Málheild sem hluti af orðabókarlýsingu

Sigrún Helgadóttir

Stofnun Árna Magnússonar í íslenskum fræðum

sigruhel@hi.is

Hugvísindaþing 10. mars 2012

Hvað er mörkuð málheild?

Mörkuð málheild (e. *tagged corpus*)

- Safn fjölbreyttra tölvutækra textabúta sem hafa verið greindir á málfræðilegan hátt
- Hverjum texta fylgja upplýsingar um textann sem búturinn er úr
- Hverri orðmynd fylgja þær málfræðilegu upplýsingar sem málheildin á að geyma
- Málheildin er skráð í stöðluðu sniði

Hvers konar málheild?

Textaval

Sérhæfð málheild

- Dæmi: Dagblaðatextar, textar um læknisfræði o.s.frv.
- Notkun: kanna málfar á tilteknu sviði

Málheild með fjölbreyttum textum (*balanced*)

- Dæmi: Mörkuð íslensk málheild (MÍM)
- Notkun: kanna almennt málfar

Hvers konar málheild?

Tímarammi:

Samtíma málheild (*synchronous*)

- Allir textar skrifaðir á tilteknu tímabili, t.d. 2000-2009 (MÍM)
- Notkun: kanna málfar á tilteknu tímabili

Söguleg málheild (diachronic)

- Textar frá ýmsum tímum. Dæmi: Íslenski trjábankinn (IcePaHC)
- Notkun: kanna málbreytingar í tíma

MÍM: Textarnir

- Ríflega 26 milljón lesmálsorð af frumsömdum textum ritaðir á árunum 2000–2009 eftir höfunda sem hafa íslensku að móðurmáli
- Textar úr útgefnum bókum eru styttir um 20%; aðrir textar eru óstyttir
- Aðeins var safnað textum sem voru aðgengilegir í rafrænu formi
- Leyfis var aflað hjá rétthöfum til þess að fá að nota alla texta í safninu

MÍM: Höfundarréttur

- Leyfis var aflað til þess að fá að nota alla texta (að undanskildum textum frá opinberum aðilum sem eru ekki verndaðir af höfundarrétti, 9. gr. 73/1972)
- Rétthafar fengu upplýsingabækling um málheildina, uppkast að notkunarleyfi og samþykkisyfirlýsingu (sjá http://www.arnastofnun.is/page/arnastofnun_ord_malheild)
- Rétthafar prentaðra bóka fengu gögn í pósti, aðrir í tölvupósti
- Ein áminning var send
- Gert var samkomulag við Rithöfundasamband Íslands, Hagþenki og Félag íslenskra bókaútgefenda

Textaflokkar í Markaðri íslenskri málheild	fjöldi orða	%
Ræður fluttar á Alþingi	250.000	1,0
Blogg	1.964.495	7,6
Dagblöð (Morgunblaðið, Fréttablaðið)	7.222.133	27,9
Dómar	316.134	1,2
Efni til upplestrar	222.872	0,9
Fréttir útvarps og sjónvarps	287.554	1,1
Skýrslur og greinargerðir af vefsetrum ráðuneyta	1.658.618	6,4
Frumvörp og lög af vef Alþingis	747.914	2,9
Lokaritgerðir háskólastúdenta	485.165	1,9
Stúdentsprófsritgerðir í íslensku	178.949	0,7
Af vefsetrum fyrirtækja, samtaka og stofnana	1.594.504	6,2
Textavarp	42.520	0,2
Safnaðarblöð	6.472	0,0
Texti um tónlist	24.357	0,1
Prentuð tímarit af ýmsu tagi	2.243.084	8,7
Vefmiðlar	243.750	0,9
Veftímarit	145.399	0,6
Tölvupóstlistar	121.164	0,5
Pistlar af Vísindavef	1.770.184	6,8
Textar úr bókum	5.770.545	22,3
Talmál	574.732	2,2
Samtals	**25.870.545**	**100,0**

MÍM: Undirbúningur texta fyrir leitarkerfi

- Textar dregnir úr umbrotssniði og hreinsaðir
- Textum er skipt í setningar
- Hverri setningu er skipt í lesmálsorð
- Hverju orði er úthlutað greiningarstreng (mark) sem segir til um orðflokk og aðrar beygingarlegar upplýsingar
- Hverju orði er einnig úthlutað flettimynd (nefnimynd, lemmu), t.d. nf. et. nafnorða, nh. sagna
- Textum fylgja bókfræðilegar upplýsingar

