Sigrún Helgadóttir

Mörkun íslensks texta

1 Inngangur

Í ýmsum tungutækniverkefnum[1] 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, Friðrik Magnússon og Stefán Briem 1991).

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998–1999 (Rögnvaldur Ólafsson, Þorgeir Sigurðsson og Eiríkur Rögnvaldsson 1999) lagði m.a. til að „unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði“. Í anda tillögunnar var gerð málfræðilegs markara fyrir íslensku eitt af þeim verkefnum sem var styrkt af tungutækniverkefni menntamálaráðuneytisins[2] í apríl 2002. Markmið verkefnisins var að finna aðferðir til þess að

[1] Orðið *tungutækni* er hér notað um það sem á ensku nefnist venjulega **language engineering**. Einnig má nota orðið *máltækni*.

[2] Verktakar við verkið voru Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir) og Orðabók Háskólans. Verkefnisstjóri var Eiríkur Rögnvaldsson en Sigrún Helgadóttir mótaði

Orð og tunga 9 (2007), 75–107.

greina íslenskan texta vélrænt í orðflokka og eftir beygingu. Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni.

Verkefnið þróaðist á þann veg að prófaðar voru fjórar aðferðir við mörkun íslensks texta. Í greininni verður gerð grein fyrir málfræðilegri mörkun og nokkrum aðferðum við vélræna greiningu. Greint verður frá tilraun til þess að nota fjórar aðferðir við vélræna mörkun íslensks texta. Við tilraunirnar var notað textasafn sem var búið til vegna *Íslenskrar orðtíðnibókar*. Einnig verður greint frá tilraunum til þess að bæta mörkun, m.a. tilraunum til þess að sameina niðurstöður þriggja markara eftir tilteknum reglum til þess að ná sem bestum árangri við mörkun. Að lokum er greint frá tilraunum til þess að marka texta sem eru ekki hluti af textasafni Orðtíðnibókarinnar. Lokaskýrslu var skilað til menntamálaráðuneytisins í febrúar 2004.

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[3]. 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.

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 mál-

vinnulag við prófun markaranna og vann meginhluta vinnunnar ásamt félögum í Málgreiningarhópnum.

[3] 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.

fræðilega mörkun. Þetta svið 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 var beitt voru regluaðferðir. Orðasafn var notað til þess að merkja sérhvert orð í texta með öllum hugsanlegum greiningarstrengjum. Síðan voru notaðar 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 allar á því að nota textasafn sem hefur verið markað og mörkunin yfirfarin handvirkt þannig að hún sé eins rétt og kostur er. Forrit er síðan látið læra af gögnunum á tiltekinn hátt. Í þeirri vinnu sem hér er greint frá voru gerðar tilraunir með þrjár mismunandi gagnaaðferðir: tölfræðilegar aðferðir, aðferð sem mætti kalla leiðréttingaaðferð (e. *transformation-based learning*) og minnisaðferð (e. *memory-based method*). Forrit eða kerfi sem nota fyrir fram greint textasafn til þess að læra af mætti kalla námfúsa markara. Í greininni er sagt frá tilraun til þess að láta fimm mismunandi námfúsa markara læra að marka íslenskan texta. Tveir markaranna nota tölfræðilegar aðferðir, tveir nota leiðréttingaaðferð og einn notar minnisaðferð. Í 4. kafla er gerð grein fyrir þessum aðferðum og forritum.

3 Efniviður

Í þeirri vinnu sem hér er lýst var notað textasafn sem var gert fyrir vinnslu *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991) sem Orðabók Háskólans gaf út 1991. Vinna við undirbúning textasafnsins hófst 1985 og er safninu lýst nákvæmlega í formála Orðtíðnibókarinnar. Í textasafninu eru brot úr 100 textum sem voru gefnir út á tímabilinu 1980–1989, hvert með um 5.000 lesmálsorðum. Textarnir voru valdir úr 5 textaflokkum: íslenskum skáldverkum (20 textar), þýddum skáldverkum (20 textar), ævisögum og minningum (20 textar), fræðslutextum (10 á sviði hugvísinda, 10 á sviði raunvísinda) og barna og unglingabókum (10 frumsamdir textar, 10 þýddir textar).

Í formála Orðtíðnibókarinnar er *lesmálsorð* skilgreint sem samfelld röð af bókstöfum og/eða tölustöfum og táknum sem aðgreind eru með stafbili eða greinarmerkjum. Notuð var sú regla að reynt var að fella eins langan stafastreng undir töluorð og kostur var. Plúsar, mínusar og prósentumerki fylgdu þannig lesmálsorðunum. Blendingar tölustafa og annarra rittákna, t.d. efnafræðiformúlur og stærðfræðiformúlur teljast eitt lesmálsorð. Vert er að benda á að skammstafanir eru í flestum tilvikum greindar eins og lesið er úr þeim.

Í textasafninu eru 590.297 lesmálsorð sem birtast í 59.358 mismunandi orðmyndum að meðtöldum greinarmerkjum. Lesmálsorðunum fylgja 639 mismunandi greiningarstrengir að meðtöldum greinarmerkjum. Þegar fengist er við vélræna málfræðilega greiningu er erfiðast að eiga við orðmyndir sem geta haft fleiri en eina greiningu. Í textasafni Orðtíðnibókarinnar hafa 15,9% orðmynda fleiri en eina hugsanlega greiningu. Margræðasta orðmyndin er *minni* sem hefur 24 greiningarstrengi í textasafninu, en fleiri eru mögulegir (ég *minni* þig á það; ég geri þetta eftir *minni*; Nonni er *minni* en Siggi; o.s.frv.)

orð	**nefnimynd**	**mark**	**skýring**
ég	ég	fp1en	f: fn; p: pfn; 1: 1. pers.; e: et.; n: nefnifall
stökk	stökkva	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
á	á	aa	a: ao.; a: stýrir ekki falli
eftir	eftir	aþ	a: ao.; þ: stýrir þágufalli
stræto	stræto	nkeþ	n: no.; k: kk.; e: et.; þ: þgf.
og	og	c	c: samtenging
veifaði	veifa	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
,	,	,	komma
vagnstjórinn	vagnstjóri	nkeng	n: no.; k: kk.; e: et.; n: nf.; g: með greini
sá	sjá	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
mig	ég	fp1eo	f: fn.; p: pfn.; 1: 1. pers.; e: et.; n: þolfall
og	og	c	c: samtenging
stoppaði	stoppa	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
.	.	.	punktur

Mynd 1. Greining orða í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Hverju lesmálsorði var síðan komið fyrir í sérstakri línu. Í þeirri línu var einnig komið fyrir greiningarstreng orðsins eða marki og nefnimynd (flettimynd) þess. Í mynd 1 er sýnd ein setning úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson og hvernig hún er greind. Til glöggvunar er sýnd skýring á greiningarstrengjum.

Í formála Orðtíðnibókarinnar er gerð grein fyrir vélrænni greiningu sem notuð var við gerð bókarinnar (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991). Vélræna greiningin byggist á greiningu 54.000 lesmálsorða sem höfðu verið greind handvirkt og notuð við orðtíðnikönnun (Friðrik Magnússon 1988). Stefán Briem (1990) gerir grein fyrir aðferðum sem var beitt við vélrænu greininguna. Höfundar Orðtíðnibókarinnar telja að um 80% lesmálsorða hafi fengið rétta greiningu að öllu leyti með vélrænu greiningunni. Nokkrum árum seinna var forritið endurbætt á grundvelli greiningar alls textans. Fékkst þá tæplega 90% nákvæmni (Stefán Briem, munnlegar upplýsingar). Athyglisvert er að bera þá niðurstöðu saman við niðurstöðu tilraunarinnar sem hér verður greint frá.

Í greiningu lesmálsorða sem notuð var í Orðtíðnibókinni er greint á milli átta orðflokka: nafnorða, lýsingarorða, fornafna, lauss greinis, töluorða, sagna, atviksorða og samtenginga. Orð sem ekki flokkast í þessa orðflokka voru annað hvort talin erlend orð eða ógreind orð. Helstu frávik frá venjulegri orðflokkagreiningu voru þau að forsetningar voru taldar með atviksorðum. Þess vegna koma fyrir atviksorð sem stýra falli. Upphrópanir voru einnig taldar með atviksorðum. Nafnháttarmerki var talið með samtengingum. Í viðauka A er yfirlit yfir greiningarstrengi sem voru notaðir.

4 Aðferðir og markarar

Í þeirri könnun sem hér er greint frá voru eingöngu prófaðar gagnaaðferðir. Þær byggjast á því að forrit býr til líkan út frá fyrir fram greindu textasafni. Þetta safn kallast þjálfunarsafn. Aðferðin er síðan prófuð á sérstöku prófunarsafni. Til þess að prófa tiltekna mörkunaraðferð þarf að hafa aðgang að nokkuð stóru textasafni sem hefur verið greint í lesmálsorð og hverju lesmálsorði gefinn greiningarstrengur í samræmi við þá greiningu sem óskað er að fá fram. Textasafninu er skipt í tvo hluta og er annar hlutinn kallaður þjálfunarsafn og hinn hlutinn prófunarsafn. Þjálfunarsafnið er oft um 90% af textasafninu

sem er til umráða og prófunarsafnið um 10%. Búið er til líkan með aðstoð þjálfunarsafnsins og það síðan prófað á prófunarsafninu. Þar sem prófunarsafnið er líka fullgreint má reikna út hversu nákvæm aðferðin er.

