Mörkun texta og markaðar málheildir

Sigrún Helgadóttir
Stofnun Árna Magnússonar í íslenskum fræðum
sigruhel@hi.is

Útdráttur

Lengi hefur mörkun texta og gerð markaðra málheilda verið mikilvægur hluti máltækni. Í þessari grein verður sagt frá vélrænum aðferðum við mörkun og sérstaklega gerð grein fyrir tilraunum við vélræna mörkun íslensks texta. Einnig verður greint frá vinnu við að koma upp markaðri íslenskri málheild.

1 Inngangur

Í ýmsum máltækniverkefnum þar sem unnið er úr texta er ávinningur að því að orð í textanum séu greind í orðflokka og beygingarmyndir. Má þar nefna greiningu texta í setningahluta, orðtöku úr texta fyrir gerð orðasafns, upplýsingaheimt, talkennsl, talgervingu, vélrænar þýðingar, orðabókargerð, fyrirspurnarkerfi og leiðréttingarforrit. Einnig er nauðsynlegt að orð í texta séu greind eftir orðflokkum og beygingu ef gera á tíðnikönnun á texta eins og þá sem birt er í *Íslenskri orðtíðnibók* (Jörgen Pind *o.fl.*, 1991). Greiningin er venjulega sett fram sem greiningarstrengur, **mark** (e. *tag*), sem sýnir orðflokk og málfræðileg atriði eins og fall, tölu og kyn fallorða og persónu, tölu og tíð sagna.

Mörkuð málheild (e. *tagged corpus*) er safn fjölbreyttra texta sem eru geymdir í stöðluðu sniði í rafrænu formi. Hverri orðmynd fylgir greiningarstrengurinn, markið, og auk þess **nefnimynd** (e. *lemma*) sem er t.d. nefnifall í eintölu fyrir fallorð og nafnháttur sagna. Hverjum texta í málheildinni fylgja jafnframt bókfræðilegar upplýsingar um verkið sem textinn er úr.

Úr mörkuðum málheildum má lesa ýmiss konar gagnlegan fróðleik. Þar má nefna upplýsingar um tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerð, merkingu o.fl. Notendur málheilda eru einstaklingar, fyrirtæki og stofnanir sem vinna að margvíslegum máltækniverkefnum, rannsóknum á íslensku nútímamáli og orðabókargerð.

Í greininni verður fyrst greint frá aðferðum við vélræna mörkun (2) og sérstaklega við mörkun íslensks texta (2.2). Í 3 verður síðan greint frá gerð *Markaðrar íslenskrar málheildar (MÍM)*.

2 Málfræðileg mörkun texta

Með **mörkun** (e. *tagging*) er átt við það að merkja orð í samfelldum texta á kerfisbundinn hátt, t.d. með málfræðilegum upplýsingum, nefnimynd orðsins og upplýsingum um setningafræðilegt hlutverk. Í þessari grein er orðið mark notað um málfræðilegt mark[1]. Málfræðilegt mark er greiningarstrengur sem er tengdur orði í texta og segir til um orðflokk orðsins og önnur málfræðileg atriði, t.d. kyn, tölu og fall fallorða og persónu, tölu og tíð 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.

Mörkun er nauðsynlegt fyrsta skref í ýmsum máltækniverkefnum þar sem unnið er úr texta og þess vegna hefur verið lögð mikil áhersla á að þróa góðar aðferðir til þess að marka íslenskan texta.

2.1 Mörkunaraðferðir

Elsta aðferð við málfræðilega mörkun er handvirk greining texta eftir orðflokkum og beygingu. Sú aðferð er þó mjög tímafrek og þess vegna hefur lengi verið fengist við að þróa vélrænar aðferðir við málfræðilega mörkun. Þetta svið

[1] Í ensku eru notuð orðin *POS tag*, *part-of-speech tag* og *morphological tag* um það sem hér er kallað málfræðilegt mark. Þó að *POS* eða *part-of-speech* sé venjulega notað um orðflokk eru þessi orð oft einnig látin ná yfir beygingarlegar myndir.

hefur því fengið mikla umfjöllun á undanförnum áratugum hjá þeim sem vinna við máltækni.

Vélrænar aðferðir við mörkun eru venjulega flokkaðar í tvo flokka, **regluaðferðir** (e. *rule based methods*) og **gagnaaðferðir** (e. *data-driven methods*). Fyrstu vélrænu aðferðirnar sem voru þróaðar voru regluaðferðir. Orðasafn var notað til þess að merkja sérhvert orð í texta með öllum hugsanlegum greiningarstrengjum. Síðan voru notaðar málfræðilegar reglur til þess að skera úr um hvaða greiningarstrengur væri réttur. Þessar reglur voru byggðar á málfræði hvers tungumáls og venjulega samdar af málfræðingum. Forrit sem notuðu reglurnar voru háð því tungumáli sem reglurnar voru gerðar fyrir.

