Sigrún Helgadóttir

Mörkuð íslensk málheild

Sigrún Helgadóttir

Mörkuð íslensk málheild

Inngangur

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998-1999 benti á að til þess að unnt sé að útbúa forrit sem nýta og nota tungumál þurfi að vera fyrir hendi miklar og nákvæmar upplýsingar um tungumálið og notkun þess. Þar má m.a. nefna upplýsingar um tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerð og merkingu.

Þessar upplýsingar má fá úr textaheild eða málheild (e. *corpus*), þ.e. safni tölvutækra texta af ýmsu tagi svo sem blaðatexta, fræðitexta af ýmsum sviðum, bókmenntatexta og talmáls. Starfshópurinn lagði því m.a. til að komið yrði á fót slíkri málheild, sem gæti nýst „fyrirtækjum sem hráefni í afurðir".

Víða í grannlöndum okkar eru stórar málheildir þegar til eða verið er að koma á fót slíkum söfnum sem eru aðgengileg fyrir margvíslegar rannsóknir á málinu og til afnota fyrir þá sem búa til ýmiss konar tungutæknitól. Þar má nefna *British National Corpus* (BNC)[1] í Bretlandi, *Korpus 2000*[2] í Danmörku og *American National Corpus* (ANC)[3] í Bandaríkjunum. Það fer eftir aðstæðum í hverju landi hvernig notkun og aðgangi er háttað.

Breska málheildin BNC er stærst þessara málheilda og komin lengst á veg. Í henni eru 100 milljón orð af breskri ensku. BNC-málheildin var búin til á fyrri hluta 10. áratugar tuttugustu aldar. Að því stóðu útgefendur og háskólastofnanir og fékk verkefnið umtalsverða opinbera styrki. Í málheildinni eru fjölbreyttir textar í tilteknum hlutföllum auk talaðs máls sem er um 10% af safninu.

Hvað er mörkuð málheild?

Með **markaðri málheild** (e. *tagged corpus*) er átt við safn fjölbreyttra textabúta sem hafa verið greindir á málfræðilegan hátt. Málheildin er í rafrænu formi og venjulega geymd í stöðluðu sniði. Hverjum textabút fylgja upplýsingar um textann sem búturinn er úr og hverri orðmynd fylgir **nefnimynd** (*lemma*) og greiningarstrengur, sem kallast **mark** (e. *tag*) og sýnir orðflokk og beyging-

[1] http://www.natcorp.ox.ac.uk/
[2] http://korpus.dsl.dk/korpus2000/
[3] http://americannationalcorpus.org/

armynd orðsins. Nefnimynd nafnorða og fornafna er nefnifall og nafnháttur er nefnimynd sagna. Taka má sem dæmi setningarbrotið *ég sagði*. Nefnimynd fornafnsins *ég* er *ég* og markið verður `fp1en`, þar sem `f` táknar fornafn, `p` táknar persónufornafn, `1` táknar fyrstu persónu, `e` táknar eintölu og `n` táknar nefnifall. Nefnimynd sagnarinnar *sagði* er *segja* og markið verður `sfg1eþ` þar sem `s` táknar sagnorð, `f` táknar framsöguhátt, `g` táknar germynd, `1` táknar fyrstu persónu, `e` táknar eintölu og `þ` táknar þátíð.

Hvernig eru málheildir notaðar?

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, orðasambönd, setningargerð og merkingu eins og þegar er getið. Málheildir gefa einnig upplýsingar um hvernig tiltekið tungumál er notað á tilteknum tíma. Þær gefa vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti.

Mynd 1

Mörkuð málheild er því undirstaða fyrir þróun þýðingarforrita og mikilvæg fyrir nútíma orðabókargerð. Margir útgefendur orðabóka byggja nú gerð orðabóka á stórum mörkuðum málheildum. Upplýsingar sem fást úr markaðri málheild má einnig nota við gerð ýmissa tungutæknitóla, t.d. fyrir talgreiningu og talgervingu. Einnig eru slíkar upplýsingar nauðsynlegar við þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði. Mörg tungutæknitól af þessu tagi nýtast sérstaklega fyrir blinda, heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika.

