Málgreiningarhópur
Orðabók Háskólans

Sigrún Helgadóttir

Gerð markara fyrir íslenskan texta

Febrúar 2004

Útgáfa 1.0

Endurskoðuð gerð febrúar 2006

0. Útdráttur

Í þessari skýrslu er greint frá niðurstöðum tilrauna við að marka íslenskan texta vélrænt. Til er textasafn þar sem hverju orði fylgir greiningarstrengur sem segir til um orðflokk orðsins og beygingarmynd þess. Þetta textasafn varð til við undirbúning Íslenskrar orðtíðnibókar. Í textasafninu eru 590.297 lesmálsorð sem birtast í 59.358 mismunandi orðmyndum að meðtöldum greinarmerkjum. Lesmálsorðunum fylgja 639 mismunandi greiningarstrengir að meðtöldum greinarmerkjum. Textasafnið var notað til þess að kenna þremur forritum, fnTBL, TnT og MXPOST, að greina íslenskan texta á sama hátt. Úr textasafninu voru búin til 10 pör þjálfunar- og prófunarsafna þar sem prófunarsöfnin eru óháð en þjálfunarsöfnin skarast um 80%. Forritin voru prófuð á þessum efnivið. Bestur árangur náðist með TnT-forritinu eða 90,36% nákvæmni að meðaltali. Prófað var að einfalda greiningarstrengi þannig að aðeins var greindur orðflokkur atviksorða og samtenginga og fornafnaflokkum var slegið saman. Þá fékkst 91,83% nákvæmni með TnT-forritinu. Síðan voru prófaðar aðferðir til þessa að kjósa á milli greiningarstrengja sem kerfin þrjú úthlutuðu lesmálsorðum. Þegar vegið er með heildarnákvæmni forritanna og kosið um einfaldaða greiningarstrengi fékkst 92,56% nákvæmni. Nákvæmni hækkaði í 92,82% þegar kostir MXPOST-forritsins við að greina á milli falla nafnorða voru nýttir. Einnig voru gerðar tilraunir með að nota hjálparorðasafn. Búið var til orðasafn sem átti að geyma u.þ.b. helming óþekktra orða í hverju prófunarsafni með tilliti til samsvarandi þjálfunarsafns. TnT- forritið og fnTBL-forritið gefa kost á að nota slíkt hjálparorðasafn. Þegar markað var með aðstoð slíks orðasafns og síðan beitt sömu aðgerðum og taldar voru upp hér að framan fékkst **93,65%** nákvæmni. Líklegt er að unnt sé að ná meiri nákvæmni með því að nýta beygingarlýsingu sem gerð hefur verið á Orðabók Háskólans. Einnig voru gerðar tilraunir við að marka ólíka texta.

1. Inngangur

Í apríl 2002 veitti menntamálaráðuneytið styrk að upphæð 6 milljónir króna til þess að gera málfræðilegan markara fyrir íslensku. Um styrkinn sóttu Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir) og Orðabók Háskólans. Gerður var samningur milli ráðuneytisins og verktakanna dags. 18. október 2002. Þessi skýrsla er lokaskýrsla um verkefnið. Í skýrslunni er greint frá verkaskiptingu verktaka, aðferðum sem beitt var og lokaniður-stöðu verkefnisins. Enn fremur er gerð tillaga um framhald verkefnisins.

Í ýmsum tungutækniverkefnum þar sem unnið er úr texta er ávinningur af því að orð í textanum séu greind í orðflokka og eftir beygingu. Má þar m.a. nefna greiningu texta í setningahluta (e. *partial parsing*), nám orða úr texta fyrir gerð orðasafns (e. *lexical acquisition*), upplýsingaheimt, talkennsl, talgervingu, vélrænar þýðingar, orðabókargerð, fyrirspurnarkerfi og leiðréttingaforrit. Einnig er nauðsynlegt að orð í texta séu greind eftir orðflokkum og beygingu ef gera á tíðnikönnun á texta eins og birt er í Orðtíðnibókinni (Megyesi 2002:17).