MÍM: Tilreiðsla, skipting í setningar, mörkun, lemmun

- Notaður er hugbúnaður sem er hluti af *IceNLP toolkit* (Hrafn Loftsson) til þess að skipta texta í orð og setningar
- Mörkun er gerð með *CombiTagger,* forriti sem sameinar niðurstöðu fjögurra markara
- Nákvæmni mörkunar á textum MÍM hefur verið metin 88-95% eftir textum
- Lemmun er gerð með Lemmaldi Antons K. Ingasonar (nákvæmni hefur verið metin um 90% fyrir texta úr Morgunblaðinu)
- Sérstakt forrit (WorkerBranch) setur af stað alla þrjá verkþætti, tilreiðslu, mörkun og lemmun.

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

Mynd 1. Greining orða í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Hvernig verður málheildin notuð?

Mörkuð íslensk málheild verður aðgengileg á tvennan hátt:

1. Á vefsetri Stofnunar Árna Magnússonar í íslenskum fræðum verður leitarbær útgáfa hennar
2. Textar verða aðgengilegir sem xml-skrár í sniði fyrir málheildir sem er skilgreint af *Text Encoding Initiative* (TEI). Bókfræðilegar upplýsingar, mörk og lemmur verða hluti af textunum. Notendur sem vilja fá textana í tölvur sínar skrifa undir notkunarleyfi (sjá dæmi á vef SÁ).

MÍM: Leitarkerfi

Hluti af málheildartextum er þegar leitarbær á vefsetrinu http://mim.hi.is

- Mörkuð íslensk málheild (alls 17.692.940 lesmálsorð)
  - Morgunblaðið, Bækur (154 bækur), Vísindavefur (39 höfundar), RÚV (fréttir útvarps og sjónvarps), Fréttablaðið, Blogg (bloggfærslur almennra bloggara, guðfræðinga og stjórnmálamanna)
- Orðtíðnibók = textar Orðtíðnibókar (þeir textar sem leyfi hefur fengist fyrir); handleiðrétt mörk
- Fornrit = 44 sögur úr útgáfu Svarts á hvítu
  - Heimskringla, Sturlunga, Landnámabók, Íslendingasögur
- Talmál (Verður aðgengilegt í sumar, líka með hljóði)

MÍM: Leitarkerfi

- Leitarkerfið er byggt á ***Glossa*** frá háskólanum í Osló (http://www.hf.uio.no/tekstlab/glossa.html)
- Í *Glossa* er notast við leitarvélina IMS Corpus Workbench (CWB) frá háskólanum í Stuttgart (http://cwb.sourceforge.net/)
- Bráðabirgðaútgáfa af leitarviðmóti er tilbúin:
  - Leitarmöguleikar hafa verið lagaðir að íslenskum mörkum
  - Vefviðmótið hefur verið þýtt
  - Bókfræðilegar upplýsingar verða tiltækar með vorinu
  - Gerðar verða ráðstafanir til þess að notendur geti valið texta til þess að leita í

Hvernig má nota málheildina sem orðabók?

- Greining orða í málheildinni leyfir markvissa leit:
  - Leita má að öllum orðmyndum tiltekins orðs út frá nefnimynd (flettimynd)
  - Leita má að t.d. nafnorðum í þgf.
  - Leita má að fleiri en einu orði í einu
  - Leita má að tveimur eða fleiri orðum þar sem ótiltekin orð eru á milli þeirra
- Atriði sem bætt verður við í vor og sumar:
  - Velja texta til leitar
  - Sjá úr hvaða texta leitarniðurstaða er fengin

Hvernig má nota málheildina sem orðabók?

- Niðurstöður birtast sem orðstöðulykill
  - Tengill verður framan (eða aftan) við niðurstöðulínu þar sem fá má upplýsingar um textann sem dæmið er tekið úr
  - Möguleiki til þess að raða dæmunum
  - Möguleiki til þess að fá upplýsingar um tíðni orðmynda eða “orðasambanda”

Hverjir nota málheildina sem orðabók?