Gagnaaðferðir byggjast á því að nota textasafn sem hefur verið markað og mörkunin yfirfarin handvirkt þannig að hún sé eins rétt og kostur er. Forrit þar sem beitt er tiltekinni aðferð er látið læra af gögnunum. Útkoma úr þjálfun gagnamarkara er mörkunarlíkan fyrir tiltekið tungumál sem má nota við mörkun nýs texta. Markaða textasafninu er venjulega skipt í tvo ójafna hluta. Stærri hlutinn er notaður sem þjálfunarsafn fyrir forritið til þess að læra af og minni hlutinn er notaður sem prófunarsafn til þess að prófa líkanið sem búið var til eftir þjálfunarsafninu. Gagnaaðferðir eru óðháðar tungumáli.

2.2 Mörkun íslensks texta

Haustið 1998 skipaði menntamálaráðherra starfshóp til að gera könnun á því hvernig mætti efla tungutækni[2] hér á landi. Starfshópurinn skilaði lokaskýrslu (Rögnvaldur Ólafsson *o.fl.*, 1999) árið 1999. Í framhaldi af starfi vinnuhópsins var veitt opinbert fé til verkefna á sviði tungutækni undir merkjum tungutækniverkefnis menntamálaráðuneytisins.

Meðal fyrstu verkefna sem voru styrkt af tungutækniverkefni ráðuneytisins var gerð málfræðilegs markara fyrir íslensku og hófst verkið haustið 2002. Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni. Valið stóð á milli þess að búa til reglumarkara eða beita gagnaaðferðum á greint textasafn. Upplýsingar lágu fyrir um að gerð reglumarkara væri mjög tímafrek. Það tók t.d. um 7 mannár að búa til reglumarkara fyrir norsku sem byggðist á **hömlumálfræði** (e. *constraint grammar*) (Hagen *o.fl.*, 2000). Sú leið þótti því ekki fýsileg.

Orðabók Háskólans hafði yfir að ráða gagnasafni með um 590 þúsund lesmálsmálsorðum sem var notað við gerð *Íslenskrar orðtíðnibókar* (Jörgen Pind *o.fl.*, 1991). Í textasafninu eru textabrot sem valin höfðu verið úr 100 textum, 5000 orð úr hverjum texta. Allir textarnir voru gefnir út á árunum 1980-1989. Af þessum textum voru 20 úr frumsömdum íslenskum skáldverkum, 20 úr þýddum skáldverkum, 20 úr ævisögum, 20 úr nytjatextum (10 úr hugvísindum og 10 úr raunvísindum) og 20 úr barnabókum, þar af 10 frumsömdum á íslensku og 10 þýddum.

Þau skilyrði voru sett að aðeins einn texti væri notaður eftir hvern höfund eða þýðanda.

Markamengi með um 700 mörkum var skilgreint og hverju orði í textunum úthlutað marki. Textarnir voru markaðir með forriti (markara) sem notaði málfræðireglur og tíðnitölur (Stefán Briem, 1990) og síðan var mörkunin yfirfarin handvirkt. Nefnimynd fylgir einnig hverju orði í textasafninu.

Áður en hafist var handa við gerð málfræðilegs markara fyrir íslensku höfðu verið gerðar tilraunir með að þjálfa gagnamarkara á íslenskum textum veturinn 2001-2002 í tengslum við norræn námskeið í máltækni sem nokkrir Íslendingar sóttu. Notað var úrtak úr greindum textum orðtíðnibókarinnar, þ.e. barnabækurnar. Prófuð var svo kölluð leiðréttingaaðferð (TBL, *transformation based learning*) með því að beita forritinu *μ-TBL* (Lager, 1999) og Markovslíkan með tölfræðimarkaranum *TnT* (Brants, 2000) (Sigrún Helgadóttir, 2002a og 2002b; Eiríkur Rögnvaldsson o.fl. 2002). Miðað við þau gögn sem notuð voru fékkst betri niðurstaða með leiðréttingaaðferðinni.

Upphaflegar hugmyndir um að gera málfræðilegan markara fyrir íslensku beindust þess vegna að því að nota leiðréttingaaðferðina en sú aðferð byggist á því að finna heppileg „sniðmát“ (e. *templates*) sem skilgreina aðgerðir á grundvelli tiltekins umhverfis orðanna sem verið er að marka. Þegar verkefnið hófst var þó ákveðið að prófa nokkrar gagnaaðferðir.

Til þess að prófa ólíkar aðferðir við mörkun er gjarnan notuð aðferð sem byggist á því að hafa til umráða tíu pör af þjálfunar- og prófunarsöfnum þar sem hvert prófunarsafn er um tíundi hluti textasafnsins sem er til ráðstöfunar. Búin voru til 10 pör úr textaskrám *Orðtíðnibókarinnar* þannig að í hverri skrá væru textabútar úr öllum 100 textunum sem mynda textasafn bókarinnar. Hverjum texta í *Orðtíðnibókinni* var skipt upp í

[2] Þegar tungutækniverkefni menntamálaráðuneytisins var sett af stað var orðið *tungutækni* notað um það sem á ensku kallast *language technology* (eldra *language engineering*). Mörgum finnst orðið *máltækni* ná merkingunni betur og er það notað annars staðar í þessari grein.

tíu nokkurn veginn jafna hluta. Hver þessara tíu hluta myndar eitt prófunarsafn og samstætt þjálfunarsafn hefur að geyma hina hlutana níu í hvert sinn. Prófunarsöfnin skarast ekki en þjálfunarsöfnin hafa um 80% sameiginlega texta. Allir markarar voru prófaðir á öllum 10 pörum og fundin meðalnákvæmni (þessi aðferð er kölluð á ensku *ten-fold cross-validation*).