Handvirk greining texta eftir orðflokkum og beygingu er mjög seinvirk og frekar leiðinleg iðja. Þess vegna hefur lengi verið fengist við að beita vélrænum aðferðum við þetta starf. Þetta svið hefur því fengið mikla umfjöllun hjá þeim sem vinna við máltækni.

Það að greina orð eftir orðflokkum og beygingu er kallað „part-of-speech tagging“ eða aðeins „tagging“ í ensku og greiningarstrengurinn kallast „pos tag“ eða aðeins

„tag“. Forrit eða kerfi sem framkvæmir þetta verk kallast á ensku „tagger“. Lagt er til að greiningarstrengurinn kallist „mark“ á íslensku, aðgerðin verði kölluð „mörkun“ og forritið eða kerfið kallist „markari“.

Vélrænar aðferðir við mörkun eru venjulega flokkaðar í tvo flokka, reglubyggðar aðferðir (e. *rule-based methods*) og gagnaaðferðir (e. *data-driven methods*). Fyrstu vélrænu aðferðirnar sem var beitt voru reglubyggðar aðferðir. Orðasafn var notað til þess að úthluta orðum í texta hugsanlegum greiningarstrengjum. Síðan voru notaðar málfræðilegar reglur til þess að skera úr um hvaða greiningarstrengur væri réttur. Þessar reglur voru venjulega handskrifaðar og byggðar á málfræði hvers tungumáls.

Gagnaaðferðir byggjast á fyrir fram greindu textasafni. Forrit er látið búa til líkan á grundvelli gagna sem þegar hafa verið greind. Gagnaaðferðum má skipta í nokkra flokka. Má þar m.a. nefna tölfærðilegar aðferðir og leiðréttingaaðferð(e. *transformation-based learning*) sem var prófuð á íslenskum texta og er lýst í umsókn um styrk til verkefnisins og byggist á því að skipta um greiningarstreng þegar ákveðnum skilyrðum í umhverfi orðsins er fullnægt. Einnig má nefna svokallaða minnisaðferð.

Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni.

2. Verkaskipting

Verktakar skiptu þannig með sér verkum að Orðabók Háskólans tók að sér bókhalds-umsjón, lét í té skrifstofuaðstöðu og aðgang að tölvukerfi og veitti enn fremur aðgang að þeim gögnum sem voru notuð. Málgreiningarhópurinn vann verkið að öðru leyti. Eiríkur Rögnvaldsson var verkefnisstjóri. Kristín Bjarnadóttir og Auður Þórunn Rögnvaldsdóttir sáu um undirbúning á tölvuskrám Orðtíðnibókar fyrir frekari vinnslu. Sigrún Helgadóttir prófaði þá markara sem voru notaðir í tilrauninni, gerði ýmsa útreikninga og skrifaði lokaskýrslu. Vinna við verkefnið hófst haustið 2002 og lýkur með ritun þessarar skýrslu í febrúar 2004. Málgreiningarhópurinn hélt reglulega fundi á meðan unnið var að verkefninu.

3. Forsaga

Haustið 2001 sóttu félagar í Málgreiningarhópnum fjarnámskeið í málgreiningu (*Natural Language Processing*, NLP) við háskólann í Gautaborg. Sem verkefni í því námskeiði var gerð tilraun með að beita mörkunarkerfi sem kallast **µ-TBL** og er samið af Torbjörn Lager (Lager 1999) á íslenskan texta. Við tilraunina var notað textasafn sem er um einn tíundi af textasafni Orðtíðnibókarinnar (textabrot úr barnabókum). Í umsókn um styrk til verkefnisins er gerð ítarleg grein fyrir þessari tilraun og þeim aðferðum sem kerfið beitir. Niðurstaðan var það góð að ástæða þótti til þess að gera frekari tilraunir til þess að marka íslenskan texta með vélrænum aðferðum (Auður Þórunn Rögnvaldsdóttir 2002), (Eiríkur Rögnvaldsson 2002), (Kristín Bjarnadóttir 2002), (Sigrún Helgadóttir 2002a). Sigrún Helgadóttir prófaði einnig tölfræðilegan markara, **TnT**, á sama efni (Sigrún Helgadóttir 2002b). Á grundvelli þessara prófana var ákveðið að sækja um styrk til þess að prófa þessar aðferðir og sambærilegar aðferðir við að marka íslenskan texta.