Mörkuð íslensk málheild

Stefnt er að því að setja saman á næstu þremur árum málheild með íslenskum textum sem hafa að geyma um 25.000.000 orð. Árið 2002 veitti menntamálaráðuneytið styrk til verkefnis sem fólst í því að gera tilraunir með búnað til að marka íslenskan texta á vélrænan hátt.[4] Í tilraununum var notað textasafn sem var gert vegna *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991). Í því safni eru um 500.000 orð og fylgir hverri orðmynd nefnimynd og mark og hefur greining orða í textasafninu verið leiðrétt handvirkt. Textasafn orðtíðnibókarinnar verður því notað sem fyrsti vísir að málheildinni. Stefnt er að því að safna auk þess efni úr 900-1.000 textabútum sem skiptast á tiltekinn hátt eftir uppruna og efni. Hámarksstærð hvers textabúts verður 40.000 orð en aldrei er tekinn heill texti. Ef texti er styttri en 40.000 orð er 10% af textanum sleppt.

Valdir verða textar úr ritum sem gefin hafa verið út frá árinu 2000. Stefnt er að því að um 60% textanna komi úr bókum, 25% úr blöðum og tímaritum, 5-10% verði úr öðru útgefnu efni, 5-10% verði óútgefið efni og minna en 5% verði efni sem er skrifað til upplestrar. Enn fremur er stefnt að því að um 25% af textunum séu skáldverk og um 75% verði nytjatexti sem skiptist milli texta um hagnýtt vísindi, náttúrufræði, þjóðfélagsfræði, heimsmál, viðskipti, listir, trúarbrögð, heimspeki og tómstundir.

Orð í textunum verða greind á vélrænan hátt og er stefnt að um 90% nákvæmni. Hverri orðmynd í málheildinni á að fylgja nefnimynd orðsins og mark. Stefnt er að því að mörkun um einnar milljónar lesmálsorða í málheildinni verði leiðrétt handvirkt, þ.e. um 500.000 orð til viðbótar þeim 500.000 orðum sem þegar hafa verið greind.

Málheildir eru venjulega skráðar með stöðluðu sniði til þess að tryggja að sem flestir sem nota ólíkar tölvur og hugbúnað geti nýtt efnið. Notuð verður XML-útgáfa af sniði fyrir málheildir sem TEI-samtökin (TEI: *Text Encoding Initiative*) hafa skilgreint. Í þessu sniði er gert ráð fyrir að hverjum textabút fylgi haus þar sem skráðar eru margvíslegar upplýsingar um textann, höfund hans o.fl.

[4] Sjá greinargerð um gerð markara fyrir íslenskan texta annars staðar í þessu hefti

Í mynd 1 er sýnt er dæmi um skráningu textabrots með þremur setningum úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson. Fremst er haus þar sem eru upplýsingar um textann, síðan koma orðin í textanum ásamt nefnimynd þeirra og marki. Ekki er víst að þetta dæmi sýni endanlega gerð þess sniðs sem notað verður fyrir málheildina.

Söfnun texta

Til þess að raunhæft sé að koma þessu í kring er nauðsynlegt að semja við rétthafa texta um að fá efni án þess að greitt sé fyrir það. Rétthöfum verður því gerð grein fyrir hvernig aðgangur verður veittur að málheildinni þegar hún verður komin í notkun.

Ráðgert er að nýta það tækifæri sem nú gefst til textaöflunar til þess að bæta einnig textasöfn Orðabókar Háskólans. Textum verður fyrst komið fyrir í textasafni Orðabókarinnar og síðan verða textabrot sótt þangað til nota í málheildinni.

Til þess að geta valið texta er ráðgert að fá lista úr bókasafnskerfinu Gegni yfir efni sem gefið var út árið 2000 og síðar. Beita þarf öðrum aðferðum til þess að finna óútgefið efni og efni sem er ætlað til upplestrar. Einungis verður valinn tölvutækur texti.

Talmáli verður ekki safnað sérstaklega en reynt verður að fá afnot af tölvutækum skrám sem þegar eru til og geyma umritað talað mál.

Mörkun og hjálparskrár

Eins og þegar hefur verið getið veitti menntamálaráðuneytið styrk árið 2002 til verkefnis sem fólst í tilraunum til að marka íslenskan texta á vélrænan hátt. Vinna við verkið hófst síðla árs 2002 og var lokið í upphafi árs 2004. Niðurstöður verkefnisins verða nýttar við mörkun texta í málheildinni.