Prófaðar voru fjórar gagnaaðferðir og fimm forrit eða markarar sem unnt er að þjálfa á íslenskum texta og eru fáanleg án greiðslu (Sigrún Helgadóttir, 2007). Prófaðir voru tveir markarar sem nota tölfræðilegar aðferðir, TnT (Brants, 2000) sem byggist á Markovslíkani og *MXPOST* (Ratnaparkhi, 1996) sem byggist á svo kölluðu **hámarksóreiðulíkani** (e. *Maximum Entropy Model*). Prófaðir voru tveir markarar sem byggjast á leiðréttingaaðferð, µ-TBL (Lager, 1999) og *fnTBL* (Florian og Ngai, 2002). Einnig var prófaður einn markari, *MBT* (Daelemans *o.fl.*, 2003), sem byggist á minnistækni. Markarinn µ-TBL sem hafði áður verið prófaður á litlum útdrætti úr textasafni *Orðtíðnibókarinnar* og gefið viðunandi niðurstöðu virtist ekki ráða við allt textasafn *Orðtíðnibókarinnar*. Honum voru því ekki gerð frekari skil. Minnismarkarinn gaf einnig slæma útkomu fyrir íslensku. Besta niðurstaða náðist með TnT markaranum eða 90,4% nákvæmni í mörkun. Vert er að benda á að gerð er sú krafa að allir stafir í greiningarstreng séu réttir.

Til samanburðar má geta þess að náðst hefur 96–97% nákvæmni við mörkun ensks texta með gagnamarkaranum CLAWS og með því að nota markamengi með um 133 mörkum (Garside, 1987). Megyesi (2002) lýsir tilraunum til þess að marka sænskan texta þar sem notaður var Stockholm-Umeå Corpus til þess að þjálfa nokkra gagnamarkara. Markamengið sem var notað hefur 153 mörk. TnT markarinn náði 93,55% nákvæmni í mörkun þegar hann var þjálfaður á þeirri málheild. Ekki er líklegt að unnt verði að ná svipaðri nákvæmni við mörkun íslensku eins og við mörkun ensku þar sem markamengi íslenskunnar er mun stærra en markamengi ensku. En síðar í greininni verður sýnt fram á að unnt er að ná sambærilegri nákvæmni í mörkun íslensku og sænsku, t.d. með því að minnka markamengið.

Gerðar voru ýmsar tilraunir til þess að bæta mörkunina. Helsta vandamál við mörkun er hvernig á að fara með óþekkt orð, þ.e. orð sem ekki koma fyrir í því textasafni sem viðkomandi markari hefur lært af. Meðalhlutfall óþekktra orða í pörunum tíu var 6,8%. Ein leið til þess að bæta niðurstöður mörkunar er að nýta viðbótarorðasafn og lækka þannig hlutfall óþekktra orða. Önnur leið er að sameina eða samþætta niðurstöður tveggja eða fleiri markara, t.d. með því að kjósa á milli þeirra eftir tilteknum reglum. Niðurstaða tilraunanna varð sú að gerð var tillaga um að nota þrjá markara, TnT, MXPOST og fnTBL, kjósa milli niðurstaðna þeirra og nota viðbótarorðasafn (Sigrún Helgadóttir, 2007). Einnig var bent á að markaskrá orðtíðnibókarinnar væri mjög stór og að sú greining sem þar er notuð væri ekki endilega sú eina rétta. Nákvæmni í mörkun eykst ef slakað er á kröfu um ítarlega greiningu. Sumar máltæknilausnir geta nýtt sér greiningu sem er ekki jafn ítarleg og sú greining sem kemur fram í markamengi *Orðtíðnibókarinnar*. Ef aðeins er gerð krafa um að fyrsti stafur í greiningarstreng sé réttur (orðflokkur) nær TnT markarinn 98,16% nákvæmni.