4. Efniviður

Í þessu verkefni var notað textasafn sem var búið til fyrir vinnslu *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991) sem Orðabók

Háskólans gaf út 1991. Vinna við bókina hófst 1985. Textasafninu er lýst nákvæmlega í formála Orðtíðnibókarinnar en helstu atriða verður getið hér.

Í textasafninu eru 100 textar, hver með um 5.000 lesmálsorðum. Flestir textarnir eru hluti af stærra ritverki. Textarnir voru valdir úr ritverkum sem voru gefin út á tímabilinu 1980–1989. Textarnir voru valdir úr 5 textaflokkum, tuttugu textar úr hverjum flokki. Flokkarnir eru þessir:

1. Íslensk skáldverk.
2. Þýdd skáldverk.
3. Ævisögur og minningar.
4. Fræðslutextar (10 á sviði hugvísinda, 10 á sviði raunvísinda).
5. Barna- og unglingabækur (10 frumsamdir textar, 10 þýddir textar).

Hver höfundur eða þýðandi má hvorki vera höfundur né þýðandi annars texta í safninu. Textarnir er í mörgum tilvikum teknir úr upphafi meginmáls. Fyrirsögnum og kaflanúmerum í upphafi texta er sleppt en ekki inni í textum. Myndum og öðru sundurlausu efni er sleppt. Texti endar alltaf á heilli setningu.

Allir textarnir voru vandlega lesnir og augljósar ritvillur leiðréttar, en óvenjulegum rithætti var haldið. Fyrir greiningu fyrir Orðtíðnibókina voru greinarmerki fjarlægð. Þeim var síðan bætt inn í textaskrárnar aftur vegna mörkunartilraunar. Hástaf var breytt í lágstaf í upphafi setningar nema þar sem setning hófst á sérnafni. Þessi breyting var nýtt við mörkunartilraunirnar. Ritun orða sem voru rituð með bókstafnum *z* var breytt í samræmi við núgildandi stafsetningarreglur. Einnig var bókstafurinn *t* felldur niður í fjórum tilvikum um leið og nýjum stafsetningarreglum var beitt.

Í formála Orðtíðnibókarinnar er hugtakið *lesmálsorð* skilgreint. Þar segir: „Hugtakið *lesmálsorð* nær til samfelldrar raðar af bókstöfum og/eða tölustöfum og táknum sem aðgreind eru með stafbili eða greinarmerkjum.“ Oftast er augljóst hvernig lesmálsorð eru aðgreind en það getur þó reynst flókið þar sem tölur og ýmis tákn koma fyrir. Við aðgreiningu lesmálsorða Orðtíðnibókarinnar var notuð 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. Skammstafanir eru í flestum tilvikum greindar eins og lesið er úr þeim.

Hverju lesmálsorði var síðan komið fyrir í sérstakri línu. Í þeirri línu var einnig komið fyrir *greiningarstreng* orðsins og flettimynd/nefnimynd (e. *lemma*) þess. Greiningarstrengurinn er runa bókstafa (og tölustafa) sem segja til um orðflokk og ýmiss konar málfræðilega greiningu lesmálsorðsins. Greiningarstrengurinn er kallaður *mark* í þeirri könnun sem hér verður fjallað um. Auð lína var höfð á milli setninga. Þessi lína var fjarlægð á einhverju stigi við úrvinnslu fyrir orðtíðnibókina en þurfti síðan að bæta við aftur seinna. Til þess var þá skrifað sérstakt forrit.

Texti Orðtíðnibókarinnar var markaður í þremur atrennum. Í fyrstu var byggt á greiningu 54.000 lesmálsorða sem höfðu verið greind handvirkt og notuð við orðtíðnikönnun (Friðrik Magnússon 1988). Búið var til forrit byggt á mörkun þessa texta sem notaði beygingarfræðilegar upplýsingar, málfræðireglur og tölfræðilegar aðferðir. Þetta forrit var notað til þess að marka lesmálsorð í fyrstu 50 textunum. Greiningin var síðan leiðrétt og forritið endurbætt. Síðustu 50 textarnir voru síðan markaðir með endurbættu forrit og niðurstaðan leiðrétt. Höfundar telja að um 80%

lesmálsorða hafi fengið rétta greiningu að öllu leyti. 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. Nákvæm lýsing á greiningunni er að öðru leyti í formála Orðtíðnibókarinnar. Í viðauka B er yfirlit yfir greiningarstrengi sem notaðir voru.

