

Mörkuð íslensk málheild

Hvað er málheild?

Málheild (e. *corpus*) er safn fjölbreyttra texta sem eru geymdir í stöðluðu sniði í rafrænu formi. Til þess að textarnir verði sem gagnlegastir við málrannsóknir eru þeir greindir á margvíslegan hátt. Oft er hverri orðmynd látinn fylgja greiningarstrengur, **mark** (e. *tag*), sem sýnir orðflokk og önnur málfræðileg atriði. t.d. kyn, tölu og fall fallorða, og persónu, tölu og tíð sagna. Einnig fylgir oft hverri orðmynd **nefnimynd** (e. *lemma*) sem er t.d. nefnifall í eintölu fyrir fallorð og nafnháttur sagna. Málheild þar sem hverri orðmynd fylgir mark og nefnimynd kallast **mörkuð málheild**. Hverjum texta í málheildinni fylgja einnig bókfræðilegar upplýsingar um verkið sem textinn er úr.

Notkun og gagnsemi málheilda

Málheildir fela í sér upplýsingar um hvernig tiltekið tungumál er notað á tilteknum tíma. Þær gefa vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti. Úr málheildinni má lesa ýmiss konar gagnlegan fróðleik, t.d. um tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerðir og merkingu orða.

Málheildir eru mikilvægar heimildir við gerð orðabóka og handbóka um mál og málnotkun og margir útgefendur orðabóka leggja mikið upp úr því að orðabókaritstjórar á þeirra vegum hafi aðgang að stórum mörkuðum málheildum. Má m.a. nefna að þrjú öflug orðabókaförlög (Oxford University Press, Addison-Wesley Longman og Larousse Kingfisher Chambers) tóku þátt í gerð bresku málheildarinnar *BNC* (British National Corpus) sem var sett saman í byrjun 10. áratugar síðustu aldar.

Markaðar málheildir gegna veigamiklu hlutverki við þróun þýðingarforrita og upplýsingar sem úr þeim fást má einnig nota við gerð ýmiss konar tungutækni-búnaðar, t.d. fyrir talgreiningu og talgervingu. Slíkar upplýsingar eru einnig nauðsynlegar við þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði. Ýmis tungutækni-búnaður af þessu tagi nýtist sérstaklega fyrir blinda, heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika.

Mörkuð íslensk málheild

Árið 2004 var hafist handa við gerð markaðrar málheildar með íslenskum textum. Verkefnið er samstarfsverkefni menntamálaráðuneytisins sem fjármagnar verkið og orðfræðisviðs *Stofnunar Árna Magnússonar í íslenskum fræðum* (áður Orðabók Háskólans) sem annast gerð málheildarinnar.

Gert er ráð fyrir að í málheildinni verði í fyrstu um 25 milljón orð úr 900–1000 fjölbreyttum textum frá tímabilinu 2000–2006. Dæmin verða fengin úr bókum, blöðum, tímaritum og ýmiss konar smáprenti auk óútgefina texta og talmáls. Textarnir fjalla um ólík málefni og meðal þeirra verða bæði skáldverk og ýmiss konar umræðu-, fræðslu- og upplýsingaefni. Engin verk verða þó tekin upp í heilu lagi í málheildina. Síðar er stefnt að því að auka málheildina þannig að í framtíðinni verði í henni um 100 milljón orð.

Gerð og umsýsla *Markaðrar íslenskrar málheildar*

Í lokaskýrslu starfshóps um tungutækni á vegum menntamálaráðherra sem prentuð var vorið 1999 (sjá <http://www.tungutaekni.is/news/Skyrsla.pdf>) var gerð stórra rafræna málsafna með íslenskum textum talin meðal mikilvægustu forgangsverkefna til að stuðla að því að íslenska yrði gjaldgeng í upplýsingatækni nútímans og tryggja þar með stöðu málsins til framtíðar. Í kjölfarið lagði menntamálaráðuneytið fram fjár-

magn til ýmissa brýnna verkefna á sviði tungutækni, þ.á m. til að styrkja gerð *Markaðrar íslenskrar málheildar* (alls 18,5 milljónir króna). *Orðabók Háskólans* tók að sér að vinna verkið með samningi við ráðuneytið árið 2004. Haustið 2006 tók til starfa ný stofnun, *Stofnun Árna Magnússonar í íslenskum fræðum*, sem tók við eignum og skuldbindingum *Orðabókar Háskólans*, þ.á m. gerð málheildarinnar sem unnin er á orðfræðisviði hinnar nýju stofnunar. Stofnunin leggur til húsnaði, aðstöðu og sérfræðipækkingu og annast m.a. samninga við rétthafa textanna um afnot þeirra í tungutækniverkefnum og við rannsóknir.