Prófaðar voru fjórar gagnaaðferðir og fimm forrit eða markarar sem unnt er að þjálfa á íslenskum texta og eru fáanleg án greiðslu. Prófaðir voru tveir tölfræðimarkarar, **TnT** sem byggist á Markovslíkani og **MXPOST** sem byggist á svo kölluðu hámarksóreiðulíkani (e. *Maximum Entropy Model*). Prófaðir voru tveir markarar sem byggjast á leiðréttingaaðferð, μ**-TBL** og **fnTBL.** Einnig var prófaður einn markari, **MBT**, sem byggist á minnistækni. Alla þessa markara má kalla á íslensku gagnamarkara eða námfúsa markara. Markarinn μ-TBL hafði áður verið prófaður á litlum úrdrætti úr textasafni Orðtíðnibókarinnar og fékkst þá viðunandi niðurstaða (Sigrún Helgadóttir 2002). En markarinn virtist ekki ráða við allt textasafn Orðtíðnibókarinnar. Honum eru því ekki gerð frekari skil. Hér á eftir verður lauslega lýst þeim fjórum aðferðum við mörkun sem voru notaðar í tilrauninni og þeim mörkurum sem voru valdir til prófunar á íslenskum texta.

4.1 Falin Markovslíkön

Í þessum flokki var valinn markarinn TRIGRAMS'N'TAGS (TnT) sem Thorsten Brants samdi (Brants 2000a og 2000b). Aðferðinni verður lýst með því að skýra í stórum dráttum hvernig TnT-kerfið starfar.

TnT notar annars stigs Markovslíkön fyrir mörkun. Ástönd (e. *states*) standa fyrir mörk og færslulíkur eru háðar markapörum. Forritið metur færslulíkur og frálagslíkur út frá mörkuðu textasafni. Kerfið notar sennileikalíkur (e. *maximum likelihood probabilities*) sem eru reiknaðar út frá hlutfallslegri tíðni. TnT notar Viterbi-algrím með geislaleit (e. *beam search*) við mörkun til þess að flýta fyrir vinnslu.

TnT vinnur úr óþekktum orðum með því að greina endingar (viðskeyti) eins og lagt er til í Samuelsson (1993) þar sem líkur á mörkum eru stilltar í samræmi við endingar orða. Lengsta ending sem TnT notar er 10 stafa löng (10 er sjálfgildi í forritinu). Líkindadreifing fyrir tiltekna endingu er búin til með því að skoða öll orð í þjálfunarsafni sem hafa sameiginlega endingu af tiltekinni hámarkslengd. Forritið vinnur aðeins úr endingum orða sem hafa tiltekna lágmarkstíðni og var tíðnin 10 valin út frá reynslu. Forritið heldur einnig tvo lista yf-

ir endingar, einn fyrir orð sem hefjast á lágstaf og einn fyrir orð sem hefjast á hástaf.

TnT er beitt á nýtt mál eða nýtt svið í tveimur þrepum:

1. Líkan er búið til
2. Texti er markaður

Líkanið er búið til út frá þjálfunarsafninu. Tvær skrár verða til í því skrefi: skrá með tíðni orða og marka sem þau geta fengið og skrá með tíðni tveggja eða þriggja marka sem standa saman. Þessar skrár eru síðan notaðar þegar forritið markar nýjan texta. Forritið gefur einnig kost á að nota viðbótarorðasafn. Finnist orð ekki í orðasafninu, sem var búið til þegar líkanið var gert, er leitað að því í viðbótarorðasafninu.

4.2 Hámarksóreiðuaðferð

Í þessum flokki var valinn markarinn **MXPOST** (Ratnaparkhi 1996). Í Ratnaparkhi (1997) er inngangur að því hvernig hámarksóreiðulíkön (e. *Maximum Entropy Models*) eru notuð við málgreiningu. Ratnaparkhi segir þar að mörg málgreiningarverkefni megi endurskilgreina sem tölfræðileg flokkunarverkefni. Verkefnið felst í því að meta líkur á að flokkur *a* komi fyrir í „samhenginu“ *b*, eða *p(a,b)*. Í málgreiningarverkefnum eru orð venjulega hluti af „samhenginu“. Í sumum verkefnum er „samhengið“ aðeins eitt orð en í öðrum getur *b* verið nokkur orð og greiningarstrengir þeirra. Í stórum textasöfnum fæst nokkur vitneskja um hvenær *a* og *b* koma fyrir saman. En ekkert textasafn hefur nægilegar upplýsingar til þess gefa upplýsingar um *p(a,b)* fyrir öll hugsanleg pör *(a,b)* þar sem orðin í *b* eru sjaldgæf. Vandamálið snýst um að meta á öruggan hátt líkindalíkanið *p(a,b)* með því að nota ófullkomnar upplýsingar um *a*-in og *b*-in.

Þjálfunarsafninu er lýst sem miklum fjölda af sérkennaþáttum (e. *features*). Þessir sérkennaþættir eru tvígild föll af „sögum“ (e. *histories*) (samhengi orða og greiningarstrengja) og greiningarstrengjum. Í útgáfu Ratnaparkhis eru sérkennaþættir orðið sem verið er að fjalla um, næstu tvö orð á undan, næstu tvö orð á eftir og greiningarstrengur (mark) næstu tveggja orða á undan. Sérkennaþættir sjaldgæfra og óþekktra orða (koma ekki fyrir í þjálfunarsafni) hafa einnig fyrstu og síðustu fjóra stafi orðs og upplýsingar um hvort orðið hafi hástaf, bandstrik eða tölustaf. Sérkennaþættir óþekktra orða eru búnir til úr

sérkennaþáttum sjaldgæfra orða. Litið er á sérkennaþætti, sem koma fyrir sjaldnar en 10 sinnum í þjálfunarsafni, sem óáreiðanlega. Markarinn notar geislaleit til þess að finna líklegustu runu marka og sú röð sem hefur hæst líkindi er valin. MXPOST-forritið gefur ekki kost á því að nota viðbótarorðasafn. Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

4.3 Leiðréttingaaðferð

Brill (1994 og 1995) hefur lýst leiðréttingaaðferðinni og hvernig má beita henni við mörkun texta. Með þessari aðferð er málfræðileg þekking skráð í nokkrum einföldum reglum. Fyrst er hverju orði í þjálfunarsafninu gefinn sá greiningarstrengur sem er líklegastur miðað við þjálfunarsafnið sjálft. Þessi mörk eru síðan borin saman við rétt mörk. Forritið lærir leiðréttingareglur sem er beitt til þess að komast nær hinni réttu greiningu. Forritið lærir reglurnar út frá sniðmátum sem lýsa aðgerð (breyta greiningarstreng A í greiningarstreng B) á grundvelli tiltekins umhverfis (orð og mörk í samhengi), þ.e. hvaða orð og mörk eru næst á undan og eftir því marki sem verið er að skoða. Upphaflega gerði Brill ráð fyrir því að aðeins væru skoðuð mörk í næsta nágrenni við markið sem var skoðað. Síðar bætti hann við sniðmátum þar sem gert var ráð fyrir að orð væru skoðuð líka. Forritið sem lærir leiðréttingareglurnar beitir öllum leiðréttingum, telur hversu margar villur hver leiðrétting lagar og velur þá leiðréttingu sem lagar flestar villur. Ákveðið er fyrir fram hver er minnsti fjöldi leiðréttinga sem regla þarf að hafa í för með sér til þess að vera valin. Þegar engar leiðréttingar finnast sem fækka villum um þann fjölda hættir forritið að læra reglur. Á þennan hátt verður til raðað mengi af leiðréttingareglum, hver regla endurspeglar tiltekið sniðmát.

Í fyrstu tilraunum sínum gerði Brill (1994) ráð fyrir því að engin óþekkt orð væru í þeim texta sem átti að marka. Síðar þróaði Brill aðferð til þess að greina óþekkt orð. Aðferðin byggist líka á því að láta forrit læra leiðréttingareglur. Óþekktum orðum eru gefin mörk. Brill gefur óþekktum orðum sem hefjast á lágstaf mark sem venjuleg nafnorð og óþekktum orðum sem hefjast á upphafsstaf mark sem sérnöfn. Síðan eru skilgreind sniðmát. Sniðmát Brills fela í sér að skoðaðir eru fyrstu og síðustu fjórir stafir í orði. Athugað er hvort orðið hafi forskeyti eða viðskeyti sem er eins til fjögurra stafa langt, hvort unnt sé

að taka í burtu eða bæta við eins til fjögurra stafa forskeyti eða viðskeyti og fá út nýtt orð, hvort orðið hafi tiltekinn staf eða hvort tiltekið orð sé til vinstri eða hægri við orðið. Á grundvelli þessara sniðmáta lærir forritið reglur til þess að beita. Í Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir (2002) er góð lýsing á aðferðinni.

Nokkur forrit bjóðast sem nota aðferðir Brills. Valið var að nota forritið **fnTBL** (*Fast Transformation-Based Learning Toolkit*) eftir Florian og Ngai (2002). Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

4.4 Minnistækni

Námfús markari sem byggist á minnisaðferð lærir af mengi dæma sem eru geymd í gagnasafni þar sem hvert dæmi hefur verið flokkað á tiltekinn hátt og markað í samræmi við það. Dæmin eru tekin úr dæmasafni sem hefur verið handmarkað. Þegar flokka á ný dæmi leitar kerfið í gagnasafninu að dæmi eða mengi dæma sem líkjast sem mest nýja dæminu. Þetta má orða þannig að leitað sé að „næsta nágranna“ nýja dæmisins. Aðferðin er dregin af tækni sem kennd er við „*k* næstu nágranna“ og aðeins *k* næstu nágrannar eru skoðaðir. Oft er *k* látið vera 1 en í tilraunum með minnisaðferð er það eitt af markmiðunum að finna besta gildi fyrir *k*.

Til þess að prófa þessa aðferð var valinn **MBT**-markarinn (Daelemans o.fl. 2003) sem notar Tilburg Memory-Based Learner (*TiMBL*, Daelemans o.fl. 2004). MBT-markarinn býr til aðskilin gagnasöfn fyrir þekkt orð og óþekkt orð. *TiMBL*-kerfið notar svokallað IGTREE-algrím fyrir þekkt orð og IB1-algrím fyrir óþekkt orð. Forritinu er beitt á nýtt mál eða svið á líkan hátt og lýst var fyrir TnT-forritið.