5. Aðferðir og markarar

Þær aðferðir sem verða prófaðar eru allar dæmi um gagnaaðferðir, þ.e. aðferðir þar sem reynt er að læra af fyrir fram greindum gögnum (e. *data-driven methods*). Aðferð þeirri sem um ræðir er beitt með forriti sem 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.

Ákveðið var að prófa 4 gagnaaðferðir. Alls voru prófaðir 5 markarar sem unnt er að þjálfa á íslenskum texta og eru fáanlegir 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 aðferð sem mætti kalla leiðréttingaaðferð (e. *error-driven transformation-based learning*), **µ-TBL** og **fnTBL.** Einnig var prófaður einn markari, **MBT**, sem byggist á minnistækni (e. *memory-based technique*). Alla þessa markara má kalla á íslensku gagnamarkara eða námfúsa markara.

Prófunum er nú lokið. Markarinn µ-TBL hafði áður verið prófaður á litlum úrdrætti úr textasafni Orðtíðnibókarinnar og fékkst þá viðunandi niðurstaða (Sigrún Helgadóttir 2002a). En markarinn virtist ekki ráða við allt textasafn Orðtíðnibókarinnar. Mjög léleg niðurstaða fékkst með MBT-markaranum. Prófuð var ný útgáfa af markaranum í nóvember 2005. Í viðauka C er greint frá niðurstöðu þeirrar tilraunar.

Í þessum kafla verður lauslega lýst þeim þremur aðferðum við mörkun sem voru notaðar í tilrauninni og þeim mörkurum sem voru valdir til prófunar á íslenskum texta.

5.1 Falin Markovslíkön

Í þessum flokki var valinn markarinn TRIGRAMS'N'TAGS (TnT) sem Thorsten Brants samdi (Brants 2000a). 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 yfir 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 mörkuðu textasafni, þjálfunarsafni. 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.

5.2 Hámarksóreiðuaðferð

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.

Til þess að prófa þessa aðferð á íslenskum texta var notað forritið MXPOST eftir Ratnaparkhi (Ratnaparkhi 1996). 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 (e. *beam search*) 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.

5.3 Leiðréttingaaðferð

Brill (Brill 1994 og Brill 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 (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.

Nokkur forrit bjóðast sem nota aðferðir Brills. Valið var að nota forritið fnTBL (*Fast Transformation-Based Learning Toolkit*) eftir Radu Florian og Grace Ngai (Florian og Ngai 2002).

6. 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.

6.1 Skrár

Tölvuskrár Orðtíðnibókarinnar voru skipulagðar þannig að í hverri skrá er textabútur úr einni heimild. Hverri skrá var skipt í 10 nokkurn veginn jafna búta. Úr þessum 10 bútum voru búin til 10 pör af skrám þannig að skrárnar í hverju pari skarast ekki. Í hverju pari er ein skrá með um 90% af lesmálsorðum úr textasafninu og önnur með um 10% af lesmálsorðum úr textasafninu. Stærri skráin er notuð sem þjálfunarsafn og sú minni sem prófunarsafn í tilraununum. Í hverju pari eru því textar sem eiga að vera dæmigerðir fyrir alla textaflokka í textasafninu. Prófunarsöfnin 10 eru óháð hvert öðru en þjálfunarsöfnin hafa um 80% sameiginlega texta. Allir markarar í tilrauninni voru prófaðir á öllum 10 pörum og fundin meðalnákvæmni mörkunar (e. *ten-fold cross-validation*).

Úr skrám Orðtíðnibókarinnar var unnt að velja lesmálsorðin þannig að öll orð hefjist á lágstaf nema sérnöfn. Skil á milli setninga höfðu glatast í vinnslu skránna en voru sett inn aftur með sérstöku forriti fyrir tilraunir með fnTBL- og MXPOST-kerfunum.