Stofnun Árna Magnússonar í íslenskum fræðum hefur einnig verið falið að vista málheildina þegar hún verður tilbúin. Aðgangur verður veittur að *Markaðri íslenskrri málheild* í rafrænu formi (sbr. 2. lið hér að neðan). Allir notendur skrifa undir notkunarskilmála þar sem notkun málheildarinnar er skilgreind og takmörkuð. Ekki verður tekið gjald fyrir aðgang að málheildinni en heimilt verður að taka gjald sem nemur kostnaði við umsýslu og dreifingu gagnanna.

Leyfi frá rétthöfum

Til þess að málheildin komi að tilætluðum notum þarf hún að geyma sem fjölbreytilegasta texta. Það er því von aðstandenda verkefnisins að sem flestir gefi leyfi til þess að textar þeirra séu nýttir í þessu skyni. Leitað er eftir tvenns konar leyfi frá rétthöfum:

1. Leyfi til þess að varðveita heil verk í textasafni *Stofnunar Árna Magnússonar í íslenskum fræðum*. Textasafnið nýtist til rannsókna á íslensku máli og orðaforða og verður aðgengilegt til leitar á vefsetri stofnunarinnar. Í niðurstöðum slíkrar leitar birtast einungis stutt dæmi úr textunum, þ.e. leitarstrengurinn ásamt nánasta samhengi (allt að 500 bókstafir, 5–6 línur), auk bókfræðilegra upplýsinga um viðkomandi texta. Ekki verður unnt að nálgast lengri texta með leit í textasafni stofnunarinnar.
2. Leyfi til þess að nota hluta úr textum í *Markaða íslenska málheild*. Hámarkslengd texta í málheildinni verður 40.000 orð (120–140 bls. í dæmigerðri skáldsögu). Ef útgefið verk eftir nafngreindan höfund er styttra en 40.000 orð verður a.m.k. 20% af textanum sleppt. Aðgangur að málheildinni verður á vefsetri *Stofnunar Árna Magnússonar í íslenskum fræðum*. Boðið verður upp á opna leit sem skilar einungis stuttum dæmum úr textum (sbr. 1 hér að ofan). Auk þess verður hægt að sækja um notendanafn og aðgangsorð að málheildinni gegn því að samþykkja notkunarskilmála og gefur það rýmri aðgang og möguleika til sveigjanlegri leitar. Loks mun þeim sem hyggjast nota málheildina í umfangsmiklum rannsóknum eða í tungutæknilausnum gefast kostur á að fá hana afhenta í heild gegn greiðslu umsýslugjalds og með undirritun notkunarleyfis þar sem ítarlegir og strangir notkunarskilmálar eru tilgreindir.

Stofnun Árna Magnússonar í íslenskum fræðum mun ekki láta textana í hendur þriðja aðila nema eins og lýst hefur verið hér að framan og notendur málheildarinnar skuldbinda sig til þess að láta textana ekki í hendur þriðja aðila né heldur að birta textana eða brot úr þeim. Tilvitnanir eru þó heimilar í samræmi við höfundalög.

Haldin verður skrá yfir alla þá sem fá notendanafn og aðgangsorð að málheildinni á vefsetri svo og yfir nöfn og heimilisföng þeirra sem fá afrit af henni og hafa þar með skuldbundið sig til þess að fara eftir þeim skilmálum sem fram koma í notkunarleyfi. Rétthafar efnis (þ.e. höfundar textanna) geta hvenær sem er fengið aðgang að þessum skrám.

Helstu spurningar

Hér verður farið yfir nokkur atriði sem rétthafar kynnu að vilja spyrja um.

Hver hagnast á málheildinni og hvernig verður hún notuð?

Menntamálaráðuneytið styrkir verkið þar sem gert er ráð fyrir að það muni efla mjög fræðilegar og hagnýtar rannsóknir á íslensku máli. Markmiðið með gerð málheildarinnar er ekki beinn fjárhagslegur ávinningur en þó má segja að allir hagnist á því að verkið verði til. Málheildin verður sérstaklega gagnleg fyrir þá sem smíða margs konar hugbúnað og tól sem tengjast málnotkun. Má þar nefna þýðingarforrit og búnað sem fylgir ritvinnslukerfum fyrir leiðréttingu og ábendingar um stafsetningu og málfræði. Málheildin getur enn fremur verið góð stoð fyrir þá sem semja orðabækur, málfræðibækur og margs konar kennslufni um íslensku. Aðgangur að málheildinni gerir auðveldara að greina merkingu orða út frá textasamhengi, skoða hvernig einstök orð eru notuð, afmarka orðasambönd, fá málfræðilegar upplýsingar o.s.frv.

Er tryggt að rafrænar textar í málheildinni verði ekki misnotaðir?