5 Mörkun íslensks texta

Í þessum kafla verður gerð grein fyrir tilraunum við að marka texta Orðtíðnibókarinnar. Skipulagi skráa við tilraunina verður einnig lýst og jafnframt hvernig niðurstöður verða metnar.

5.1 Skrár

Til þess að prófa mismunandi aðferðir við mörkun er oft notuð aðferð sem byggist á því að hafa til umráða tíu pör af þjálfunar- og prófunarsöfnum. Í hverri tölvuskrá Orðtíðnibókarinnar er textabútur úr einni heimild. Pörin voru búin til þannig að hverri skrá var skipt upp í tíu nokkurn veginn jafna hluta. Hver þessara tíu hluta myndar eitt prófunarsafn og samstætt þjálfunarsafn hefur að geyma hina hlutana níu í hvert sinn. Stærri skráin er notuð sem þjálfunarsafn og sú minni sem prófunarsafn. Prófunarsöfnin skarast því 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*).

5.2 Mælikvarðar fyrir nákvæmni

Til þess að finna hversu nákvæmlega markari úthlutar mörkum eru mörk hans borin saman við „rétt mörk“ sem hafa verið yfirfarin handvirkt. Þessi réttu mörk eru oft kölluð á ensku „gold standard“. Mjög erfitt er að ná 100% réttri mörkun þar sem ýmis álitamál koma upp. Tveir einstaklingar mörkuðu textasafn Orðtíðnibókarinnar og hafa þeir efalaust borið sig saman. Nákvæmnin hefur ekki verið rannsökuð sérstaklega.

Til þess að geta metið árangur tiltekins markara þarf að hafa viðmiðun um lægstu nákvæmni, *grunnmörkun* (e. *baseline tagging*), sem unnt er að ná án þess að nota markarann. Grunnmörkun er gerð með tilteknu orðasafni sem hefur upplýsingar um orðmyndir og mörk þeirra og með því að nota tiltekna aðferð við mörkun óþekktra orða. Hér er notuð ein af fjórum skilgreiningum á grunnmörkun sem kemur fram í Megyesi (2002:55). Búið er til orðasafn úr hverju þjálfunarsafni og þeim orðmyndum í viðkomandi prófunarsafni sem koma líka fyrir í þjálfunarsafninu gefið algengasta mark þeirrar orðmyndar. Óþekkt orð rituð með litlum staf fá algengustu greiningu nafnorða (*nken*) og óþekkt orð rituð með upphafsstaf fá algengustu greiningu sérnafna (*nken-m*). Þessi mörk eru síðan borin saman við rétt mörk. Meðalnákvæmni slíkrar mörkunar fyrir öll prófunarsöfnin reyndist **76,63%.** Markari sem nær ekki þessari nákvæmni bætir því engu við það sem fæst með grunnmörkun eingöngu.

Frammistaða hvers markara er metin með því að reikna út hittni (e. *accuracy*) miðað við rétta greiningu (handmörkun) og reiknuð sem

> *hittni=(fjöldi rétt greindra lesmálsorða)/(heildarfjöldi lesmálsorða í safni)*

Tökum sem dæmi að í prófunarsafni séu 59.169 lesmálsorð. Tiltekinn markari markar 53.101 lesmálsorð eins og gert var með handmörkun. Hittni markarans fyrir öll orð er því 53.101/59.169 eða 89,74%.

Einnig má athuga hvernig markaranum tekst að greina einstaka greiningarflokka. Þá má nota mælikvarðana nákvæmni (e. *precision*), griphlutfall (e. *recall*) og *F*-gildi. Þessa mælikvarða má nota til þess að kanna hvaða villur markararnir gera. *Nákvæmni* segir til um hversu rétt markarinn greinir tiltekið mark, en *griphlutfall* segir til um hlutfall hvers marks af þeim mörkum sem markarinn finnur (Megyesi 2002: 53–54). Megyesi skilgreinir þessar stærðir þannig fyrir tiltekið mark X:

> *nákvæmni* (*P*)= (*fjöldi rétt greindra lesmálsorða sem hafa mark* X)/(*heildarfjöldi lesmálsorða sem markari greinir með mark* X)
>
> *griphlutfall (R)=(fjöldi rétt greindra orða sem hafa mark X)/(heildarfjöldi orða með mark X í safni)*

F-gildið er vegið umhverfumeðaltal (harmonic mean) af *P* og *R*.
Manning og Schütze (1999: 269) skilgreina *F*-gildi sem
$F=1/(\alpha*(1/P)+(1-\alpha)*(1/R))$
og Megyesi (2002: 32) skilgreinir *F*-gildið sem
$F=(\beta^2+1)*P*R/(\beta^2*P+R)$
Ef β=1 og α=0,5 er *F*-gildið hreint umhverfumeðaltal af *P* og *R*:
$F=P*R/((R+P)/2)=2*P*R/(P+R)$
Í Manning og Schütze (1999: 267–269) er góð lýsing á því hvernig á að finna *P* og *R*.
Tökum sem dæmi að við viljum finna *P*, *R* og *F* fyrir hvernig tiltekinn markari, t.d. TnT, greinir atviksorð í einu prófunarsafni íslenska verkefnisins.
Í prófunarsafninu eru 11.660 atviksorð. TnT greinir 11.451 af þeim rétt (tp, *true positives* samkvæmt Manning og Schütze).
Fjöldi orða sem TnT greinir sem atviksorð = 11.716 (*selected* með orðalagi Manning og Schütze).

Fjöldi orða sem TnT greinir rangt sem atviksorð = 11.716-11.451=265 (fp, *false positives* með orðalagi Manning og Schütze)
Fjöldi orða sem eru atviksorð en TnT greinir sem eitthvað annað = 11.660-11.451=209 (fn=*false negatives* með orðalagi Manning og Schütze)
Þá er
P=tp/(tp+fp)=tp/(valið)=11.451/11.716 = 0,977
R=tp/(tp+fn)=tp/(það sem átti að velja)=11.451/11.660 = 0,982
F=2**P***R*/(*P*+*R*)= = 0,980
Þessar stærðir má reikna fyrir hvaða greiningarstreng sem er.

Í íslensku hefur ekki skapast sú hefð að gera greinarmun á hittni (*accuracy*) og nákvæmni (*precision*) heldur er orðið *nákvæmni* notað um hvort tveggja. Þar sem ekki er hætta á ruglingi er þeirri hefð fylgt í þessari grein.

6 Prófanir

Allir markararnir sem voru valdir voru þjálfaðir á þjálfunarsöfnunum 10 og prófaðir á samsvarandi prófunarsöfnum. Í upphaflegu tilraunum sem voru gerðar 2002–2004 fengust niðurstöður með þremur mörkurum, TnT, MXPOST og fnTBL. Tilraunin með MBT-markarann var gerð í nóvember 2005 (Sigrún Helgadóttir og Örvar Hafsteinn Kárason 2005).

	Meðalnákvæmni		
Markari	**Óþekkt orð** **%**	**Þekkt orð** **%**	**Öll orð** **%**
fnTBL	54,02	91,36	88,80
MXPOST	62,51	91,04	89,08
TnT	71,62	91,74	90,36
MBT	56,86	89,21	87,00

Tafla 1. Niðurstaða af þjálfun og mörkun 10 para skráa

Niðurstöður prófana eru sýndar í töflu 1. Eins og sést á töflunni eiga markararnir fjórir misjafnlega auðvelt með að greina óþekkt orð, þ.e. orð sem koma ekki fyrir í viðkomandi þjálfunarsafni og þeir hafa því ekki séð áður. Fundið var hlutfall óþekktra orða í hverju prófunarsafni

og reiknað meðaltal fyrir prófunarsöfnin 10 og reyndist það 6,84%. Markararnir nota mismunandi aðferðir við greiningu óþekktra orða. TnT-markarinn virðist hafa yfir að ráða betri aðferð en hinir markararnir við að greina óþekkt orð og fær því besta heildarniðurstöðu eða **90,36**%.

Vert er að benda á að mark er talið rangt þó að aðeins eitt af 6 atriðum í greiningarstreng sé rangt.

Mismunur á mörkunarnákvæmni TnT og fnTBL er 1,56 prósentustig. Við það að nákvæmni hækkar úr 88,80% í 90,36% fækkar villum um 14%.

Dreifing orðmynda eftir orðflokkum er ólík meðal óþekktra orða og allra orða. Nafnorð, lýsingarorð og sagnir, eru að meðaltali um 44,3% af öllum orðum í prófunarsöfnunum en um 95,9% að meðaltali af óþekktum orðum.

Gert var parað t-próf á hlutfalli rangt greindra orða til þess að kanna hvort tölfræðilega marktækur munur væri á árangri þeirra þriggja markara sem náðu bestum árangri. Niðurstaða prófsins fyrir pörin fnTBL/TnT, MXPOST/TnT og fnTBL/MXPOST er sýnd í töflu 2. Munur á mörkurum er marktækur í öllum tilvikum ($p<0,05$).

Samanburður	**t**	**frítölur**
fnTBL/TnT	40,16	9
MXPOST/TnT	30,94	9
fnTBL/MXPOST	5,37	9

Tafla 2. Parað t-próf á mismuni á hlutfalli rangt greindra orða

7 Greining á niðurstöðum

Niðurstöður þeirra þriggja markara (TnT, MXPOST og fnTBL) sem náðu bestum árangri voru skoðaðar nánar.

Fyrir hvern greiningarstreng var reiknuð nákvæmni (*precision*, *P*), griphlutfall (*recall*, *R*) og *F*-gildi. Niðurstöður útreikninganna eru ekki sýndar hér þar sem þær taka of mikið pláss. Í töflu 3 er sýndur sambærilegur útreikningur fyrir orðflokka.