Í töflu 1 sést yfirlit yfir efniviðinn.

Tafla 1. Yfirlit yfir 10 pör þjálfunar- og prófunarsafna

	Þjálfunarsafn			Prófunarsafn		
	Lesmálsorð	Orðmyndir	Mörk*	Lesmálsorð	Orðmyndir	Mörk*
01	531.128	55.188	639	59.169	13.279	552
02	531.330	55.622	636	58.967	12.618	562
03	531.220	55.586	637	59.077	12.550	548
04	531.230	55.700	637	59.067	12.433	548
05	531.222	55.774	634	59.075	12.381	555
06	531.161	55.643	636	59.136	12.622	564
07	531.188	55.697	635	59.109	12.572	567
08	531.316	55.577	638	58.981	12.804	549
09	531.154	55.657	639	59.143	12.697	547
10	531.724	55.722	635	58.573	12.326	553
Allt safnið	590.297	59.358	639			

* Að meðtöldum greinarmerkjum

6.2 Grunnmörkun

Til þess að geta metið árangur markara þarf að hafa einhverja viðmiðun. Venjulega er útkoma markara borin saman við bæði efri og neðri mörk. Efri mörk fást með því að handmarka sama texta. Nákvæmni markarans er þá hlutfall þeirra orða sem markarinn markar eins og gert var í handmörkun. En handmörkun getur verið misjafnlega nákvæm. Jurafsky og Martin (2000:308) segja frá nokkrum rannsóknum þar sem handmörkun er könnuð. Einn rannsakandi (Marcus et al. 1993) fundu t.d. að þeir sem unnu við mörkun Penn Treebank útgáfu Brown-textasafnsins voru sammála í 96–97% tilvika um mörk. Fyrir það textasafn er því ekki mögulegt að ná 100% nákvæmni. Annar rannsakandi hefur sýnt fram á að handmörkun getur orðið 100% nákvæm ef tveir menn mega bera sig saman (Voutilainen 1995, p. 174). Almennt er álitið að nákvæmni mörkunar textasafns Orðtíðnibókarinnar sé nálægt 100%. Það hefur þó ekki verið rannsakað sérstaklega en tveir einstaklingar mörkuðu safnið og hafa efalaust borið sig saman.

Einnig er nauðsynlegt hafa viðmiðun um lægstu nákvæmni sem unnt er að sætta sig við af hálfu markara. Þessi lægsta nákvæmni er venjulega kölluð *grunnmörkun*. Hér er stuðst við skilgreiningar á grunnmörkun sem koma fram í doktorsritgerð Beötu Megyesi (Megyesi 2002:55). Grunnmörkun er fundin fyrir hvert par af þjálfunar- og prófunarsöfnum sem lýst var að framan. Fyrir hvert par er búið til orðasafn úr þjálfunarsafninu þar sem fram kemur hvert orð og algengasta mark orðsins í því safni. Síðan er athugað hvaða orð í prófunarsafninu koma fyrir í þjálfunarsafninu. Þeim orðum sem koma fyrir í þjálfunarsafninu er gefið algengasta mark sem fannst þar. Óþekkt orð fá þrenns konar mismunandi greiningu:

1. Óþekkt orð eru álitin ranglega greind.
2. Óþekkt orð fá það mark sem kemur oftast fyrir í þjálfunarsafninu, hér *aa*.
3. 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. Í töflu 2 er sýnd niðurstaða grunnmörkunar fyrir öll 10 pörin og hlutfall óþekktra orða fyrir hvert prófunarsafn.

