Orðabók Háskólans

Mörkuð íslensk málheild

Ráðstefna um íslenska tungutækni

Mörkuð íslensk málheild

Sigrún Helgadóttir, tölfræðingur

Orðabók Háskólans

sigrunh@lexis.hi.is

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 Svavarsdóttir, Kristín Bjarnadóttir.

Efnisyfirlit

- Hvað er mörkuð málheild?
- Íslenska málheildin
- Hvernig eru málheildir notaðar?
- Efnissöfnun, skipting efnis eftir uppruna og innihaldi
- Hvar má finna efnið?
- Dæmi (Íslenskri orðtíðnibók breytt í markaða málheild)
- Aðgangur að íslenkri málheild – leyfi til notkunar texta
- Erlendar málheildir, dæmi

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 lengri

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
- Stefnt verður að 90% nákvæmni í mörkun

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 orðabókargerð, 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 tiltekið tungumál er notað á tilteknum tíma, vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti.

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.
- Mörg tungutæknitól nýtast sérstaklega fyrir blinda, heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika.

Hvernig eru textar valdir?

Útgáfutími: 2000 og seinna

Uppruni

- 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

Hvernig eru textar valdir?

Efnisval

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

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.)

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)

```
<?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>

Aðgangur

- 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)
- Bútar úr textum fara í Markaða íslenska málheild
  - Leitaraðgangur á vefsetri OH með sérstöku uppflettiforriti (skoða má heila textabúta)
  - Dreifing á geisladiskum eða með því að sækja skrár á vefsetur OH

Leyfi

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

Það er verkefni næstu mánaða að finna leið til þess að afla þessara leyfa án þess að greitt sé fyrir afnotin

Mismunandi er eftir textum hvernig höfundarrétti er háttað

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.

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

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