Markararnir hegða sér á líkan hátt nema fyrir þá orðflokka sem hafa fá orð, þ.e. *e* (erlend orð), *g* (greinir) og *x* (ógreint). TnT fær t.d. hærri nákvæmni en griphlutfall fyrir greininn þar sem markarinn greinir tiltölulega fá orð sem greini en MXPOST fær hærra griphlutfall

en nákvæmni þar sem sá markari greinir fleiri orð sem greini en ættu að fá þá greiningu.

Orðflokkar	Fjöldi í safni	fnTBL			MXPOST			TnT		
		P	R	F (β=1)	P	R	F (β=1)	P	R	F (β=1)
a (atviksorð)	116.112	98,02	98,31	98,16	97,53	98,25	97,89	98,04	98,11	98,07
c (samtengingar)	60.256	98,64	99,05	98,84	98,39	98,95	98,67	98,41	98,92	98,67
e (erlend orð)	411	54,20	37,71	44,48	72,19	56,20	63,20	85,53	63,26	72,73
f (fornöfn)	74.315	98,99	98,71	98,85	99,14	98,25	98,69	98,81	98,84	98,82
g (greinir)	632	82,15	84,49	83,31	78,77	87,50	82,91	94,22	77,37	84,97
l (lýsingarorð)	35.669	89,69	86,15	87,88	88,90	86,00	87,43	93,48	91,74	92,60
n (nafnorð)	122.621	96,31	96,73	96,52	96,63	96,98	96,80	98,48	98,57	98,53
s (sagnorð)	103.136	96,54	97,00	96,77	97,27	97,52	97,39	97,76	98,22	97,99
t (tölуorð)	5.901	92,85	95,03	93,93	94,44	93,90	94,17	95,02	93,12	94,06
x (ógreint)	127	63,64	44,09	52,09	70,49	33,86	45,74	58,40	57,48	57,94

Tafla 3. Nákvæmni (P), griphlutfall (R) og F-gildi fyrir orðflokka

Í töflu 4 er griphlutfall greint í sundur eftir því hvort mörkurunum tekst að greina öll atriði í greiningarstreng rétt eða a.m.k. orðflokkinn rétt. Hlutfallstölur eru reiknaðar af heildarfjölda lesmálsorða í orðflokki í safninu.

Fyrsti dálkur fyrir hvern markara sýnir hlutfall rétt greindra strengja af heildarfjölda slíkra strengja í safninu, annar dálkur sýnir hlutfall þar sem orðflokkur er réttur en einhver greiningaratriði röng og síðasti dálkurinn sýnir summu þessara dálka sem er griphlutfallið fyrir orðflokkinn eins og sýnt er í töflu 3. Fyrir utan sjaldgæfa og erfiða orðflokka (*e*, *g* og *x*) virðast allir markararnir eiga í mestum erfiðleikum með að greina lýsingarorð rétt. Þetta virðist eiga við um orðflokkinn sjálfan og einnig virðist erfitt að greina rétt hinar ýmsu greiningarmyndir. Lýsingarorð í íslensku geta fræðilega haft 120 beygingarmyndir. Sumar eru mjög sjaldgæfar þannig að það kemur ekki á óvart að markararnir eigi erfitt með að búa til reglur um hvernig eigi að greina þær.

Orðflokkur	Fjöldi í safni	fnTBL			MXPOST			TnT		
		Grein. str. réttur	Orðfl. réttur	R	Grein. str. réttur	Orðfl. réttur	R	Grein. str. réttur	Orðfl. réttur	R
a (atviksorð)	116.112	93,54	4,77	98,31	92,83	5,41	98,25	92,22	5,89	98,11
c (samtengingar)	60.256	97,71	1,34	99,05	97,09	1,86	98,95	97,14	1,79	98,92
e (erlend orð)	411	37,71	0,00	37,71	56,20	0,00	56,20	63,26	0,00	63,26
f (fornöfn)	74.315	89,38	9,33	98,71	88,15	10,10	98,25	89,46	9,38	98,84
g (greinir)	632	66,77	17,72	84,49	66,14	21,36	87,50	64,72	12,66	77,37
l (lýsingarorð)	35.669	64,09	22,05	86,15	66,99	19,01	86,00	72,88	18,86	91,74
n (nafnorð)	122.621	78,97	17,76	96,73	80,19	16,79	96,98	84,48	14,09	98,57
s (sagnorð)	103.136	91,89	5,12	97,00	92,94	4,58	97,52	92,64	5,58	98,22
t (tölуorð)	5.901	69,17	25,86	95,03	71,65	22,25	93,90	73,34	19,78	93,12
x (ógreint)	127	44,09	0,00	44,09	33,86	0,00	33,86	57,48	0,00	57,48

Tafla 4. Sundurliðun griphlutfalls fyrir orðflokka eftir því hvort allur greiningarstrengur er rétt greindur eða a.m.k. orðflokkur. Hlutfallstölur eru reiknaðar af fjölda lesmálsorða í hverjum orðflokki í safni

Niðurstöður mörkunar voru skoðaðar og greindar til þess að finna hvers konar villur markararnir gera og hvernig mætti bæta árangurinn.

Skipta má villum sem markarar gera í tvo flokka. Í fyrsta lagi eru villur sem verða vegna margræðni, þ.e. tiltekin orðmynd getur haft fleiri en eina greiningu. Í öðru lagi verða villur í greiningu óþekktra orða þegar sú aðferð við greiningu óþekktra orða sem markarinn notar gefur ekki rétta greiningu.

Í töflu 5 eru sýndar 20 algengustu villur sem hver markari gerir. Í töflu 6 eru sýndar 20 algengust villur sem allir markarar eru sammála um.

fnTBL				MXPOST				TnT			
markari> rétt mark	Tíðni	%	Safn-tíðni %	markari> rétt mark	Tíðni	%	Safn-tíðni %	markari> rétt mark	Tíðni	%	Safn-tíðni %
aþ>ao	1.568	2,37	2,37	aþ>ao	2.218	3,44	3,44	aþ>ao	1.734	3,05	3,05
ao>aþ	1.522	2,30	4,68	ao>aþ	1.514	2,35	5,79	ao>aþ	1.489	2,62	5,66
nveo>nveþ	830	1,26	5,93	aa>ao	616	0,96	6,75	ao>aa	1.045	1,84	7,50
nveþ>nveo	824	1,25	7,18	ao>aa	599	0,93	7,68	aþ>aa	911	1,60	9,10
sng>sfg3fn	672	1,02	8,19	nveþ>nveo	586	0,91	8,59	nveþ>nveo	887	1,56	10,66
nheo>nhen	594	0,90	9,09	nveo>nveþ	547	0,85	9,44	nveo>nveþ	865	1,52	12,18
nhen>nheo	582	0,88	9,97	sfg3eþ>sfg1eþ	503	0,78	10,22	aa>ao	689	1,21	13,39
sfg3eþ>sfg1eþ	572	0,87	10,84	nhen>nheo	489	0,76	10,98	ssg>sþghen	671	1,18	14,57
aa>ao	562	0,85	11,69	sfg3fn>sng	446	0,69	11,67	nheo>nhen	659	1,16	15,73
nkeo>nkeþ	500	0,76	12,45	c>ct	392	0,61	12,28	nhen>nheo	638	1,12	16,85
aa>aþ	462	0,70	13,14	aa>aþ	378	0,59	12,86	sng>sfg3fn	599	1,05	17,91
ao>aa	449	0,68	13,82	nheo>nhen	371	0,58	13,44	sfg3eþ>sfg1eþ	584	1,03	18,93
lhensf>lheosf	441	0,67	14,49	nkeþ>nkeo	360	0,56	14,00	sþghen>ssg	570	1,00	19,93
nvfo>nvfn	420	0,64	15,13	nvfn>nvfo	337	0,52	14,52	nkeþ>nkeo	509	0,89	20,83
nkeþ>nkeo	412	0,62	15,75	fpkeþ>fpveþ	335	0,52	15,04	lhensf>lheosf	490	0,86	21,69
fohen>foheo	401	0,61	16,36	sng>sfg3fn	334	0,52	15,56	c>aa	437	0,77	22,46
nheog>nheng	392	0,59	16,95	aþ>aa	330	0,51	16,07	nvfo>nvfn	437	0,77	23,23
ct>c	369	0,56	17,51	nkeo>nkeþ	327	0,51	16,58	nkeo>nkeþ	434	0,76	23,99
ssg>sþghen	359	0,54	18,05	ct>c	324	0,50	17,08	nvfn>nvfo	424	0,75	24,73
nvfn>nvfo	356	0,54	18,59	fohen>foheo	321	0,50	17,58	ct>c	393	0,69	25,42

Tafla 5. Tuttugu algengustu villur sem hver markari gerir

Algengustu villur sem markararnir gera eru af fyrri gerðinni, þ.e. orsakast af margræðni. Langalgengustu villurnar felast í því að rugla saman fallstjórn forsetninga. Algengast er að rugla saman þolfalli og þágufalli. Þar sem prófaðir eru gagnamarkarar í þessari rannsókn hafa þeir ekki innbyggðar reglur sem segja til um samræmi í fallstjórn forsetninga og falli eftirfarandi nafnorðs. Markararnir gætu hins vegar búið sér til slíkar reglur út frá gögnunum. Sá þáttur hefur ekki verið kannaður til hlítar. Næstalgengastur er ruglingur á milli beygingarmynda nafnorða sem hafa sömu mynd. Má þar nefna þolfall og þágufall kvenkynsorða í eintölu (þf. *konu*; þgf. *konu*) og nefnifall og þolfall hvorugkynsorða í eintölu (nf. *barn*; þf. *barn*). Ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum er líka algengur þar sem þessar beygingarmyndir líta eins út (*ég fer*; *hann fer*). Einnig má nefna

nafnhátt og þriðju persónu fleirtölu í nútíð en þessar beygingarmyndir líta eins út (*að fara; þeir fara*).