Tafla 2. Grunnmörkun

			Grunnmörkun		
Par	Fjöldi lesmálsorða	Hlutf. óþ. orða (%)	Óþ. orð greind röng (%)	Óþ. orð greind aa (%)	Óþ. orð greind nken, nken-m (%)
01	59.169	7,57	75,43	75,50	75,81
02	58.967	6,79	76,17	76,24	76,43
03	59.077	6,88	76,36	76,43	76,66
04	59.067	6,69	76,85	76,91	77,18
05	59.075	6,51	76,93	76,98	77,24
06	59.136	6,70	76,40	76,46	76,69
07	59.109	6,70	76,81	76,87	77,12
08	58.981	6,93	75,83	75,91	76,11
09	59.143	6,80	76,03	76,07	76,42
10	58.573	6,83	76,29	76,36	76,66
Samtals	590.297	6,84	76,31	76,37	76,63

Af töflunni sést að meðalhlutfall óþekktra orða í öllu safninu er 6,84%.

Meðalnákvæmni fyrir öll prófunarsöfnin er:

1. Óþekkt orð álitin ranglega greind: **76,31%** orða rétt greind.
2. Óþekkt orð fá markið *aa*: **76,37%** orða rétt greind.
3. Óþekkt orð rituð með litlum staf fá algengustu greiningu nafnorða (*nken*) og óþekkt orð rituð með upphafsstaf fá algengustu greiningu sérnafna (*nken-m*): **76,63%** orða rétt greind.

Markari sem nær ekki þessari nákvæmni bætir því engu við það sem fæst með grunnmörkun eingöngu.

6.3 Mælikvarðar fyrir nákvæmni

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 í handmörkun. Hittni markarans fyrir öll orð er því 53.101/59.169 eða 89,74%.

Síðan 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 (Manning og Schütze 1999: 269) skilgreina F-gildi sem

$F=1/(\alpha*(1/P)+(1-\alpha)*(1/R))$

og Megyesi (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:

$F\text{-}gildi=P*R/((R+P)/2)=2*P*R/(P+R)$

Í Manning og Schütze (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 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 (97,74%)

R=tp/(tp+fn)=tp/(það sem átti að velja)=11.451/11.660 (98,21%)

F=2*P*R/(P+R)=2*97,74*98,21/(97,74+98,21)= 97,97%.

Þessar stærðir má reikna fyrir hvaða greiningarstreng sem er.

6.4 Tilraunir

Hver markari var prófaður á öllum tíu pörunum sem höfðu verið búin til. MXPOST-markarinn gefur ekki kost á að breyta sjálfgefnum gildum í vinnslu. TnT-markarinn var notaður með sjálfgefnum gildum. Skilgreina þarf sniðmát fyrir fnTBL-markarann. Eftir nokkrar tilraunir var ákveðið að nota sniðmát sem forritið lét í té fyrir samhengi orða og marka. Þetta eru alls 40 sniðmát en Brill notaði 26 sniðmát í tilraunum sínum. Fyrir óþekkt orð voru notuð sniðmát Brills að viðbættum sniðmátum til þess að finna 5-stafa viðskeyti og forskeyti. Öll þessi sniðmát eru sýnd í Viðauka A. Óþekkt orð rituð með lágstaf fengu í upphafi algengasta mark nafnorða, nken, og óþekkt orð rituð með hástaf fengu algengasta mark sérnafna, nken-m.

Allir markararnir sem voru valdir voru þjálfaðir á þjálfunarsöfnunum 10 og prófaðir á samsvarandi prófunarsöfnum. Niðurstöður eru sýndar í töflu 3. Af niðurstöðunum í töflunni má draga ýmsar ályktanir.

Í fyrsta lagi er ljóst að TnT-markarinn nær bestum árangri, þá MXPOST-markarinn og fnTBL-markarinn nær lélegustum árangri. Með því að bera saman töflur 2 og 3 sést að frammistaða markaranna er háð hlutfalli óþekktra orða í texta. Besta niðurstaðan, 90,74% fyrir öll orð, fæst með TnT-markaranum fyrir par 05 þar sem hlutfall óþekktra orða í prófunarsafni miðað við þjálfunarsafn er lægst eða 6,51%.

Versta niðurstaðan fæst með fnTBL-markaranum fyrir par 01 þar sem hlutfall óþekktra orða er hæst eða 7,57%. Heildarniðurstaða markaranna er einnig háð því hversu vel þeim tekst að greina óþekkt orð. TnT-markarinn stendur sig áberandi best við greiningu óþekktra orða. Sá markari nær einnig hæstri meðalnákvæmni við greiningu þekktra orða og þar af leiðandi bestri heildarnákvæmni. MXPOST-markarinn fær betri heildarniðurstöðu en fnTBL-markarinn þar sem sá markari nær betri árangri við mörkun óþekktra orða þó að fnTBL-markarinn standi sig betur við mörkun þekktra orða.

