Mörkuð íslensk málheild (MÍM)

Sigrún Helgadóttir

Stofnun Árna Magnússonar í íslenskum fræðum

Erindi flutt á málstofunni Stefnumót: Á mörkum málfræði og tölvutækni á Hugvísindaþingi 26. mars 2011.

sigruhel@hi.is

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

MÍM: Textarnir

- Ríflega 26 milljón lesmálsorð af frumsömdum textum frá 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 um tölvupóst
- 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: Hreinsun texta

- Textarnir fengust í alls konar sniði
- Texti var dreginn úr pdf-skjölum og Word-skjölum
- Textar voru hreinsaðir handvirkt og með forritum
- Mörg “erfið” tákn þurfti að laga, t.d. gæsalappir, úrfellingarmerki, bandstrik,…
- Notuð eru táknin « og » fyrir allar gæsir
- Í sumum textum þurfti að aðgreina fyrirsagnir
- Textum verður dreift í UNICODE (UTF8) stafatöflu…….

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 nokkurra 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.

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)
- Orðtíðnibók = textar Orðtíðnibókar (þeir textar sem leyfi hefur fengist fyrir); handleiðrétt mörk
- Fornrit = 44 sögur úr útgáfu Svarts á hvítu (1.659.385 lesmálsorð)

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 leita?

- 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