Eins og sést af töflu 5 gera markararnir misjafnlega margar villur. Þeir gera einnig misjafnlega fjölbreytilegar villur. TnT gerir 5.373 mismunandi villur, fnTBL gerir 5.897 mismunandi villur og MXPOST gerir 7.115 mismunandi villur. Þar sem markaskrá Orðtíðnibókarinnar er mjög stór er unnt að gera mjög margvíslegar villur. Fræðilega má gera 552*552 = 304.704 mismunandi villur ef tiltekið safn sem á að marka hefur 552 ólík mörk.

Af töflu 5 sést að fyrstu 20 villur sem TnT gerir skýra um 25% af villum sem markarinn gerir, fyrir fnTBL er þessi tala rúmlega 18% og rúmlega 17% fyrir MXPOST.

	Tíðni	**%**	**Safntíðni %**
markari>rétt			
aþ>ao	499	3,82	3,82
sfg3eþ>sfg1eþ	457	3,50	7,32
ao>aþ	361	2,77	10,09
sng>sfg3fn	235	1,80	11,89
nveþ>nveo	214	1,64	13,53
nveo>nveþ	212	1,62	15,15
sfg3eþ>svg3eþ	203	1,55	16,71
ao>aa	190	1,46	18,16
lhensf>lheosf	170	1,30	19,46
fpkeþ>fpveþ	167	1,28	20,74
nhen>nheo	163	1,25	21,99
aa>ao	148	1,13	23,13
fohen>foheo	144	1,10	24,23
ct>c	141	1,08	25,31
c>aa	128	0,98	26,29
nheo>nhen	126	0,97	27,25
nkeo>nkeþ	118	0,90	28,16
nken-m>nkeo-m	112	0,86	29,02
lvensf>lhfnsf	111	0,85	29,87
nkeþ>nkeo	110	0,84	30,71

Tafla 6. Tuttugu algengustu villur sem allir markarar gera

Af töflu 6 sést að næstalgengasta villa sem allir markarar gera sameinginlega er ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum. Þegar litið er á sameiginlegar villur er þessi villa algengari en ruglingur á milli þolfalls og þágufalls eintölu af kvenkynsnafnorð-

um. Ef leitað verður leiða til þess að koma í veg fyrir sumar algengustu villurnar væri e.t.v. skynsamlegt að hafa þennan lista að leiðarljósi.

Mynd 2 sýnir hvernig þrír markarar, TnT, MXPOST og fnTBL greina orð í setningunni sem einnig er sýnd í mynd 1. Þar sjást vel dæmi um helstu atriði sem markararnir eiga erfitt með að greina rétt. TnT og fnTBL markararnir greina báðir rangt fallstjórn orðsins *eftir* og einnig rangt fall nafnorðsins *strætó* en það er eins í nefnifalli, þolfalli og þágufalli. Allir markararnir greina rangt persónu sagnarinnar *veifaði*. Í töflu 6 sást einmitt að önnur algengasta sameiginlega villa allra markaranna er að rugla saman fyrstu og þriðju persónu sagna í þátíð.

orð	**mark**	**tnt**	**mxp**	**fnt**
ég	fp1en	fp1en	fp1en	fp1en
stökk	sfg1eþ	sfg1eþ	sfg1eþ	sfg1eþ
á	aa	aa	aa	aa
eftir	aþ	**ao**	aþ	**aa**
strætó	nkeþ	**nkeo**	nkeþ	**nkeo**
og	c	c	c	c
veifaði	sfg1eþ	**sfg3eþ**	**sfg3eþ**	**sfg3eþ**
,	,	,	,	,
vagnstjórinn	nkeng	nkeng	nkeng	nkeng
sá	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
mig	fp1eo	fp1eo	fp1eo	fp1eo
og	c	c	c	c
stoppaði	sfg3eþ	sfg3eþ	sfg3eþ	sfg3eþ
.	.	.	.	.

Mynd 2. Greining þriggja markara á orðum í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Fundið var hversu oft markararnir þrír voru sammála og fengu annaðhvort rétta eða ranga niðurstöðu. Tafla 7 sýnir hversu oft allir þrír markarar komast að réttri niðurstöðu, hversu oft tveir komast að réttri niðurstöðu og hversu oft aðeins einn hefur rétt fyrir sér. Í 95,47% tilvika hefur a.m.k. einn markari fundið rétt mark. Það er því hæsta fræðilega nákvæmni sem má ná fyrir þann efnivið sem prófaður var með því að sameina niðurstöður tveggja eða fleiri markara.

	Tíðni	**%**	**Safntíðni %**
3 réttir	484.294	82,04	82,04
2 réttir	51.322	8,69	90,74
1 réttur	27.941	4,73	95,47
Enginn réttur	26.740	4,53	100,00

Tafla 7. Hversu margir markarar eru sammála um rétt mark?

Í töflu 8 er sýndur paraður samanburður á mörkurum. Þar sést að TnT og fnTBL eru oftar sammála um rétt mark (og rangt mark) en önnur pör. Það gæti bent til þess að niðurstöður TnT og fnTBL séu með einhverjum hætti líkar þó að TnT gefi umtalsvert betri niðurstöðu. Það gæti því verið að unnt sé að bæta niðurstöðu TnT með niðurstöðu MXPOST.

Par	**Sama mark rétt** %	**Sama mark rangt** %	**Samtals** %
TnT og MXPOST	85,11	3,03	88,14
TnT og fnTBL	85,56	3,64	89,20
MXPOST og fnTBL	84,15	3,14	87,29

Tafla 8. Samanburður á markarapörum

8 Frammistaða markara bætt

Ýmsum aðferðum má beita til þess að bæta niðurstöður mörkunar. Stundum er reynt að bæta frammistöðu einstakra markara og einnig má sameina niðurstöðu tveggja eða fleiri markara. Í þeirri könnun sem hér er greint frá var gerð tilraun til þess að nota orðasafn til þess að bæta frammistöðu einstakra markara. Einnig var beitt tveimur aðferðum við að sameina niðurstöður markara.

8.1 Áhrif aukaorðasafns á mörkun

Við tilraunina voru notuð forritin TnT og fnTBL þar sem þau gefa kost á að nota viðbótarorðasafn.

Búið var til orðasafn sem hefur um helming þeirra orða sem eru óþekkt í hverju prófunarsafni miðað við samsvarandi þjálfunarsafn og það notað sem viðbótarorðasafn við mörkun með TnT og fnTBL.

Orðasafnið var gert þannig að búinn var listi yfir orð í hverju prófunarsafni sem voru óþekkt miðað við samstætt þjálfunarsafn og listarnir síðan sameinaðir í eitt safn. Síðan var tekið annað hvert orð úr þessu safni og notað sem viðbótarorðasafn. Safnið ætti að geyma um helming óþekktra orða í hverju prófunarsafni. Í töflu 9 sést niðurstaða fyrir mörkun með þessu orðasafni. Til samanburðar eru tölur fyrir mörkun án orðasafns hafðar með í töflunni.

	Meðalnákvæmni án orðasafns			**Meðalnákvæmni með orðasafni***		
Markari	**Óþekkt orð %**	**Þekkt orð %**	**Öll orð %**	**Óþekkt orð %**	**Þekkt orð %**	**Öll orð %**
fnTBL	54,02	91,36	88,80	70,44	91,50	90,06
TnT	71,62	91,74	90,36	86,31	91,93	91,54
*Notað er orðasafn sem hefur um helming þeirra orða sem álitin eru óþekkt frá sjónarhóli hvers prófunarsafns						

Tafla 9. Niðurstaða af þjálfun og mörkun 10 para skráa

Mörkun óþekktra orða batnar umtalsvert og hefur það áhrif á heildarniðurstöðu. Mörkun þekktra orða batnar einnig aðeins og er það sennilega afleiðing af bættri mörkun óþekktu orðanna. Þegar fleiri óþekkt orð fá rétta greiningu gefa þau betri vísbendingar um rétta mörkun þekktu orðanna í kring. Heildarnákvæmni með mörkun fnTBL hækkar meira en heildarnákvæmni með TnT. Ástæðan gæti verið sú að fnTBL-markarinn virðist eiga erfiðara með að marka óþekkt orð og þess vegna batnar mörkun óþekktra orða ef orðasafn er til staðar til þess að greina þau. Með því að nota viðbótarorðasafn nær TnT-markarinn **91,54%** nákvæmni og villum fækkar um 12%. Þessar niðurstöður sýna að mörkun ætti að batna ef unnt er að nota orðasafn. Þeir markarar sem voru prófaðir nota orðasöfn sem hafa tiltekið snið. Nauðsynlegt er að í viðbótarorðasafni séu upplýsingar um hlutfallslegt vægi mismunandi greiningarstrengja þeirra orðmynda sem geta haft fleiri en einn greiningarstreng.

8.2 Sameina niðurstöður markara

Nefna má þrjár aðferðir sem koma til greina við að sameina niðurstöður tveggja eða fleiri markara.

1. Kosið er um hvaða markari er valinn

2. Nýr markari er þjálfaður á grundvelli niðurstaðna úr tveimur eða fleiri mörkurum

3. Notaðar eru málfræðireglur

Í þessu verkefni voru prófaðar tvær af þessum aðferðum, þ.e. að kjósa á milli markara og að nota málfræðireglur.

Í Halteren o.fl. (2001) er yfirlit yfir aðferðir við að sameina niðurstöður tveggja eða fleiri markara. Markmiðið er að ná meiri nákvæmni en fæst með þeim einstökum markara sem gefur bestar niðurstöður. Í greininni er gerð grein fyrir tveimur mismunandi aðferðum við að sameina niðurstöður. Í fyrsta lagi er greint frá nokkrum aðferðum við að kjósa á milli markara. Í öðru lagi er greint frá leiðum til þess að þjálfa nýjan markara á grundvelli niðurstaðna markaranna og rétts marks. Í þessari rannsókn voru aðeins prófaðar aðferðir við að kjósa á milli markara.