Mismunur á hæstu og lægstu mörkunarnákvæmni fyrir öll orð er 2,41 prósentustig. En við það að nákvæmni hækkar úr 88,32% í 90,74% fækkar villum úr 6.910 í 5.474 eða um 21%.

Tafla 3. Niðurstaða af þjálfun og mörkun 10 para skráa

		Lesmálsorð			Rétt greind lesmálsorð			Nákvæmni		
		Öll	Þekkt	Óþekkt	Öll	Þekkt	Óþekkt	Öll orð	Þekkt orð	Óþekkt orð
Markari	Par							%	%	%
fnTBL	01	59.169	54.687	4.482	52.259	49.821	2.438	88,32	91,10	54,40
	02	58.967	54.961	4.006	52.266	50.137	2.129	88,64	91,22	53,15
	03	59.077	55.014	4.063	52.455	50.290	2.165	88,79	91,41	53,29
	04	59.067	55.113	3.954	52.542	50.474	2.068	88,95	91,58	52,30
	05	59.075	55.227	3.848	52.850	50.732	2.118	89,46	91,86	55,04
	06	59.136	55.172	3.964	52.437	50.268	2.169	88,67	91,11	54,72
	07	59.109	55.148	3.961	52.568	50.485	2.083	88,93	91,54	52,59
	08	58.981	54.891	4.090	52.325	50.060	2.265	88,72	91,20	55,38
	09	59.143	55.122	4.021	52.380	50.246	2.134	88,57	91,15	53,07
	10	58.573	54.570	4.003	52.119	49.865	2.254	88,98	91,38	56,31
Meðal nákvæmni								**88,80**	**91,36**	**54,02**
MXP	01	59.169	54.687	4.482	52.301	49.512	2.789	88,39	90,54	62,23
	02	58.967	54.961	4.006	52.442	50.015	2.427	88,93	91,00	60,58
	03	59.077	55.014	4.063	52.764	50.212	2.552	89,31	91,27	62,81
	04	59.067	55.113	3.954	52.719	50.190	2.529	89,25	91,07	63,96
	05	59.075	55.227	3.848	52.857	50.428	2.429	89,47	91,31	63,12
	06	59.136	55.172	3.964	52.578	50.094	2.484	88,91	90,80	62,66
	07	59.109	55.148	3.961	52.805	50.348	2.457	89,33	91,30	62,03
	08	58.981	54.891	4.090	52.461	49.883	2.578	88,95	90,88	63,03
	09	59.143	55.122	4.021	52.520	50.097	2.423	88,80	90,88	60,26
	10	58.573	54.570	4.003	52.416	49.838	2.578	89,49	91,33	64,40
Meðal nákvæmni								**89,08**	**91,04**	**62,51**
TnT	01	59.169	54.687	4.482	53.141	50.008	3.133	89,81	91,44	69,90
	02	58.967	54.961	4.006	53.191	50.364	2.827	90,20	91,64	70,57
	03	59.077	55.014	4.063	53.390	50.468	2.922	90,37	91,74	71,92
	04	59.067	55.113	3.954	53.511	50.623	2.888	90,59	91,85	73,04
	05	59.075	55.227	3.848	53.602	50.814	2.788	90,74	92,01	72,45
	06	59.136	55.172	3.964	53.440	50.599	2.841	90,37	91,71	71,67
	07	59.109	55.148	3.961	53.576	50.751	2.825	90,64	92,03	71,32
	08	58.981	54.891	4.090	53.225	50.268	2.957	90,24	91,58	72,30
	09	59.143	55.122	4.021	53.290	50.448	2.842	90,10	91,52	70,68
	10	58.573	54.570	4.003	53.037	50.141	2.896	90,55	91,88	72,35
Meðal nákvæmni								**90,36**	**91,74**	**71,62**

Dreifing orðflokka er ólík meðal óþekktra orða og allra orða. Orð í svokölluðum opnum orðflokkum, þ.e. nafnorð, lýsingarorð og sagnir, eru að meðaltali um 44,3% af öllum orðum í prófunarsöfnunum en um 95,9% að meðaltali af óþekktum orðum.