Til þess að komast hjá ýmsum vandamálum tengdum notkun gagnamarkara var búinn til reglumarkarinn *IceTagger* (Hrafn Loftsson, 2008). Markarinn byggist á aðferðum við að skera úr um hvaða greiningarstrengur er réttur ef fleiri en einn kostur kemur til greina. Mikilvægur hluti af IceTagger er giskarinn *IceMorphy* (Hrafn Loftsson, 2008) sem giskar á möguleg mörk fyrir „óþekkt orð“, þ.e. orð sem koma ekki fyrir í þjálfunarsafninu. IceMorphy getur einnig fyllt upp í eyður í markasafni tiltekinna orða ef öll möguleg mörk tiltekins orðs hafa ekki komið fyrir í þjálfunarsafni. Textasafn *Orðtíðnibókarinnar* var notað við þjálfun og prófun IceTagger og náðist um 91.6% nákvæmni. Með því að hækka nákvæmni í mörkun úr 90,4% í 91,6% fækkar villum í mörkun um 12,5%. Það er því eftir miklu að slægjast við að hækka nákvæmni í mörkun þó ekki sé nema um eitt prósentustig.

Gerð hefur verið tilraun til þess að búa til samþættan markara með því að láta Markovsmarkara einræða mörk orða sem IceTagger réði ekki við. Í því skyni var búinn til markarinn *TriTagger* sem hefur sömu virkni og TnT. Með þessari aðferð (Ice+HMM) fékkst um 91,8% nákvæmni (Hrafn Loftsson, 2006).

Hrafn Loftsson (2006) gerði einnig tilraunir til að sameina niðurstöður markara með kosningu.

Drezde og Wallenberg (2008) notuðu tvíátta flokkunaraðferð og náðu um 92,1% nákvæmni með því að nota textasafn *Orðtíðnibókarinnar*. Þeir notfærðu sér að gagnamarkarar geta fundið orðflokk orða með nokkuð mikilli nákvæmni. Þeir skiptu mörkuninni því í tvö stig. Á fyrra stigi var búinn til orðflokkamarkari (WC) og orð greind í orðflokka og á síðara stigi eru aðeins

skoðuð mörk sem heyra til orðflokki tiltekins orðs. Með þessu móti fækkar þeim mörkum sem þarf að skoða í hverju skrefi tvíátta mörkunaralgrímsins. Flestar villur í mörkun eru villur í greiningu falls fallorða. Drezde og Wallenberg bjuggu þess vegna til sérstakan markara sem úthlutar falli fyrir nafnorð, lýsingarorð og fornöfn (CT). Sameinaður markari þeirra (BI+WC+CT) náði 92,06% nákvæmni.

Hrafn Loftsson og fl. (2009) greina frá nýrri aðferð við mörkun sem byggist á því að samþætta reglumarkarann IceTagger og Markovsmarkarann TriTagger. Hugmyndin byggist á þeirri athugun að Markovsmarkari eins og TnT nær um 98,16% nákvæmni við að marka eftir orðflokkum en IceTagger nær aðeins 97,61% sambærilegri nákvæmni. Hugmyndin um að marka fyrst eftir orðflokkum er því fengin að láni frá Drezde og Wallenberg. Hugmyndin er að forvinna textann með TriTagger og nota síðan IceTagger til þess að úthluta marki miðað við þann orðflokk sem hefur verið valinn en þó með því að nota allan greiningarstrenginn. Með þessum nýja markara (HMM+Ice) náðist 92,19% nákvæmni.

Einnig var gerð tilraun til þess að sameina aðferðirnar Ice+HMM og HMM+Ice. Fyrstu skrefin eru eins og fyrir HMM+Ice en í seinasta skrefinu er markarinn Ice+HMM notaður í stað þess að nota aðeins IceTagger. Þessi aðferð er kölluð HMM+Ice+HMM og gefur 92,31% nákvæmni.

Markari	Óþekkt orð	Þekkt orð	Öll orð
TnT	71,82	91,82	91,45
IceTagger	75,30	92,78	91,59
Ice+HMM	75,63	93,01	91,83
BI+WC+CT	69,74	93,70	92,06
HMM+Ice	76,10	93,36	92,19
HMM+Ice+HMM	76,04	93,49	92,31

Tafla 1: Meðalmörkunarnákvæmni (%) með því að nota textasafn orðtíðnibókar

Í sömu grein er einnig greint frá tilraun til þess að nota gögn úr gagnasafni *Beygingarlýsingar íslensks nútímamáls* (Kristín Bjarnadóttir, 2005) til þess að lækka hlutfall óþekktra orða. Með því að nota þessi gögn lækkar hlutfall óþekktra orða úr 6,8% í 1,2% og nákvæmni HMM+Ice markarans hækkar í 92,99%. Líklegt er að hækka megi nákvæmni enn þá meir þegar tíðniupplýsingum hefur verið bætt við gagnasafn Beygingarlýsingarinnar.