Í því verki sem hér er lýst var gerð tilraun með fjögur afbrigði af kosningaaðferðinni. Einfaldasta aðferðin byggist á því að velja það mark sem flestir velja. Ef ekki er unnt að velja þannig er notuð slembitala til þess að velja á milli marka. Annað afbrigðið byggist á því að vega með fyrir fram þekktri heildarnákvæmni hvers markara. Í þriðja afbrigðinu er vegið með nákvæmni fyrir hvert mark. Einnig má vega með nákvæmni og griphlutfalli hvers marks. Vegið er með *nákvæmni* sem segir til um hvernig markarinn stendur sig og (*1-griphlutfalli*) sem segir til um hve oft markaranum mistekst að finna rétta markið. Hvert mark fær *nákvæmni* (precision) þess markara sem leggur markið til og (*1-griphlutfall*) marksins hjá þeim mörkurum sem leggja það ekki til.

Hæsta nákvæmni fékkst með því að nota heildarnákvæmni sem vog. Er það sama niðurstaða og fékkst í Halteren o.fl. (2001) fyrir hollenskan texta. Í töflu 10 eru sýndar niðurstöður kosningar. Þar sést að með því að kjósa milli markara og vega með heildarnákvæmni markaranna fæst **91,54**% nákvæmni. Það er marktækt hærri niðurstaða en fæst með því að nota TnT-markarann eingöngu ($p<0{,}001$). Notaðir eru þrír markarar í íslensku tilrauninni. Allar aðferðir þar sem kosningu er beitt felast því í eftirfarandi: Valið er það mark sem tveir eða fleiri eru sammála um. Ef allir eru ósammála er beitt mismunandi aðferðum við að velja markið. Þegar beitt er meirihlutakosningu er mark valið af handahófi. Þegar vegið er með heildarnákvæmni (accuracy) markarans er valið mark þess markara sem hefur hæsta heildarnákvæmni,

í þessu tilviki mark TnT. Þegar vegið er með nákvæmni hvers marks fyrir hvern markara er valið það mark sem fær hæsta nákvæmni. Þegar vegið er með nákvæmni og griphlutfalli er valið það mark sem fær hæsta summu af nákvæmni þess markara sem leggur markið til og (1-griphlutfall) marksins hjá þeim mörkurum sem leggja það ekki til.

	Óþekkt orð		**Þekkt orð**		**Öll orð**	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Alls	40.392	100,00	549.905	100,00	590.297	100,00
MXPOST	25.246	62,50	500.617	91,04	525.863	89,08
fnTBL	21.823	54,03	502.378	91,36	524.201	88,80
TnT	28.919	71,60	504.484	91,74	533.403	90,36
Meirhlutakosning	27.889	69,05	510.903	92,91	538.792	91,27
Vegið með heildarnákvæmni	29.003	71,80	511.348	92,99	540.351	91,54
Vegið með með nákv. marks	27.808	68,85	511.088	92,94	538.896	91,29
Vegið með nákv. og griphlutfalli.	28.738	71,15	511.440	93,01	540.178	91,51
Vegið með heildarnákvæmni*	34.331	84,97	512.044	93,12	546.375	92,56
* Kosið um mörk þegar viðbótarorðasafn er notað við mörkun með TnT og fnTBL						

Tafla 10. Nákvæmni þriggja markara og nákvæmni sem fæst með þremur mismunandi aðferðum við að kjósa á milli niðurstöðu markaranna

Einnig var gerð tilraun til þess að kjósa um mörk þegar viðbótarorðasafn var notað við mörkun með TnT og fnTBL. Neðsta línan í töflu 10 sýnir niðurstöðu þegar vegið er með heildarnákvæmni TnT en þá fæst **92,56%** nákvæmni. Í töflu 9 sést að nákvæmni þegar markað er með TnT og viðbótarorðasafn notað er 91,54%. Með því að kjósa á milli markaranna fækkar villum um 12% frá þeirri niðurstöðu sem fæst með TnT eingöngu.

Í töflu 11 er niðurbrotinn samanburður á mörkurunum þremur. Þar sést að líklegast er að TnT gefi rétta niðurstöðu ef markararnir gefa ólíkar niðurstöður. Af töflunni sést enn fremur að markararnir TnT og fnTBL eru í einhverjum skilningi líkari heldur en TnT og MXOPOST eða fnTBL og MXPOST. Þess vegna er líklegt að nota megi niðurstöðu MXPOST til þess að bæta niðurstöðu mörkunar.

	Tíðni	**%**	**Safntíðni %**
allir eins og réttir	484.294	82,04	82,04
allir eins og rangir	13.055	2,21	84,25
tnt=fnt=rétt, mxp=rangt	20.783	3,52	87,77
tnt=fnt=rangt, mxp=rétt	8.434	1,43	89,20
tnt=mxp=rétt, fnt=rangt	18.112	3,07	92,27
tnt=mxp=rangt, fnt=rétt	4.850	0,82	93,09
fnt=mxp=rétt, tnt=rangt	12.427	2,11	95,20
fnt=mxp=rangt, tnt=rétt	5.479	0,93	96,13
allir ólíkir, tnt=rétt	4.735	0,80	96,93
allir ólíkir, fnt=rétt	1.847	0,31	97,24
allir ólíkir, mxp=rétt	2.596	0,44	97,68
allir ólíkir og rangir	13.685	2,32	100,00
Samtals	590.297	100,00	

Tafla 11. Samanburður á mörkurum

Lars Borin (2000) hefur rannsakað hvernig megi endurnota efnivið og tungutæknitól, sem þegar eru til, á nýjan hátt. Hann skoðar hvernig megi nota tilbúna markara á efni sem þeir voru ekki þjálfaðir fyrir og þar sem ekki er til reiðu þjálfunarsafn. Borin bendir einnig á hvernig sameina megi niðurstöður markara fyrir þýsku með því að nota málfræðilegar reglur þannig að mörkunarnákvæmni sameinaðra markara verði hærri en nákvæmni þess markara sem nær mestri nákvæmni.

Þó að þessar aðstæður eigi ekki fullkomlega við íslenska verkefnið var aðferðin könnuð nánar.

Tvennt þarf að vera til staðar til þess að unnt sé að bæta nákvæmni með því að sameina niðurstöður tveggja eða fleiri markara.

1. Markararnir gera ekki sömu vitleysurnar, þ.e. þeir bæta hver annan upp (*complementarity*)

2. Mismunur er kerfisbundinn en ekki tilviljunarkenndur

Borin flokkar þá aðferð sem hann leggur til sem „knowlege-rich“, þ.e. rannsakendur þekkja gögnin vel. Málfræðilegar reglur eru skilgreindar til þess að nýta mismun markara til þess að sameina niðurstöður þeirra. Borin setti fram þessar tilgátur:

1. Þegar markararnir eru sammála hafa þeir örugglega rétt fyrir sér.

2. Villur sem markararnir gera eru ólíkar. Í mörgum tilvikum hefur annar markarinn rétt fyrir sér en hinn rangt (Borin skoðaði tvo markara). Mikilvægt er að sá markari sem gefur lægri nákvæmni hafi stundum rétt fyrir sér í slíkum tilvikum.

3. Mismunur á milli markaranna er kerfisbundinn á einhvern hátt. Þennan kerfisbundna mismun má nota til þess að bæta mörkun með því að sameina niðurstöður markaranna.

Fyrsta tilgátan var ekki prófuð. Í töflu 11 sést þó að allir þrír markarar voru sammála og höfðu rétt fyrir sér í 82,04% tilvika og voru allir sammála en höfðu rangt fyrir sér í 2,21% tilvika, þegar prófaðir voru þrír markarar (MXPOST, fnTBL og TnT) í íslensku rannsókninni. Það má því ekki ganga út frá því sem gefnu að niðurstaða sé rétt þó að allir markararnir séu sammála.

Gerð var tilraun til þess að líta á niðurstöðu kosningar sem útkomu úr markara. Athugað var hvort nota mætti niðurstöðu MXPOST, fnTBL eða TnT til þess að bæta þá niðurstöðu. Hæsta nákvæmni, 91,54%, fékkst þar sem kosið var um mörk sem þrír markarar höfðu úthlutað og vegið með heildarnákvæmni þess markara sem hafði staðið sig best, í þessu tilviki TnT. Í töflu 12 sést samanburður á þessari niðurstöðu og niðurstöðum markaranna þriggja.

Á töflunni sést að niðurstöður MXPOST myndu bæta mestu við niðurstöðu með kosningu og gefa 96,37% nákvæmni ef tækist að finna reglur til þess að nýta öll tilvik þar sem MXPOST gefur rétta niðurstöðu en kosning ranga. Með kosningu er þegar búið að nýta kosti TnT og því ekki líklegt að unnt sé að gera betur með þeim markara.

Kannað var hvaða reglum mætti beita til þess að nýta þau tilvik þar sem MXPOST getur gert betur en útkoma úr kosningu gefur. Skoðuð voru tilvik þar sem mark sem kosning gefur er ólíkt marki MXPOST. Fundið var hversu oft MXPOST gefur betri niðurstöðu en kosning í þessum tilvikum.

