menntamálaráðuneytis um tungutækni)

- Mörkuð málheild (e. *tagged corpus*)
  - Orð í textum greind eftir orðflokkum og beygingu
  - Hverju orði fylgir mark (e. *tag*) og nefnimynd (*lemma*)
  - Haus sem gefur upplýsingar um textann, höfund hans o.fl.
- Mörkuð málheild er geymd í stöðluðu sniði (XML-útgáfa af TEI-sniði)

Brot úr haus fyrir skáldsöguna *Mín káta angist* eftir Guðmund Andra Thorsson

```
<title>Mín káta angist.</title>
<author born="1957" sex="m">Guðmundur Andri
  Thorsson</author>
<imprint>
<publisher>Uglan, íslenski
  kiljuklúbburinn</publisher>
<pubPlace>Reykjavík</pubPlace>
<date value="1990">1990</date>
</imprint>
```

7 orð úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson,mörkuð og í XML-sniði

```
<w type="ao" lemma="um">um</w>
<w type="nveo" lemma="leið">leið</w>
<w type="c" lemma="og">og</w>
<w type="fp1en" lemma="ég">ég</w>
<w type="sfg1eþ" lemma="láta">lét</w>
<w type="nkeog" lemma="miði">miðann</w>
<w type="sng" lemma="detta">detta</w>
<c type="punktur">.</c>
</s>
```

Notendur og notkun

Notendur: einstaklingar, fyrirtæki og stofnanir sem vinna að orðabókargerð, margvíslegum tungutækniverkefnum og rannsóknum á íslensku nútímamáli.

Undirstaða fyrir:

- þróun þýðingarforrita
- nútíma orðabókargerð
- þróun tungutæknitóla, t.d. fyrir talgreiningu og talgervingu
- þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði.

Nýtist sérstaklega fyrir blinda, heyrnarskerta og hreyfihamlaða og þá sem glíma við skriftar- og lestrarörðugleika.

Greining orða eftir orðflokkum og beygingu

tagging – **mörkun**

Greiningarstrengurinn

tag – **mark**

Forrit sem úthlutar mörkum

tagger – **markari**

Til þess að koma upp markaðri íslenskri málheild þarf **markara fyrir íslensku**

Markari getur einnig nýst í öðrum verkefnum

Not fyrir mörkun og markaðan texta

Mörkun er oft fyrsti verkþáttur í þessum verkefnum

- greining texta í setningahluta (*partial parsing*)
- nám orða fyrir gerð orðasafns
- upplýsingaheimt
- talkennsl
- talgerving
- vélrænar þýðingar
- fyrirspurnarkerfi
- tölfræðileg greining texta (tíðnigreining)
- leiðréttingarforrit

Vélrænar aðferðir við mörkun

Regluaðferðir

- Sérhvert orð í texta merkt með öllum hugsanlegum greiningarstrengjum með aðstoð orðasafns
- Notaðar reglur til þess að skera úr um hvaða greiningarstrengur er réttur

Námfúsir markarar, nota fyrir fram greint textasafn (*data-driven methods*)

Markmið verkefnisins

Búa til markara á sem stystum tíma sem getur greint íslenskan texta með a.m.k. 92% nákvæmni

Ákveðið að prófa námfúsa markara

Aðferðir og markarar sem voru prófaðir

- Tölfræðilegar aðferðir
  - falin Markovsíkön
    - **TnT**
  - hámarksóreiðuaðferð
    - **MXPOST**
- Aðrar aðferðir
  - leiðréttingaaðferð (*transformation-based learning*)
    - **µ-TBL** og **fnTBL**
  - minnistækni (*memory-based technique*)
    - **MBT**

Námfús markari þarf þjálfunarsafn

Býr til orðasafn með öllum orðmyndum í þjálfunarsafninu þar sem fram koma allar hugsanlegar greiningarmyndir (mörk)