Í fyrrnefndri grein (Hrafn Loftsson *o.fl.*, 2009) er einnig greint frá tilraunum til þess að minnka það markamengi sem var skilgreint fyrir *Íslenska orðtíðnibók* og mest hefur verið notað við mörkun íslensks texta. Tvenns konar aðferðir koma til greina þegar markamengi er minnkað. Í fyrsta lagi má breyta mörkum í þjálfunarsafninu og þjálfa markarana á nýjum mörkum. Í örðu lagi má halda markamenginu óbreyttu fyrir þjálfun en varpa mörkum í ný mörk áður en nákvæmni er metin. Seinni aðferðin var notuð í því verki sem lýst er í greininni. Þegar mörkum var fækkað var haft í huga að taka stafi úr greiningarstreng sem lýstu aðgreiningu sem erfitt er fyrir markarana að gera. Gerðar voru tilraunir með einföldun marka með því greina ekki sérnöfn, greina ekki milli mismunandi fornafna, greina ekki fallstjórn forsetninga og gera ekki greinarmun á samtengingum, nafnháttarmerki og tilvísunartengingum. Með þessu móti fækkar mörkum í um 450 og með HMM+Ice+HMM markarum fékkst þá 93,63% nákvæmni.

Í töflu 1 eru sýndar niðurstöður fyrir þá markara sem fjallað hefur verið um.

Líklegt er að áfram verði unnið við að bæta mörkun íslensks texta. Bent hefur verið á eftirfarandi atriði sem mætti huga að (Hrafn Loftsson *o.fl.*, 2009):

- Lækka mætti hlutfall óþekktra orða með því að nota stórt orðasafn sem yrði unnið úr gagnasafni Beygingarlýsingar íslensks nútímamáls þar sem tíðniupplýsingum hefur verið bætt við.
- Búa mætti til markara (og þáttara) sem byggist á hömlumálfræði.
- Gera mætti frekari tilraunir til þess að samþætta regluaðferðir og gagnaaðferðir.
- Finna mætti betri leiðir til þess að greina fallstjórn sagna.

Það er enn nokkuð langt í land að ná 97–98% greiningu en lengra verður tæplega komist. Eftir það fer málfræðinga að greina á um hvað er rétt greining (Eiríkur Rögnvaldsson *o.fl.*, 2002).

3 Mörkuð íslensk málheild

Gögn sem eru unnin upp úr mörkuðum málheildum eru m.a. nýtt við gerð tíðniorðalista, orðabókargerð, gerð leiðréttingarforrita, þýðingarforrita, búnaðar fyrir talgreiningu og talgervingu og gerð hjálparforrita fyrir blinda, heyrnarskerta, hreyfihamlaða og þá sem glíma við skriftar- og lestarörðugleika.

Í lokaskýrslu starfshóps menntamálaráðherra um hvernig mætti efla tungutækni hér á landi (Rögnvaldur Ólafsson *o.fl*, 1999) segir m.a.:

> Öll hagnýting tungutækni við meðferð ritaðs máls er mjög háð tilvist tvenns konar gagnasafna sem byggja þarf upp á skipulegan hátt fyrir hvert tungumál. Annars vegar þarf að koma upp textaheild (corpus) málsins, og hins vegar orðasafni (lexicon) þess.

Gerð íslenskrar málheildar (eða textaheildar) var því forgangsverkefni að mati starfshópsins. Slík söfn eru til á nokkrum tungumálum. Þar má nefna *British National Corpus* (BNC)[3] í Bretlandi og *Korpus 2000*[4] í Danmörku .

Þegar vinnu lauk við markaraverkefnið vorið 2004 var hafist handa við gerð markaðrar íslenskrar málheildar og var verkið styrkt af tungutækniverkefni menntamálaráðuneytisins. Verkið var unnið samkvæmt samningi milli menntamálaráðuneytisins og Orðabókar Háskólans en verður lokið undir merkjum Stofnunar Árna Magnússonar í íslenskum fræðum. Meginmarkmið verkefnisins er að bæta forsendur fyrir þróun íslenskrar máltækni. Stefnt var að því að í málheildinni yrðu um 25 milljónir lesmálsorða úr margvíslegum textum sem hafa verið gefnir út frá árinu 2000.

Formlegri efnisöflun lauk í byrjun október 2009. Þá voru í málheildinni tæplega 26 milljónir lesmálsorða.

Málheildin á að endurspegla íslenskt samtímamál. Textavalið miðaðist við það að málheildin gefi sem raunsannasta mynd af málinu og sýni sem best fjölbreytni í málnotkun, t.d. eftir uppruna textanna og viðfangsefnum.

Textarnir voru fengnir frá útgefendum, fyrirtækjum, stofnunum og einstaklingum í rafrænu formi. Margir textar fengust beint af veraldarvefnum. Rík áherslu var lögð á gott samstarf við rétthafa og útgefendur. Gert var sérstakt samkomulag við Rithöfundasamband Íslands, Hagþenki og Félag íslenskra bókaútgefenda en öll þessi samtök mæltu með því við félagsmenn sína að þeir tækju upp samstarf við stofnunina um afhendingu texta fyrir málheildina. Aflað var formlegs samþykkis frá rétthöfum allra texta sem safnað var fyrir málheildina. Verkefnið er ekki unnið í ábataskyni og ekki var greitt fyrir afnot af textum. Hver einstakur texti verður aðeins lítið brot af allri málheildinni en þörf er á textum af margvíslegri gerð er taka til margvíslegra viðfangsefna til þess að málheildin endurspegli sem best hvernig málið er notað af ólíkum málnotendum og við mismunandi aðstæður. Val texta ræðst þó alltaf af því hversu auðvelt er að nálgast tölvutæka gerð textans. Aðeins var sóst eftir tölvutækum textum og engir textar voru skráðir eða skannaðir sérstaklega vegna verkefnisins.