Kosning vs. MXPOST	**Tíðni**	**%**	**Safntíðni**
Bæði mörk rétt	514.833	87,22	87,22
Kosning rétt, MXPOST rangt	28.509	4,83	92,05
Kosning röng, MXPOST rétt	25.518	4,32	96,37
Mörk lík og röng	11.030	1,87	98,24
Mörk ólík og röng	10.407	1,76	100,00
Kosning vs. fnTBL			
Bæði mörk rétt	517.504	87,67	87,67
Kosning rétt, fnTBL rangt	34.230	5,80	93,47
Kosning röng, fnTBL rétt	22.847	3,87	97,34
Mörk lík og röng	6.697	1,13	98,47
Mörk ólík og röng	9.019	1,53	100,00
Kosning vs. TnT			
Bæði mörk rétt	527.924	89,43	89,43
Kosning rétt, TnT rangt	42.237	7,16	96,59
Kosning röng, TnT rétt	12.427	2,11	98,69
Mörk lík og röng	5.479	0,93	99,62
Mörk ólík og röng	2.230	0,38	100,00

Tafla 12. Samanburður á útkomu markaranna og niðurstöðu kosningar þegar vegið er með heildarnákvæmni

Gert var yfirlit yfir þau tilvik þar sem það að velja mark MXPOST fram yfir útkomu úr kosningu fækkar villum. Flestar villurnar lúta að ruglingi milli falla nafnorða og lýsingarorða. Einnig er þar að finna rugling milli greiningarmynda sagnorða. Ákveðið var að nota útkomu MXPOST fyrir tiltekna samsetningu ef MXPOST gæfi rétta greiningu fram yfir kosningu oftar en 5 sinnum. Reglurnar eru í forminu:

> *ef útkoma úr kosningu er mark1 og útkoma MXPOST er mark2*
> *þá skal velja mark2*

Þegar reglur voru valdar þannig að niðurstaða batnaði um meira en 5 mörk við það að beita reglunni fékkst nákvæmni fyrir öll orð 91,81%, nákvæmni fyrir óþekkt orð 72,13% og fyrir þekkt orð 93,25%.

Einnig var gerð tilraun með að beita reglum þegar upprunaleg mörkun með TnT og fnTBL var gerð með aðstoð orðasafns. Í töflu 10 sést að þegar kosið er um mörk markaranna þriggja sem þannig eru fengin fæst 92,56% nákvæmni. Fundnar voru reglur til þess að velja

mark MXPOST umfram útkomu úr kosningu og fékkst þá **92,69%** nákvæmni.

8.3 Áhrif markaskrár

Skrá yfir alla greiningarstrengi eða mörk sem koma fyrir í tilteknu mörkuðu textasafni er oft kölluð markaskrá (e. *tagset*). Markaskrá Orðtíðnibókarinnar er mjög stór og ítarleg eins og sjá má í viðauka A. Sú greining sem þar er notuð er ekki endilega sú eina rétta og verið getur að sumar tungutæknilausnir geti nýtt sér greiningu sem er ekki jafn ítarleg. Sum tungutækniverkefni gætu þurft mikla nákvæmni í mörkun en ekki mjög ítarlega greiningu.

Prófað var að einfalda greiningarstrengi á þrennan hátt. Einföldunin felst í því að líta aðeins á fyrsta staf í greiningarstreng fyrir atviksorð og samtengingar, þ.e. greina þessa orðflokka ekki í undirflokka, og slá saman fornafnaflokkum en láta greiningu fornafna halda sér að öðru leyti.

	Meðalnákvæmni fnTBL			**Meðalnákvæmni MXPOST**			**Meðalnákvæmni TnT**		
	Rétt (fj.)	**%**	**Safntíðni (%)**	**Rétt (fj.)**	**%**	**Safntíðni (%)**	**Rétt (fj.)**	**%**	**Safntíðni (%)**
Allur greiningarstrengur réttur	524.201	88,80	88,80	525.863	89,08	89,08	533.403	90,36	90,36
Atviksorð ekki greind	5.533	0,94	89,74	6.286	1,06	90,15	6.837	1,16	91,52
Samtengingar ekki greindar	806	0,14	89,88	1.118	0,19	90,34	1.076	0,18	91,70
Öllum fornöfnum slegið saman	600	0,10	89,98	741	0,13	90,46	782	0,13	91,83
Aðeins orðflokkur réttur	42.900	7,27	97,25	40.310	6,83	97,29	37.197	6,30	98,14
Rangur orðflokkur	16.257	2,75	100,00	15.979	2,71	100,00	11.002	1,86	100,00
Samtals	590.297	100,00		590.297	100,00		590.297	100,00	

Tafla 13. Nákvæmni mörkunar þegar markaskrá er einfölduð

Í töflu 13 er sýnd nákvæmni markaranna þegar mörk eru einfölduð á þennan hátt. Af töflunni sést að með því að sleppa greiningu atviksorða hækkar nákvæmni TnT úr 90,36% í 91,52%, villum fækkar um 12%. Með því að sleppa einnig greiningu samtenginga og slá saman fornafnaflokkum fer nákvæmni TnT í 91,83%.

Ef aðeins er litið á greiningu eftir orðflokkum nær TnT **98,14%** nákvæmni. Í sumum tungutækniverkefnum gæti greining eftir orðflokkum dugað og þá gefur TnT viðunandi niðurstöðu.

9 Aðferðirnar prófaðar á nýjum textum

Aðferðirnar við mörkun sem hér hefur verið lýst voru prófaðar á textum sem ekki voru hluti af textasafni Orðtíðnibókarinnar. Fjögur að-

skilin lítil textasöfn voru notuð[4]. Í fyrsta safninu eru brot úr 13 skáldritum frá 19. öld og fyrri hluta 20. aldar, samtals 6.022 lesmálsorð að meðtöldum greinarmerkjum. Í öðru safninu eru brot úr 9 skáldverkum frá því eftir 1980, samtals 3.601 lesmálsorð að meðtöldum greinarmerkjum. Í þriðja safninu eru textar um tölvur og tækni sem eru fengnir úr gagnasafni Morgunblaðsins, úr Fréttabréfi RHÍ og af vefsíðum ýmissa tölvufyrirtækja, samtals 2.926 lesmálorð að meðtöldum greinarmerkjum. Í fjórða safninu eru textar um lögfræði og viðskipti sem eru teknir úr Lagasafni, fréttabréfi fjármálaráðuneytis og Morgunblaðinu (viðskipti), alls 2.776 lesmálsorð að meðtöldum greinarmerkjum. Mörkun var síðan leiðrétt til þess að unnt væri að reikna út nákvæmni mörkunar með hinum ýmsu aðferðum.

Í töflu 14 sjást helstu niðurstöður mörkunar lesmálsorða í þessum textum. Hér kemur í ljós að TnT-markarinn nær bestum árangri. Markararnir MXPOST og fnTBL ná svo lélegum árangri að ekki reyndist unnt að bæta niðurstöðu TnT-markarans með því að nýta niðurstöður frá hinum mörkurunum tveimur. TnT-markarinn nær betri árangri við mörkun bókmenntatextanna heldur en við mörkun texta Orðtíðnibókarinnar sjálfrar en verri árangri við mörkun textanna um tölvur og tækni og viðskipti og lögfræði.

Ekki var notað viðbótarorðasafn þannig að óþekkt orð eru þau orð sem ekki koma fyrir í textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hátt í öllum textunum og hærra en meðalhlutfall í prófunarsöfnum sem gerð voru úr textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hæst í textanum um tölvur og tækni og þar er árangur mörkunar slakastur. TnT-markarinn nær samt alls staðar viðunandi árangri ef aðeins er gerð krafa um réttan orðflokk.

[4] Aðalsteinn Eyþórsson tók saman efni í þessi textasöfn.

	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Gamall bókmenntatexti						
Alls	524	8,70	5.498	91,30	6.022	100,00
MXP	334	63,74	4.935	89,76	5.269	87,50
fnTBL	279	53,24	4.985	90,67	5.264	87,41
TnT	393	75,00	5.209	94,74	5.602	93,03
TnT, einf.	393	75,00	5.218	94,91	5.611	93,18
MXP, orðfl.	458	87,40	5.326	96,87	5.784	96,05
fnTBL, orðfl.	409	78,05	5.374	97,74	5.783	96,03
TnT, orðfl.	472	90,08	5.430	98,76	5.902	98,01
Bókmenntatextar frá því eftir 1980						
Alls	280	7,21	3.601	92,79	3.881	100,00
MXP	182	0,00	3.217	89,34	3.399	87,58
fnTBL	157	56,07	3.262	90,59	3.419	88,10
TnT	221	78,93	3.385	94,00	3.606	92,91
TnT, einf.	221	0,00	3.385	94,00	3.606	92,91
MXP, orðfl.	236	84,29	3.512	97,53	3.748	96,57
fnTBL, orðfl	221	78,93	3.537	98,22	3.758	96,83
TnT, orðfl.	257	91,79	3.561	98,89	3.818	98,38
Textar um tölvur og tækni						
Alls	442	15,11	2484	84,89	2.926	100,00
MXP	186	42,08	2191	88,20	2.377	81,24
fnTBL	169	38,24	2190	88,16	2.359	80,62
TnT	222	50,23	2317	93,28	2.539	86,77
TnT einf.	222	50,23	2317	93,28	2.539	86,77
MXP, orðfl	364	82,35	2410	97,02	2.774	94,81
fnTBL, orðfl.	356	80,54	2437	98,11	2.793	95,45
TnT, orðfl.	395	89,37	2453	98,75	2.848	97,33
Textar um lögfræði og viðskipti						
Alls	390	14,05	2.386	85,95	2.776	100,00
MXP	236	60,51	2.042	85,58	2.278	82,06
fnTBL	213	54,62	2.041	85,54	2.254	81,20
TnT	284	72,82	2.174	91,11	2.458	88,54
TnT einf.	284	72,82	2.176	91,20	2.460	88,62
MXP, orðfl.	348	89,23	2.301	96,44	2.649	95,43
fnTBL, orðfl.	336	86,15	2.309	96,77	2.645	95,28
TnT, orðfl.	366	93,85	2.337	97,95	2.703	97,37

Tafla 14. Nákvæmni við mörkun texta sem eru ekki í textasafni Orðtíðnibókar

10 Niðurstöður og umræða

Hér á undan hefur verið greint frá tilraunum við að marka íslenskan texta með ýmsum aðferðum sem hafa verið þróaðar fyrir önnur tungumál. Fjórir markarar voru þjálfaðir og prófaðir á íslenskum texta og reynt var að finna aðferðir til þess að bæta niðurstöðu markaranna. Gerðar voru tilraunir með að nota orðasafn við mörkun, að kjósa á milli markaranna og að beita málfræðilegum reglum til þess að velja tiltekið mark fram yfir annað mark. Einnig var sýnt að með því að einfalda mörk mætti ná betri niðurstöðu. Það virðist skipta máli í hvaða röð aðgerðunum er beitt. Í töflu 15 er gefið yfirlit yfir helstu niðurstöður af því að sameina aðferðir.