Lærir af félagsskap orðanna hvaða mark sé líklegast

Lærir hvernig sé best að giska á mark orða sem koma ekki fyrir í þjálfunarsafninu

Ólíkir markarar nota ólíkar aðferðir til þess að skera úr um hvaða mark sé líklegast og hvernig sé best að marka óþekkt orð

Efniviður

Textasafn Íslenskrar orðtíðnibókar

(Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991)

590.297 lesmálsorð

59.358 orðmyndir

639 greiningarstrengir

að meðtöldum greinarmerkjum

Greining var leiðrétt handvirkt

Orðabók Háskólans

Mörkuð íslensk málheild

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
strætó	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

Markarar prófaðir

Skipta textasafni í þjálfunarsafn (90%) og prófunarsafn (10%)

Námfús markari 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

Niðurstaða af þjálfun og mörkun 10 para skráa

	Meðalnákvæmni		
	Öll orð	Þekkt orð	Óþekkt orð
Markari	%	%	%
fnTBL, meðalnákvæmni	88,80	91,36	54,02
MXPOST, meðalnákvæmni	89,08	91,04	62,51
TnT, meðalnákvæmni	90,36	91,74	71,62

Meðalhlutfall óþekktra orða: 6,84%

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

Algengustu villur sem allir markara gera

	Tíðni	%	Samanl. %
Samtals	13.055	100,00	100,00
markari_rétt			
aþ_ao	499	3,82	3,82
sfg3eþ_sfg1eþ	457	3,50	7,32
ao_aþ	361	2,77	10,09
sng_sfg3fn	235	1,80	11,89
nveþ_nveo	214	1,64	13,53
nveo_nveþ	212	1,62	15,15
sfg3eþ_svg3eþ	203	1,55	16,71
ao_aa	190	1,46	18,16

Orðabók Háskólans

Mörkuð íslensk málheild

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
.	.	.	.	.

Orðabók Háskólans

Mörkuð íslensk málheild

Hvernig má bæta niðurstöður mörkunar?

- Einfalda markaskrá
- Bæta mörkun óþekktra orða
  - Bæta aðferðir markaranna
  - Láta í té viðbótarorðasafn og ýmsa lista
- Kjósa á milli markara
- Beita reglum

Nákvæmni mörkunar þegar markaskrá er einfölduð

	Meðalnkv. fnTBL		Meðalnkv. MXPOST		Meðalnkv.i TnT	
	%	Samanl. %	%	Samanl. %	%	Samanl. %
Allur greiningarstrengur réttur	88,80	88,80	89,08	89,08	90,36	90,36
Atviksorð ekki greind	0,94	89,74	1,06	90,15	1,16	91,52
Samtengingar ekki greindar	0,14	89,88	0,19	90,34	0,18	91,70
Öllum fornöfnum slegið saman	0,10	89,98	0,13	90,46	0,13	91,83
Aðeins orðflokkur réttur	7,27	97,25	6,83	97,29	6,30	98,14
Rangur orðflokkur	2,75	100,00	2,71	100,00	1,86	100,00
Samtals	100,00		100,00		100,00	

Lokaniðurstöður

Markaskrá	Orða-safn[1]	Aðferð	Óþekkt orð %	Þekkt orð %	Öll orð %
Óbreytt	Nei	MXPOST	62,50	91,04	89,08
Óbreytt	Já	fnTBL	70,44	91,50	90,06
Óbreytt	Já	TnT	86,28	91,93	**91,54**
Óbreytt		Kosið milli markara	84,97	93,12	92,56
Einfölduð[2]		MXPOST	62,52	92,52	90,46
Einfölduð[2]		fnTBL	70,46	92,71	91,18
Einfölduð[2]		TnT	86,29	93,44	**92,95**
Einfölduð		Kosning	84,98	94,16	**93,53**
		Reglur	84,18	94,35	**93,65**
Orðflokkar		MXPOST	88,19	97,96	97,29
		fnTBL	82,15	98,35	97,25
		TnT	92,47	98,55	98,14

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

[2] 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.