Ekki eru til leiðbeiningar um samsetningu texta í verki eins og því sem hér er fjallað um. Helst var stuðst við samsetningu texta í bresku málheildinni BNC. Sú málheild var gerð í upphafi 10. áratugar síðustu aldar í Bretlandi. Um 90% af þeirri málheild er ritað mál og um 10% talmál. Mjög dýrt er að safna talmáli. Þess vegna var horfið frá því að safna talmáli fyrir íslensku málheildina. Þó tókst að afla nokkurs talmálsefnis sem safnað hefur verið vegna annarra verkefna og er það nú ríflega 2% af textum málheildarinnar.

Í töflu 2 er sýnd sundurliðun á textum málheildarinnar eftir textaflokkum (uppruna textanna).

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

Tafla 2: Sundurliðun á textum eftir textaflokkum

Stór hluti textanna er úr útgefnum bókum og blöðum. Einnig var aflað efnis af heimasíðum stofnana og fyrirtækja, af bloggsíðum einstaklinga og úr tölvupósti. Auk þess er í málheildinni nokkuð af óútgefnu efni, þ. á m. nemendaritgerðir og textar sem ætlaðir eru til upplestrar eins og stólræður presta og útvarpspistlar.

Í málheildinni eru einnig beinar umritanir eftir töluðu máli, t.d. eðlilegum samtölum. Þetta eru

[3] http://www.natcorp.ox.ac.uk/
[4] http://korpus.dsl.dk/korpus2000/indgang_til_korpusdk.php

tæplega 600 þús. orð sem safnað hefur verið í þremur verkefnum:

- ÍSTAL - *Íslenskur talmálsbanki*: Sjálfsprottin samtöl, hljóðrituð 2000; u.þ.b. 20 klst.
- *MIN - Moderne importord i sprogene i Norden*: Hópviðtöl, hljóðritað 2002; u.þ.b. 10 klst.
- *Hvernig tala ungir Íslendingar í upphafi 21. aldar?* Samtöl (stýrð) ungmenna við jafnaldra og eldra fólk, hljóðrituð 2006; u.þ.b. 4 klst.
- *Tilbrigði í setningagerð*: Óundirbúnar ræður á Alþingi, hljóðritaðar 2004-05; u.þ.b. 20 klst.

Einnig fengust frá tölvudeild Alþingis textar af ræðum sem höfðu verið fluttar á Alþingi á árunum 2000-2009 (1% af textum málheildarinnar).

Í bresku málheildinni eru um 60% af ritmálstextum tekin úr bókum og um 25% úr dagblöðum og tímaritum. Þá voru aðstæður nokkuð aðrar en á Íslandi í upphafi 21. aldar. Munar þar mestu um tilkomu veraldarvefjarins. Prentmarkaður hér er einnig mun minni en í Bretlandi og því er erfiðara að afla efnis úr bókum. Stærsti flokkurinn í íslensku málheildinni eru textar úr dagblöðum, bæði prent- og vefmiðlum eða um 7.508 þús. orð (um 28% af málheildinni). Tiltölulega auðvelt er að afla þessara texta. Aðeins þurfti að hafa samband við einn aðila til þess að fá mikið af texta. Næsti flokkur er prentaðar bækur, um 5.770 þús. orð eða 22% af textum málheildarinnar. Sá flokkur reyndist langerfiðastur viðureignar. Hafa þurfti samband við bæði rétthafa og útgefendur og einnig reyndist mikil vinna fólgin í því að gera textana nothæfan fyrir málheildina. Í sumum tilvikum hafði t.d. tekist að afla leyfis rétthafa en útgefandi hafði textann ekki tiltækan í tölvutæku formi.

Mikil vinna var lögð í að nálgast efni úr tímaritum. Í sumum tilvikum tókst að fá efni úr mörgum tímaritum frá tilteknum útgefanda (*Heimur, Birtíngur*) sem gat veitt leyfi til notkunar textanna. Í öðrum tilvikum þurfti að sækja leyfi til hvers einstaks höfundar (*Læknablaðið, Náttúrufræðingurinn* o.fl.). Textar úr dagblöðum og tímaritum, hvort sem um er að ræða prentað efni eða vefefni, eru samanlagt um 38% af öllum textum í málheildinni. Um 20% af textum málheildarinnar var safnað beint af veraldarvef, t.d. bloggi, textum af vefsetrum fyrirtækja, félagasamtaka og stofnana og vefmiðlum.

Textarnir verða geymdir í rafrænu formi með sérstöku sniði sem TEI-samtökin[5] (TEI: Text Encoding Initiative) hafa skilgreint fyrir málheildir. Þar eru skráðar bókfræðilegar upplýsingar um hvern texta og mark og nefnimynd hvers orðs. Textarnir verða flokkaðir eftir uppruna og viðfangsefni.