- Notandi sem vill finna fleiri notkunardæmi um tiltekið orð en hann finnur í orðabók getur fundið þau í málheildinni
- Orðabókarhöfundur getur leitað að notkunardæmum fyrir verk sitt
- Hugsanlegt er að tengja málheildina við rafrænar orðabækur
  - Notanda er gefinn kostur á að smella á leitarorðið og er þá vísað í málheildina og fær orðstöðulykil með dæmum
  - Einnig er möguleiki að fá upplýsingar um tíðni orðmynda
  - Dæmi: Íslex, Beygingarlýsing íslensks nútímamáls

Málföng fyrir íslensku

Málföng er þýðing á “language resources”, þ.e. bæði tól (forrit) og gagnasöfn.

Á vegum verkefnisins META-NORD er verið að búa til vefsetur:

www.málföng.is

Þar má finna ýmis verkefni sem hafa verið unnin á Íslandi og snúast um máltækni.

Leit í MÍM

Tengill er af síðunni www.málföng.is á

http://mim.hi.is/

(Þessi miðlari virkar tímabundið ekki fullkomlega, vantar “röðun” og “tíðni”)

Notum í staðinn tilraunamiðlara:

http://130.208.176.60//index.php?corpus=mim

Leit í MÍM

- Leitum að orðinu “hestur” til þess að sjá hvernig kerfið virkar
- Leitum að setningasamböndum á borð við

“björt rödd”, þ.e. hvaða lýsingarorð eru notuð með “rödd” (lo.+rödd, nefnimynd)

“borða sig saddan” (so+sig+lo.þf.)

“teyma hestinn”, þ.e. hvaða sagnir eru notaðar með “hestur” (so+hestur, nefnimynd)

“borða matinn” (borða, éta, smakka+no.þf.)

“köflóttur kjóll” (lo.+fatnaðarorð)

“í hendi sér” (fs.+no+afturbeygt fornafn)

Hvernig má nota málheildina sem orðabók?

- Notandi sem hefur flett upp orði í orðabók en óskar eftir að fá fleiri notkunardæmi getur fundið þau í málheildinni
- Orðabókarhöfundur getur leitað að notkunardæmum

MÍM: Leitarkerfi

- Mörkuð íslensk málheild (alls 17.692.940 lesmálsorð)
  - Morgunblaðið, textar úr völdum blöðum af Morgunblaðinu 2002–2008 (5.840.345 lesmálsorð)
  - Bækur, 154 bækur (6.786.611 lesmálsorð)
  - Vísindavefur, 39 höfundar (1.952.344 lesmálsorð)
  - Fréttahandrit útvarps og sjónvarps (314.203 lesmálsorð)
  - Fréttablaðið, textar úr 18 tölublöðum af Fréttablaðinu frá 2002–2007 (580.595 lesmálsorð)
  - Blogg, 2.218.842 lesmálsorð af textum úr bloggfærslum almennra bloggara, guðfræðinga og stjórnmálamanna

MÍM: Leitarkerfi

- Leitarkerfið er byggt á ***Glossa*** frá háskólanum í Osló (http://www.hf.uio.no/tekstlab/glossa.html)
- Í *Glossa* er notast við leitarvélina IMS Corpus Workbench (CWB) frá háskólanum í Stuttgart (http://cwb.sourceforge.net/)
- Bráðabirgðaútgáfa af leitarviðmóti er tilbúin:
  - Leitarmöguleikar hafa verið lagaðir að íslenskum mörkum
  - Vefviðmótið hefur verið þýtt
  - Bókfræðilegar upplýsingar verða tiltækar með vorinu
  - Gerðar verða ráðstafanir til þess að notendur geti valið texta til þess að leita í

MÍM: Leitarkerfi- eftir hverju má velja texta?

- Aldur lesenda (fullorðnir, börn, unglingar, börn/unglingar, allir)
- Textaflokkar (einhvers konar upprunaflokkun: blogg, textar úr bókum, dagblöð, prentuð og á vef, efni til upplestrar, opinberir textar (dómar, lög, reglugerðir o.fl.), skólaritgerðir, textavarp, tímarit (prentuð og á vef), tölvupóstlistar, vefsetur - ýmsir – bæklingar – fréttabréf, pistlar af vísindavef, talmál
- Velja einstaka texta eða raða textum saman eftir óskum notandans