Stefnt skal að:

- 25.000.000 orð
- 900–1.000 textabútar
- allt 40.000 orð hver, 10% sleppt ef texti lengri
- 90% nákvæmni í mörkun
- um 1.000.000 orð þar sem mörkun er leiðrétt

Val texta

Uppruni

- 60% bækur
- 25% blöð og tímarit
- 5–10% annað útgefið efni
- 5–10% óútgefið efni
- <5% efni ætlað til upplestrar

Val texta

Efnisval

- 25% skáldverk
- 75% nytjatexti
  - hagnýtt vísindi, náttúrufræði, þjóðfélagsfræði, heimsmál, viðskipti, listir, trúarbrögð, heimspeki, tómstundir,...

- Hverju orði fylgir mark og nefnimynd
- Málheildin verður skráð með stöðluðu sniði
  - XML-útgáfa af TEI-sniði fyrir málheildir
  - (TEI: *Text Encoding Initiative*)

Brot úr haus fyrir skáldsöguna *Mín káta angist* eftir Guðmund Andra Thorsson

```
<title>Mín káta angist.</title>
<author born="1957" sex="m">Guðmundur Andri
  Thorsson</author>
<imprint>
<publisher>Uglan, íslenski
  kiljuklúbburinn</publisher>
<pubPlace>Reykjavík</pubPlace>
<date value="1990">1990</date>
</imprint>
```

7 orð úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson,mörkuð og í XML-sniði

```
<w type="ao" lemma="um">um</w>
<w type="nveo" lemma="leið">leið</w>
<w type="c" lemma="og">og</w>
<w type="fp1en" lemma="ég">ég</w>
<w type="sfg1eþ" lemma="láta">lét</w>
<w type="nkeog" lemma="miði">miðann</w>
<w type="sng" lemma="detta">detta</w>
<c type="punktur">.</c>
</s>
```

Söfnun texta

- Ekki verður greitt fyrir afnot af efni
- Rétthafar fá nákvæmar upplýsingar um hvernig aðgangur verður veittur að málheildinni
- Textum fyrst safnað í textasafn OH
- Textabútar fluttir úr textasafni OH í málheild til frekari úrvinnslu

Orðabók Háskólans

Mörkuð íslensk málheild

Mörkun lesmálsorða

- Notaðar niðurstöður markaraverkefnis
- Búið til orðasafn með orðmyndum, nefnimyndum og mörkum úr beygingarlýsingu íslensks nútímamáls
- Hjálparskrár: mannanöfn, örnefni, heiti fyrirtækja og stofnana, skammstafanir...

Vinnulag

Skilgreina aðferðir, afla forrita og búa þau til

Semja við rétthafa og afla texta

Undirbúa texta fyrir vinnslu og efnisflokka þá

Forvinnsla (málsgreinaskil, lesmálsorð, tölur, skammstafanir...)

Mörkun

Leiðrétta mörkun valdra texta („bootstrapping“)

Koma textum í XML-snið

Orðabók Háskólans

Mörkuð íslensk málheild

Varðveisla og rekstur

Óráðið

Íslenskur markari

Samningur við menntamálaráðuneyti frá 18.10.2002. Verkið unnið okt. 2002–feb. 2004

Verktakar: Orðabók Háskólans og Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir)

Verkefnisstjóri: Eiríkur Rögnvaldsson

Mörkuð íslensk málheild

Samningur við menntamálaráðuneyti frá 14.6.2004. Verkið unnið júní 2004–júní 2007

Verktaki: Orðabók Háskólans

Verkefnisstjóri: Sigrún Helgadóttir

Verkefnisstjórn: Eiríkur Rögnvaldsson, Ásta Svarvarsdóttir, Kristín Bjarnadóttir

Annað starfsfólk: Auður Þórunn Rögnvaldsdóttir o.fl.