3.1 Greining textanna

Sumarið 2009 var unnið að verkefninu *Mörkun og leiðrétting nýrrar málheildar* með styrk frá Nýsköpunarsjóði námsmanna. Markmið verkefnisins var að búa til málheild með um einni milljón orða þar sem mörkun og lemmun (finna nefnimyndir) hefur verið leiðrétt að hluta til. Textarnir voru valdir úr textasafni málheildarinnar. Í verkefninu voru aðferðir við tilreiðslu (skiptingu texta í setningar og orð), mörkun og lemmun yfirfarnar og lagaðar. Einnig var beitt aðferðum við leit að villum sem Hrafn Loftsson (2009) hefur þróað. Stefnt er að því að þessi nýja málheild leysi textasafn *Orðtíðnibókarinnar* af hólmi sem málheild fyrir þróun markara og annarra máltæknitóla þar sem textar í henni eru mun fjölbreyttari en textar í textasafni *Orðtíðnibókarinnar*.

Við úrvinnslu texta málheildarinnar verður notað forrit sem var þróað fyrir þetta verkefni og sér um **tilreiðslu** (e. *tokenization*) þar sem texta er skipt í setningar og lesmálsorð, mörkun og lemmun. Mörkunarhluti forritisins markar með fimm mörkurum ((IceTagger (Hrafn Loftsson, 2008), TnT (Brants, 2000), MXPOST (Ratnaparkhi, 1996), Bidir (Drezde og Wallenberg, 2008) og fnTBL (Florian og Ngai, 2002)). Síðan er forritið *CombiTagger* (Verena Henrich *o.fl.*, 2009) látið velja það mark sem álitlegast þykir samkvæmt tilteknum skilyrðum. Aðferð þessi byggist á tilraunum sem Hrafn Loftsson (2006, 2009) hefur gert. Anton K. Ingason og fl. (2008) hafa lýst þeirri einingu forritsins sem lemmar texta.

3.2 Aðgangur að málheildinni

Ekki verður seldur aðgangur að málheildinni eins og þegar hefur verið getið. Samkvæmt samningi við rétthafa verður allt efni sem er safnað fyrir málheildina líka notað til þess að styrkja Textasafn Stofnunar Árna Magnússonar í íslenskum fræðum[6]. Texti heilla bóka getur farið í textasafnið en sníða verður a.m.k. 20% af bókum fyrir

[5] http://www.tei-c.org/Guidelines/

[6] http://www.arnastofnun.is/page/arnastofnun_gagnasafn_textasafn

málheildina vegna samninga við rétthafa og útgefendur. Þetta er gert til þess að tryggja enn frekar að ekki verði unnt að endurgera texta úr málheildinni.

Textar sem safnað er fyrir málheildina eru ekki markaðir fyrir textasafnið. Aðgangur að textasafninu er á vefsetri Stofnunar Árna Magnússonar og eru niðurstöður leitar sýndar sem orðstöðulyklar. *Beygingarlýsing íslensks nútímamáls* er notuð til þess að veita aðgang að öllum orðmyndum tiltekins orðs. Birtir eru 140 stafir af textanum. Einnig má sjá úr hvaða verki tiltekin lína er fengin. Þegar er nokkuð af efni málheildarinnar aðgengilegt í textasafninu eins og bloggtextar, textar úr Morgunblaðinu, textar af vísindavefnum og textar af tölvupóstlistum. Fleiri textar munu bætast í textasafnið á næstu vikum.

Gert er ráð fyrir að aðgangur að mörkuðum textum málheildarinnar verði tvenns konar: Í fyrsta lagi verði opin leit á vefsetri Stofnunar Árna Magnússonar í íslenskum fræðum[7] með sérstöku leitarforriti þar sem mörk og nefnimyndir nýtast til þess að gera leitina markvissari. Einnig er gert ráð fyrir að afmarka megi leitina við tiltekna textaflokka. Stefnt er að því að fá að nota leitarforrit fyrir málheildir sem nú er í smíðum á vegum *Tekstlaboratoriet*[8] við Háskólann í Osló.

Í öðru lagi geta þeir sem vilja fá að nota málheildina í eigin tölvukerfi gert um það sérstakan samning, undirritað notkunarleyfi og greitt umsýslugjald. Engin forrit fylgja. Málheildinni verður dreift í TEI-sniði. Í upphafi var gert ráð fyrir að málheildinni yrði dreift á geisladiskum en ekki er ólíklegt að hún verði gerð aðgengileg á vefsetri stofnunarinnar og menn geti sótt hana þangað og flutt í eigin tölvur gegn því að samþykkja notkunarskilmálana á einhvern hátt. Nánari útfærsla á þessu bíður síðari tíma.

4 Lokaorð

Greint hefur verið frá rannsóknarvinnu við mörkun íslensks texta. Unnið verður frekar við það svið á næstunni til þess að ná meiri nákvæmni í mörkun.