Notkunarskilmálar tilgreina nákvæmlega hvernig nota má málheildina og textana sem í henni eru. Sérstaklega er lagt bann við því að veita þriðja aðila aðgang að málheildinni í heild eða að hluta. Aðstandendum málheildarinnar er annþing um að textar í málheildinni verði ekki misnotaðir. Textar sem teknir eru með í málheildina eru áfram varðir af ákvæðum höfundarréttarlaga.

Hverjir nota málheildina?

Þess má vænta að notendur málheildarinnar verði einkum einstaklingar, fyrirtæki og stofnanir sem vinna að orðabókargerð, margvíslegum tungutækniverkefnum og rannsóknum á íslensku nútímamáli. Opin leit getur einnig nýst öllum málnotendum, t.d. rithöfundum, þýðendum og skólafólki.

Verður greitt fyrir afnot af textum?

Verkefnið er ekki unnið í ábataskyni og ekki verður greitt fyrir afnot af textum. Hver einstakur texti verður aðeins lítið brot af allri málheildinni en þörf er á textum af margvíslegri gerð og um ólík efni til þess að málheildin endurspegli sem best hvernig málið er notað af ólíkum málnotendum og við mismunandi aðstæður.

Sýnishorn af broti úr Markaðri íslenskri málheild

Hver texti í málheildinni verður merktur með titli rits, nafni höfundar, útgáfuári, textategund, aldri og kyni höfundar, markhópi o.fl. upplýsingum sem nýtast til þess að flokka textana. Textarnir verða geymdir í rafrænu formi með sérstöku XML-sniði sem hefur verið skilgreint fyrir málheildir. Sýnt er dæmi um skráningu textabrots með tveimur setningum úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson. Fyrst er brot úr hausnum þar sem eru upplýsingar um textann, síðan koma orðin í textanum ásamt nefnimynd þeirra og greiningarstreng. Ekki er víst að þetta dæmi sýni endanlega mynd þess sniðs sem notað verður fyrir málheildina.

```
<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>
```

Fyrst eru allar sniðskipanir fjarlægðar og lítur þá textinn út þannig:

Ég stökk á eftir strætó og veifaði, vagnstjórinn sá mig og stoppaði.
Ég tautaði takk og brosti til hans um leið og ég lét miðann detta.

Síðan er textinn greindur vélrænt eftir orðflokkum og beygingu, nefnimyndir fundnar og textanum komið í sérstakt snið og lítur þá út þannig:

```
<s n="1">
<w type="fplen" lemma="ég">Ég</w> <w type="sfglep" lemma="stökkva">stökk</w>
<w type="aa" lemma="á">á</w> <w type="aþ" lemma="eftir">eftir</w>
<w type="nkep" lemma="strætó">strætó</w> <w type="c" lemma="og">og</w>
<w type="sfglep" lemma="veifa">veifaði</w> <c type=",">,</c>
<w type="nkeng" lemma="vagnstjóri">vagnstjórinn</w>
<w type="sfg3ep" lemma="sjá">sá</w> <w type="fpleo" lemma="ég">mig</w>
<w type="c" lemma="og">og</w>
<w type="sfg3ep" lemma="stoppa">stoppaði</w> <c type=".">.</c>
</s>
<s n="2">
<w type="fplen" lemma="ég">Ég</w> <w type="sfglep" lemma="tauta">tautaði</w>
<w type="au" lemma="takk">takk</w> <w type="c" lemma="og">og</w>
<w type="sfglep" lemma="brosta">brosti</w> <w type="ae" lemma="til">til</w>
<w type="fpkee" lemma="hann">hans</w> <w type="ao" lemma="um">um</w>
<w type="nveo" lemma="leið">leið</w> <w type="c" lemma="og">og</w>
<w type="fplen" lemma="ég">ég</w> <w type="sfglep" lemma="láta">lét</w>
<w type="nkeog" lemma="miði">miðann</w>
<w type="sng" lemma="detta">detta</w> <c type=".">.</c>
</s>
```

Til þess að nýta pláss eru oftast höfð tvö orð í línu í þessu dæmi. Í leitarkerfi málheildarinnar verður hægt að velja hvort textinn birtist í þessu sniði eða sem hreinn texti.

Táknið `<s n="1">` merkir að fyrsta setning í textanum byrji hér, táknið `</s>` lokar setningunni.

Runan `<w type="sfglep" lemma="stökkva">stökk</w>` sýnir orðmyndina *stökk*. Nefnimyndin er *stökkva* sem er nafnháttur sagnarinnar, táknaður með alþjóðlega orðinu *lemma* fyrir **nefnimynd**. Mark orðsins er `sfglep` þar sem **s** stendur fyrir sögn, **ε** stendur fyrir framsöguhátt, **1** stendur fyrir fyrstu persónu, **e** stendur fyrir eintölu og **þ** stendur fyrir þátíð. Í þeim staðli sem er notaður er orðið **type** haft um mörkin.