MÍM: Heimildir

- TEI: Text Encoding Initiative

  http://www.tei-c.org/index.xml
- IceNLP hugbúnaðurinn er aðgengilegur hér:

  http://icenlp.sourceforge.net/
- CombiTagger er aðgengilegur hér:

  http://combitagger.sourceforge.net/
- CombiTagger sameinar niðurstöður þessara markara:
  - IceTagger (hluti af IceNLP toolkit)
  - TnT (Brants)
  - fnTBL (Ngai og Florian)
  - MXPOST (Ratnaparkhi)

MÍM: Samstarfsmenn við verkefnið

- Verkefnisstjórn: Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir
- Auður Rögnvaldsdóttir (forritun og fleira í upphafi verks)
- Eyrún Valsdóttir (textaöflun og fleira)
- Hjördís Stefánsdóttir (textaöflun, textahreinsun og fleira)
- Guðmundur Örn Leifsson (setti upp Glossa-hugbúnaðinn)
- Steinþór Steingrímsson (vinnur nú við hugbúnað málheildar)
- Stjórnendur og starfsmenn Orðabókar Háskólans og síðar Stofnunar Árna Magnússonar í íslenskum fræðum

MÍM: Fjármögnun

- Tungutækniverkefni menntamálaráðuneytisins
- Orðabók Háskólans/Stofnun Árna Magnússonar í íslenskum fræðum
- Rannís
- Nýsköpunarsjóður námsmanna
- Rannsóknarsjóður Háskólans
- META-NORD

MÍM: Afleidd verkefni

- Nýsköpunasjóður námsmanna sumarið 2006. Námsmenn: Sigrún Andrea Ásgeirsdóttir, Anton Karl Ingason og Skúli Bernharð Jóhannsson. Umsjónarmenn: Sigrún Helgadóttir o.fl.
- Vefviðmót fyrir texta Orðtíðnibókar, með því að nota **Xaira**, forrit sem fylgir BNC (British National Corpus).
- Markmið verkefnisins var að kanna hvort nota mætti Xaira til þess að búa til leitarviðmót fyrir MÍM. Kerfið sem var búið til var aðallega ætlað nemendum og gaf m.a. skemmtilegt tæki til þess að kenna málfræðilega greiningu. Ekki þótti fýsilegt að nota Xaira fyrir málheildina sjálfa. Kerfið var aldrei opnað af ýmsum ástæðum.

MÍM: Afleidd verkefni

Staðalmálheild – gullstaðall

- Nýsköpunarjsóður námsmanna 2009. Námsmaður: Jökull Huxley Yngvason. Umsjónarmenn: Hrafn Loftsson, Eiríkur Rögnvaldsson og Sigrún Helgadóttir.
- Markmið verkefnisins var að búa til nýja staðalmálheild fyrir þróun og þjálfun margvíslegra máltæknieininga.
- Textasafn ÍO (Íslenskrar orðtíðnibókar) er of lítið (590 þús. lesmálsorð) og hefur bókmenntalega slagsíðu.
- Tekið var úrtak úr MÍM með um milljón lesmálsorðum.
- Þróað var kerfið *WorkerBranch* fyrir tilreiðslu, mörkun og lemmun, þetta kerfi hefur síðan verið notað fyrir MÍM.

MÍM: Afleidd verkefni

Staðalmálheild – gullstaðall

- Leitað var að villum með því að grunnþátta textana og athuga ósamræmi í nafnliðum og sagnliðum (Hrafn Loftsson. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). Athens, Greece.).
- Villurnar voru leiðréttar og nákvæmni mörkunar var metin með því að skoða mark um hundraðasta hvers orðs. Mörkunarnákæmni var metin 88–95%. (Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir and Eiríkur Rögnvaldsson. Developing a PoS-tagged corpus using existing tools. In Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation, LREC 2010. Valetta, Malta.)

MÍM: Afleidd verkefni

Staðalmálheild – gullstaðall

- Nýsköpunarjsóður námsmanna 2010. Námsmaður: Kristján Friðbjörn Sigurðsson. Umsjónarmenn: Eiríkur Rögnvaldsson, Sigrún Helgadóttir og Hrafn Loftsson.
- Kristján fór yfir um 230 þúsund orðmyndir og leiðrétti mörk um 7,25% orðmynda.
- Kristján hélt áfram að leiðrétta með skóla í vetur og mun vinna við verkefnið í sumar. Þá er þess vænst að farið hafi verið yfir alla texta gullstaðalsins.
- Næsta skref verður að leiðrétta nefnimyndir.
- Stefnt er að því að gullstaðallinn verði tilbúinn til notkunar árið 2012.