Einnig var sagt frá gerð *Markaðrar íslenskrar málheildar* og stöðu þess verkefnis. Textar úr málheildinni hafa þegar verið notaðir við nokkur verkefni, t.d. við gerð *Merkingarbrunns fyrir íslenska máltækni* (Anna B. Nikulásdóttir, 2009) sem er hluti af rannsóknarverkefninu *Hagkvæm máltækni utan ensku*, sem hlaut öndvegisstyrk RANNÍS til þriggja ára í upphafi árs 2009.

Aðgangur að málheildinni á vefsetri Stofnunar Árna Magnússonar í íslenskum fræðum mun gjörbreyta aðstöðu til rannsókna á íslensku máli og nýtast til kennslu í íslensku í menntaskólum og háskólum. Markaðir textar verða mikilvægur efniviður í umfangsmiklar rannsóknir og gerð máltæknibúnaðar, t.d. leiðréttingaforrita.

Heimildir

Anna B. Nikulásdóttir. 2009. Merkingarbrunnur fyrir íslenska máltækni. Í xxxxxxxxxxxx. Reykjavík.

Anton K. Ingason, Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In B. Nordström and A. Ranta (eds.), *Advances in Natural Language Processing, 6th International Conference on NLP, GoTAL 2008, Proceedings*. Gothenburg, Sweden. 2008.

Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, bls. 224–231. Seattle, Washington, USA.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. MBT: *Memory-Based Tagger, Reference Guide. ILK Technical Report 03-13*, http://ilk.uvt.nl/downloads/pub/papers/ilk.0313.pdf

M. Drezde og J. Wallenberg. 2008. Icelandic Data Driven Part of Speech Tagging. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, OH, USA.

Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir. 2002. Vélræn málfræðigreining með námfúsum markara. *Orð og tunga* 6:1–9.

Florian, Radu and Grace Ngai. 2002. Fast Transformation-Based Learning Toolkit. http://nlp.cs.jhu.edu/~rflorian/fntbl/tbl-toolkit/tbl-toolkit.html

R. Garside. 1987. The CLAWS word-tegging system. In R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

K. Hagen, J. Johannessen, and A. Nøklestad. 2000. A Constraint-Based tagger for Norwegian. In C.-E. Lindberg and S.-N. Lund, editors, *17th Scandinavian Conference of Linguistics. Odense working Papers in Language and Communication*, volume 19, pages 31–48. Odense, Denmark.

[7] http://www.arnastofnun.is/

[8] http://www.hf.uio.no/tekstlab/

Van Halteren, Hans, Jakub Zavrel and Walter Daelemans. 2001. Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems. *Computational Linguistics* 27 (2), bls. 199–230.

Verena Henrich, Timo Reuter and Hrafn Loftsson. 2009. CombiTagger: A System for Developing Combined Taggers. In *Proceedings of the 22nd International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*. Sanibel Island, Florida, USA.

Beata Megyesi. 2002. *Data-Driven Syntactic Analysis – Methods and Applications for Swedish*. Ph.D.Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Hrafn Loftsson. 2006. Tagging Icelandic text: an experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, 40(2): 175–181.

Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

Hrafn Loftsson, Ida Kramarczyk, Sigrún Helgadóttir and Eiríkur Rögnvaldsson. 2009. Improving the PoS tagging accuracy of Icelandic text. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA-2009)*. Odense, Denmark.

Hrafn Loftsson. 2009. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece.

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

Kristín Bjarnadóttir. 2005. Modern Icelandic Inflections. In H. Holmboe, editor, *Nordisk Sprogteknologi 2005*, Museum Tusculanums Forlag, Copenhagen.

Lager, Torbjörn (1999) The µ-TBL System: Logic Programming Tools for Transformation-Based Learning. In *Proceedings of the Third International Workshop on Computational Natural Language Learning (CoNLL'99)*, Bergen, 1999.

Ratnaparkhi, A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Mehods in Natural Lanugage Processing (EMNLP-96)*, bls. 133–143. Philadelphia. PA.

Rögnvaldur Ólafsson, Þorgeir Sigurðsson, Eiríkur Rögnvaldsson. 1999. *Tungutækni*. Skýrsla starfshóps. Menntamálaráðuneytið.

Sigrún Helgadóttir. 2002a. The Icelandic µTBL Experiment: Learning rules from four different training corpora by using the µ-TBL System – Further developments. Term paper in NLP 1, GSLT. Óprentuð námsritgerð.

Sigrún Helgadóttir. 2002b. Statistical Tagger for Icelandic (TnT). Term paper in Statistical Methods 1, NGSLT. Óprentuð námsritgerð

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga*, 9:75–107.

Stefán Briem. 1990. Automatisk morfologisk analyse af islandsk tekst. Jörgen Pind og Eiríkur Rögnvaldsson (ritstj.). *Papers from the Seventh Scandinavian Conference of Computational Linguistics*. 1989:3–13. Institute of Lexicography, Institute of Linguistics, Reykjavík.