Orðabók Háskólans

Mörkuð íslensk málheild

Málstofa Stofnunar Árna Magnússonar í íslenskum fræðum 23. maí 2008

Mörkuð íslensk málheild

Sigrún Helgadóttir

Eyrún Valsdóttir

Anton Karl Ingason

Efni kynningar

Sigrún Helgadóttir, inngangur

Eyrún Valsdóttir, yfirlit yfir efnisöflun

Hlé til þess að fylla á glös!

Anton Karl Ingason, hvernig Xaira-forritið var notað fyrir leit í textum íslenskrar orðtíðnibókar, hugmyndir um flutning texta í Málheildina og um almennara leitarforrit

Sýnd leit með vefaðgangi að textum Orðtíðnibókar þar sem Xaira-forritið er notað og leit í textum málheildar í Textasafni SÁ

Spurningar og svör

Mörkuð íslensk málheild

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

Verktaki: Orðabók Háskólans

Verkefnisstjóri: Sigrún Helgadóttir

Verkefnisstjórn: Eiríkur Rögnvaldsson, Ásta Svavarsdóttir, Kristín Bjarnadóttir.

Hvað er mörkuð málheild (e. *tagged corpus*)?

Safn fjölbreyttra tölvutækra textabúta sem hafa verið greindir á málfræðilegan hátt

Hverjum textabút fylgja upplýsingar um textann sem búturinn er úr

Hverri orðmynd fylgja þær málfræðilegu upplýsingar sem málheildin á að geyma

Málheildin er skráð með stöðluðu sniði

Mörkuð íslensk málheild

Stefnt að 25.000.000 lesmálsorðum sem komi úr 900–1.000 textabútum

Hver bútur allt að 40.000 lesmálsorð, 10% sleppt ef texti styttri

Hverjum textabút fylgja helstu bókfræðilegar upplýsingar

Hverri orðmynd fylgir mark (e. *tag*) og nefnimynd (e. *lemma*)

- Mark segir til um orðflokk og málfræðilegar upplýsingar, t.d. kyn, tölu og fall nafnorða

Málheildin verður skráð með stöðluðu sniði

Notuð verður XML-útgáfa af TEI-sniði fyrir málheildir

(TEI: *Text Encoding Initiative*)

Notendur og notkun

Notendur málheildarinnar eru einstaklingar, fyrirtæki og stofnanir sem vinna að orðabókagerð, margvíslegum tungutækniverkefnum og rannsóknum á íslensku nútímamáli.

Úr málheildinni má lesa ýmiss konar gagnlegan fróðleik, t.d.

- Upplýsingar um tíðni orðflokka, orða og beygingarmynda,
- Upplýsingar um orðasambönd, setningargerð og merkingu
- Upplýsingar um hvernig tungumálið er notað á tilteknum tíma, vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti.

Undirstaða fyrir:

- þróun þýðingarforrita
- nútíma orðabókagerð
- þróun tungutæknitóla, t.d. fyrir talgreiningu og talgervingu
- þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði.
- Mörg tungutæknitól nýtast sérstaklega fyrir blinda (t.d vefþulan Ragga), heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika (Ragga aftur o.fl.).

Hvernig eru textar valdir í Málheildina?

Útgáfutími: 2000 og seinna

Uppruni (upphaflegar hugmyndir)

- 60% úr bókum
- 25% úr blöðum og tímaritum
- 5–10% úr öðru útgefnu efni (bæklingar, auglýsingapésar o.s.frv.)
- 5–10% úr óútgefnu efni (persónuleg bréf, dagbækur, ritgerðir, minnisblöð o.s.frv.)
- <5% úr efni ætluðu til upplestrar (pólitískar ræður, leikrit, handrit til upplestrar í útvarpi, stólræður o.s.frv.)
- Talmál (Ístal, þingræður, ...)

Hvernig eru textar valdir?

Uppruni (líkleg skipting miðað við það sem þegar hefur verið safnað)

< 45 % úr bókum

35 % úr blöðum og tímaritum

8 % blogg

8 % ýmislegt vefefni (vefsetur, Vísindavefur)

4 % ýmislegt óútg. (til upplestrar, t.d. stólræður, fréttahandrit útvarps), tölvupóstur, lög og dómar, skólaritgerðir, textavarp, ýmsir bækl.)

? Talmál (Ístal, þingræður, ...)

Ekki verður gert ráð fyrir þýðingum

(Meira frá Eyrúnu seinna)

Hvernig eru textar valdir?

Efnisval (upphaflegar hugmyndir)

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

Endanleg skipting verður ljós þegar meira efni hefur verið safnað

Hvar má finna efnið?

- Beint frá útgefendum (bækur, blöð og tímarit)
- Af vefsíðum (blöð, tímarit, ýmislegt annað útgefið og óútgefið (blogg, tölvupóstur, greinar o.s.frv.)
- Frá höfundum beint
- Nánar í greinargerð Eyrúnar

Öflun texta

- Ekki verður greitt fyrir afnot af efni
- Haft samband við rétthafa með tölvupósti eða með bréfi, rétthafar fá upplýsingar um málheildina, drög að notkunarleyfi og undirrita samþykkisyfirlýsingu
- Textum fyrst safnað í textasafn OH
- Textabútar fluttir úr textasafni OH í málheild til frekari úrvinnslu

Stofn Markaðrar íslenskrar málheildar er textasafn *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem (1991). Orðabók Háskólans.)