	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	**Tíðni**	**%**	**Tíðni**	**%**	**Tíðni**	**%**
Orðasafn notað við mörkun með fnTBL og TnT[5]						
fnTBL	28.461	70,44	503.142	91,50	531.603	90,06
MXPOST	25.252	62,50	500.611	91,04	525.863	89,08
TnT	34.859	86,28	505.511	91,93	540.370	91,54
Mörk einfölduð[6]						
fnTBL	28.467	70,46	509.788	92,71	538.255	91,18
MXPOST	25.261	62,52	508.747	92,52	534.008	90,46
TnT	34.863	86,29	513.797	93,44	548.660	92,95
Vegið með heildarnákvæmni	34.336	84,98	517.773	94,16	552.109	93,53
MXPOST fram yfir kosn. m. heildarnkv.	34.013	84,18	518.818	94,35	552.831	**93,65**

Tafla 15. Nákvæmni við mörkun íslensks texta þegar fjórum aðgerðum er beitt í röð til þess að bæta niðurstöðu mörkunar þriggja markara

Sýndar eru niðurstöður miðað við að notað sé orðasafnið sem var búið til þegar markað er með TnT og fnTBL. Hæsta nákvæmni, **93,65%**, fæst með því að nota orðasafn, einfalda mörk markaranna, kjósa á milli einfaldaðra marka og beita síðan reglum sem velja mark MXPOST þegar tilteknum skilyrðum er fullnægt. Villum fækkar um 34% miðað við niðurstöðu mörkunar með TnT eingöngu.

Niðurstaða sem fæst með því að nota hjálparorðasafn við mörkun með TnT og fnTBL sýnir að villum mun fækka þegar orðasafn er notað. Það fer að sjálfsögðu eftir eðli textanna sem á að marka og stærð hjálparorðasafnsins hversu mikið nákvæmni eykst við það. Með þeim efnivið sem hér var til ráðstöfunar er þó ljóst að þær aðferðir sem hafa verið prófaðar geta gefið um 92% nákvæmni fyrir texta sem eru líkir textum Orðtíðnibókarinnar.

[5] Orðasafn hefur u.þ.b. helming óþekktra orða

[6] Einföldun felst í að greina ekki atviksorð og ekki heldur samtengingar Fornöfn eru sett í einn flokk en að öðru leyti er greining þeirra eftir kyni, tölu og falli látin haldast.

Þessar niðurstöður benda til þess að nauðsynlegt sé að bæta árangur mörkunar óþekktra orða til þess ná viðunandi árangri í mörkun texta. Ein leið til þess að gera það er að hafa til umráða umfangsmiklar orðaskrár þar sem fram koma beygingarmyndir sem flestra orða, mörk þeirra og hlutfallsleg tíðni einstakra greiningarmynda. Nota má *Beygingarlýsingu íslensks nútímamáls* (Kristín Bjarnadóttir 2004), sem einnig var gerð var fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins, sem efnivið í slíka orðaskrá. Einnig er nauðsynlegt að hafa tiltækar skrár með ýmiss konar sérnöfnum svo sem mannanöfnum, nöfnum fyrirtækja og stofnana og örnefnum. Einnig væri æskilegt að kanna frekar hvers konar markaskrá sé heppileg fyrir hin ýmsu verkefni.[7]

Aðferðirnar voru einnig prófaðar á textum sem voru ekki hluti af textasafni Orðtíðnibókarinnar. Þá kom í ljós að TnT-markarinn nær bestum árangri við mörkun allra textanna. Aðrir markarar náðu svo lélegum árangri að ekki reyndist unnt að bæta niðurstöðu mörkunar með því að nýta þá.

Heimildir

Borin, Lars. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. *Second International Conference on Language Resources and Evaluation*, Athens 31 May – 2 June, 2000, bls. 21–26.

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

Brants, Thorsten. 2000b. TnT - A Statistical Part-of-Speech Tagger. Version 2.2. *http://www.coli.uni-sb.de/~thorsten/tnt/*

Brill, Eric. 1994. Some Advances in Rule-Based Part of Speech Tagging. *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, bls. 722–727. Seattle, Washington.

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, December 1995: 543–563.

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*

[7] Eftir að þessu verki lauk formlega bjó Hrafn Loftsson (2006) til málfræðilegan reglumarkara 2004–2005 og notaði texta *Íslenskrar orðtíðnibókar* við prófun. Hrafn náði 91,471% nákvæmni í mörkun með reglumarkara sínum (*IceTagger*). Með því að sameina niðurstöður fjögurra markara náði Hrafn 92,94% mörkunarnákvæmni.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02, http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf*

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*

Friðrik Magnússon. 1988. Hvað er títt? Tíðnikönnun Orðabókar Háskólans. *Orð og tunga* 1:1–49.

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.

Hrafn Loftsson. 2006. Tagging Icelandic text: A linguistic rule-based approach. Technical Report CS-06-04, Department of Computer Science, University of Sheffield.

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

Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. *Samspil tungu og tækni*. Menntamálaráðuneytið, Reykjavík.

Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, Massachusetts. London, England.

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

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.

Ratnaparkhi, A. 1997. A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical Report 97-08, Institute for Research in Cognitive Scince, University of Pennsylvania.

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

Samuelsson, Christer. 1993. Morphological tagging based entirely on Bayesian inference. *9th Nordic Conference on Computational Linguistics NODALIDA-93,* bls. 225–238. Stockholm University, Stockholm, Sweden.

Sigrún Helgadóttir. 2002. 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.

Sigrún Helgadóttir and Örvar Kárason. 2005. Memory-Based Learning Assignment. Term paper in Machine Learning, GSLT.

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

Lykilorð:

mark, mörkun, markari

Keywords:

part-of-speech tag, tagging, tagger

Abstract

This paper gives the results on the automatic tagging of Icelandic text, using a corpus that was prepared for the making of the *Icelandic Frequency Dictionary*. The corpus contains 590,297 running words with 59,358 word forms, including punctuation. Each running word has been supplied with a morphosyntactic tag and the tagset contains 639 tags, including punctuation tags. Five different data-driven taggers, fnTBL, TnT, MXPOST, μ-TBL and MBT were trained on the corpus by using ten-fold cross-validation. The TnT tagger obtained best results for tagging or 90.36% accuracy. The TnT and fnTBL systems allow the use of a backup lexicon. When using such a lexicon TnT reached 91.54% tagging accuracy and fnTBL 90.06%. Methods for combining the results of the taggers were also tested. A voting method where each tagger votes its overall precision gave best result of the voting methods tested or 91.54% accuracy. By utilizing the ability of the MXPOST tagger to distinguish between noun cases, rules were composed to increase tagging accuracy to 91.81%. By using a special strategy for simplifying tags, the TnT tagger gave 91.83% tagging accuracy. Finally, the different strategies for improving tagging accuracy were applied in a certain order. The best result, 93.65% accuracy, was obtained by tagging with a backup lexicon with fnTBL and TnT, simplifying the resulting tags, voting between the simplified tags and applying rules based on MXPOST. Compared with the result obtained with TnT alone, the number of errors is reduced by 34%. By using a lexicon derived from the Morphological Description of Modern Icelandic as a backup lexicon the accuracy can be further increased. Finally an experiment was made in tagging texts that are not a part of the corpus of the *Icelandic Frequency Dictionary*.

Sigrún Helgadóttir
Stofnun Árna Magnússonar í íslenskum fræðum
Neshaga 16
IS-107 Reykjavík
sigrunh@lexis.hi.is

Viðauki A

Skýring skammstafana í greiningarstrengjum Íslenskrar orðtíðnibókar

Dálkur	Formdeild	Greiningartákn-greiningaratriði
1	Orðflokkur	N-nafnorð
2	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn, X-ókyngreint
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
5	Greinir	G-með viðskeyttum greini
6	Sérnöfn	M-mannsnafn, Ö-örnefni, S-önnur sérnöfn
1	Orðflokkur	L-lýsingarorð
2	Stig	F-frumstig, M-miðstig, E-efstastig
3	Beyging	S-sterk beyging, V-veik beyging, O-óbeygt
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	Tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	F-fornafn
2	Flokkur	A-ábendingarfornafn, B-óákveðið ábendingarfornafn, E-eignarfornafn, O-óákveðið fornafn, P-persónufornafn, S-spurnarfornafn, T-tilvísunarfornafn
3	Kyn/Persóna	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	G-greinir
2	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	T-töluorð
2	Flokkur	F-frumtala
3	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	S-sögn (þó ekki lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	N-nafnh., B-boðh., F-framsöguh., V-viðtengingarh., S-sagnbót, L-lýsingarh. nútíðar
4	Tíð	N-nútíð, Þ-þátíð
5	Tala	E-eintala, F-fleirtala
6	Persóna	1-1. persóna, 2-2. persóna, 3-3. persóna
1	Orðflokkur	S-sögn (lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	Þ-lýsingarháttur þátíðar
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	Tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall

1	Orðflokkur	A-atviksorð
2	Stig	M-miðstig, E-efsta stig
3	Flokkur/-Fallstjórn	A-stýrir ekki falli, U-upphrópun/ O-stýrir þolfalli, Þ-stýrir þágufalli, E-stýrir eignarfalli
1	Orðflokkur	C-samtenging
2	Flokkur	N-nafnháttarmerki, T-tilvísunartenging
1	Flokkur	E-erlent orð
1		X-ógreint orð