Við mörkunina þarf einnig að nota ýmsar hjálparskrár og orðasöfn. Stærst þessara hjálparskráa er orðasafn sem gert hefur verið úr beygingarlýsingu íslensks nútímamáls. Beygingarlýsingin var einnig gerð fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins. Í beygingarlýsingunni eru allar beygingarmyndir um 170.000 íslenskra orða. Einnig hefur verið aflað skráa yfir mannanöfn, örnefni, heiti fyrirtækja og skammstafanir.

Hvernig verður verkið unnið?

Gert er ráð fyrir að í upphafi verksins verði lögð áhersla á að skilgreina verkþætti og afla forrita og annarra verkfæra eða búa þau til. Einnig verður unnið við að undirbúa öflun texta. Nokkur vinna mun felast í því að fá leyfi til þess að fá að nota textana í málheildinni. Á fyrsta ári verkefnisins verður lögð áhersla á að safna textum sem frjáls aðgangur er að.

Þegar búið er að ná í textana þarf að koma þeim í vinnsluhæft form, m.a. með því að

hreinsa úr þeim prentskipanir og ganga frá neðanmálsgreinum og myndatextum.

Einnig þarf að efnisflokka textana og skrá ýmiss konar upplýsingar um þá, t.d. uppruna, höfund, birtingu og tímasetningu. Þá tekur við ýmiss konar forvinnsla sem felst m.a. í því að merkja málsgreinaskil á ótvíræðan hátt, leysa úr skammstöfunum, taka ákvörðun um meðferð nafna, talna og dagsetninga.

Síðan verður beitt þeim aðferðum við mörkun sem skilgreindar voru í verkefninu sem greint var frá hér að ofan. Aðferðirnar voru þróaðar með því að láta tiltekin forrit læra greiningu af textasafni *Íslenskrar orðtíðnibókar*. Val texta í textasafn Orðtíðnibókarinnar takmarkaðist af því hvaða textar voru tiltækir í tölvutæku formi þegar textunum var safnað. Þess vegna vega bókmenntatextar þyngra en æskilegt væri. Stefnt er að því að leiðrétta handvirkt til viðbótar greiningu um 500.000 orða til þess að úthlutun marka verði nákvæmari. Stefnt er að því að hlutföll milli textategunda í því milljón orða safni þar sem mörkun hefur verið leiðrétt verði lík hlutföllum í málheildinni allri. Fyrst verður tekinn fyrir textabútur með um 100.000 orðum og orðin mörkuð með þeim aðferðum og gögnum sem þegar eru til. Mörkunin verður síðan leiðrétt handvirkt. Leiðrétti búturinn verður þá lagður við textasafn orðtíðnibókarinnar og forritin látin læra af stækkuðu textasafni. Þannig verður haldið áfram þangað til fyrir liggur safn með um 1.000.000 orðum þar sem mörkun hefur verið leiðrétt handvirkt. Sá lærdómur sem dreginn verður af þessu safni verður síðan notaður til að marka texta sem síðar verður bætt við málheildina.

Rekstur málheildar

Ekki hefur verið ákveðið hvar málheildin verður vistuð né hvernig veittur verður að henni aðgangur. Nauðsynlegt er að ákveða þetta svo gera megi rétthöfum texta grein fyrir þessum atriðum áður en þeir gefa leyfi til þess að textar þeirra verði notaðir.

Umsjón með verkinu

Verkefnið „Mörkuð íslensk málheild" er unnið af Orðabók Háskólans samkvæmt samningi við menntamálaráðuneytið frá 14. júní 2004. Verkinu á að skila í júní 2007. Orðabók Háskólans leggur m.a. til verksins aðstöðu og markað textasafn *Íslenskrar orðtíðnibókar*. Verkefnisstjóri er Sigrún Helgadóttir (sigrunh@lexis.hi.is). Skipuð hefur verið verkefnisstjórn sem vinnur með verkefnisstjóra. Í verkefnisstjórninni eiga sæti Ásta Svavarsdóttir, Eiríkur Rögnvaldsson og Kristín Bjarnadóttir.

Heimildir

Jörgen Pind (ritstj.), Friðrik Magnússon, Stefán Briem. 1991. *Íslensk orðtíðnibók.* Orðabók Háskólans, Reykjavík.