100 textar, um 5000 orð hver (500.000 lesmálsorð)

frumsamin íslensk skáldverk	20 %
þýdd skáldverk	20 %
ævisögur	20 %
nytjatexti	20 % (hugvísindi og raunvísindi)
barnabækur	20 % (frumsamdar og þýddar)

Hafðir með þeir textar sem leyfi fæst fyrir, þýðingum sleppt

Öflun leyfa

Afla þarf tvenns konar leyfa frá rétthöfum texta:

- Leyfi til þess að geyma heila texta í textasafni OH og veita að þeim takmarkaðan aðgang
- Leyfi til þess að geyma textabúta í Markaðri íslenskri málheild og veita að þeim nánast ótakmarkaðan aðgang

Aflað var leyfa frá rétthöfum texta sem ekki komu úr bókum samhliða efnisöflun

Rithöfundasamband Íslands og Hagþenkir hafa lýst yfir stuðningi við verkefnið

Stjórn félags íslenskra bókaútgefenda vísar til einstakra útgefenda

Lögfræðingur, sérhæfður í höfundarétti leiðbeindi um samningu samþykkisyfirlýsinga og notkunarleyfis.

Vinnslustig

- Rafrænn texti fenginn frá rétthafa eða útgefanda (hreinn texti, Word-skjal, pdf-skjal, annað umbrotskerfi t.d. Quark)
- Textinn dreginn úr umbroti, breytt um stafatöflu (UTF-8 > Latin1), ýmis tákn löguð o.s.frv.
- Bókfræðilegar upplýsingar skráðar (titill, höfundur, ártal, útgefandi, ýmis flokkun, o.s.frv.)

Vinnslustig

Mörkun og önnur úrvinnsla

- Tilreiðsla (tokenization, texta skipt í orð með sérstöku forriti)
- Málfræðileg mörkun (IceTagger eða samsettur markari)
- Lemmun (Lemmald, lemmari Antons)
- Texti færður í xml-snið (TEI, Text Encoding Initiative)

Skrá komið fyrir í gagnasafni

(Sjá einnig tillögur Antons á eftir)

Aðgangur að málheildinni

- Heilir textar fara fyrst í textasafn OH
  - Takmarkaður uppflettiaðgangur á vefsetri OH (t.d. 50–300 orð á undan og eftir orði sem leitað er að). Textar úr Morgunblaðinu, af Vísindavef og bloggtextar þegar aðgengilegir í Textasafni OH, sýnt á eftir
- Bútar úr textum fara í Markaða íslenska málheild
  - Leitaraðgangur á vefsetri OH með sérstöku leitarforriti, sjá dæmi á eftir
  - Dreifing á geisladiskum eða með því að sækja skrár á vefsetur OH. Notendur undirrita notkunarleyfi eða samþykkja á vefsetri

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <mimDoc id="A1A">
- <teiHeader>
- <fileDesc>
- <titleStmt>
  <title>Mín káta angist. Skáldsaga. Íslensk skáldverk.</title>
.
 <monogr>
 <title>Mín káta angist</title>
 <author born="1957">Guðmundur Andri Thorsson</author>
-      <imprint>
-        <publisher>Mál og menning</publisher>
  <pubPlace>Reykjavík</pubPlace>
 <date value="1988">1988</date>
 </imprint>
  </monogr>
.
</teiHeader>
```

<s n="4">
<w type="fp1en" lemma="ég">Ég</w>
<w type="sfm1eþ" lemma="setja">settist</w>
<w type="ao" lemma="fyrir">fyrir</w>
<w type="aa" lemma="framan">framan</w>
<w type="tfvfo" lemma="tveir">tvær</w>
<w type="lvfosf" lemma="gamall">gamlar</w>
<w type="nvfo" lemma="kona">konur</w>
<w type="ao" lemma="með">með</w>
<w type="nkfo" lemma="hattur">hatta</w>
<w type="ct" lemma="sem">sem</w>
<w type="sfg3fþ" lemma="segja">sögðu</w>
<w type="foven" lemma="hvor">hvor</w>
<w type="foveþ" lemma="annar">annarri</w>
<w type="au" lemma="já">jájá</w>
<w type="aa" lemma="þangað">þangað</w>
<w type="aa" lemma="til">til</w>
<w type="foveþ" lemma="annar">annarri</w>
<w type="sfm3eþ" lemma="hugkvæmast">hugkvæmdist</w>

Erlendar málheildir

- BNC, British National Corpus 100 M lesmálsorð (http://www.natcorp.ox.ac.uk/)
- ANC, American National Corpus, 22 M lesmálsorð (http://americannationalcorpus.org/)
- Korpus 2000, dönsk málheild 50+ M lesmálsorð http://korpus.dsl.dk/korpus2000/
- The Brown Corpus frá 1961 1M lesmálsorð
- LOB (London-Osló-Bergen), Umeå o.s.frv.