Í töflu 4 er sýnt hvernig lesmálsorð skiptast eftir orðflokkum í öllu safninu og meðal óþekktra orða. Óþekkt orð eru orð í hverju prófunarsafni sem ekki koma fyrir í samsvarandi þjálfunarsafni.

Tafla 4. Dreifing lesmálsorða eftir orðflokkum

	Allt safnið		Óþekkt orð	
Orðfl.	Tíðni	%	Tíðni	%
-	2.256	0,38	0	0,00
!	883	0,15	0	0,00
(	629	0,11	0	0,00
)	636	0,11	0	0,00
,	22.083	3,74	0	0,00
.	33.637	5,70	0	0,00
/	5	0,00	5	0,01
:	1.137	0,19	0	0,00
;	534	0,09	0	0,00
?	2.103	0,36	0	0,00
[	3	0,00	0	0,00
]	3	0,00	0	0,00
_	44	0,01	0	0,00
«	3.597	0,61	0	0,00
»	3.567	0,60	0	0,00
a (atviksorð)	116.112	19,67	516	1,28
c (samtengingar)	60.256	10,21	5	0,01
e (erlend orð)	411	0,07	275	0,68
f (fornöfn)	74.315	12,59	48	0,12
g (greinir)	632	0,11	0	0,00
l (lýsingarorð)	35.669	6,04	7.322	18,13
n (nafnorð)	122.621	20,77	27.559	68,23
s (sagnorð)	103.136	17,47	3.858	9,55
t (töluorð)	5.901	1,00	753	1,86
x (ógreint)	127	0,02	51	0,13
Samtals	590.297	100,00	40.392	100,00

Gert var parað t-próf á hlutfalli rangt greindra orða til þess að kanna hvort tölfræðilega marktækur munur væri á árangri markaranna þriggja. Niðurstaða prófsins fyrir pörin fnTBL/TnT, MXPOST/TnT og fnTBL/MXPOST er sýnd í töflu 5. Munur á mörkurum er marktækur í öllum tilvikum.

Tafla 5. Parað t-próf á mismuni á hlutfalli rangt greindra orða

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

6.4.1 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. 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 markaskrána á þrenns konar hátt.

1. Atviksorð voru ekki greind, þ.e. aðeins var litið á fyrsta staf í greiningarstreng.
2. Samtengingar voru ekki greindar, þ.e. aðeins var litið á fyrsta staf í greiningarstreng.
3. Fornafnaflokkum var slegið saman en greining fornafna látin halda sér að öðru leyti.

Í töflu 6 er sýnd nákvæmni markaranna þegar mörk eru einfölduð á þennan hátt.

Tafla 6. Nákvæmni mörkunar þegar markaskrá er einfölduð

	Meðalnákvæmni fnTBL			Meðalnákvæmni MXPOST			Meðalnákvæmni TnT		
	Rétt (fj.)	%	Samanl. %	Rétt (fj.)	%	Samanl. %	Rétt (fj.)	%	Samanl. %
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	

Af töflunni sést að með því að sleppa greiningu atviksorða hækkar nákvæmni TnT úr 90,36% í 91,52%, villum fækkar um 12,02%. Með því að sleppa einnig greiningu samtenginga og slá saman fornafnaflokkum fer nákvæmni TnT í 91,83%.

Með því að einfalda markaskrá á þennan hátt er álitið að ekki verði dregið úr gagnsemi mörkunar fyrir tungutækniverkefni.

Ef aðeins er litið á greiningu eftir orðflokkum nær TnT 98,14% nákvæmni, MXPOST 97,27% nákvæmni og fnTBL 97,25% nákvæmni. Verið getur að greining eftir orðflokkum sé í einstaka tilvikum nægileg.

6.4.2 Áhrif mismunandi texta

Í könnunum á mörkun hefur komið í ljós að mismunandi textar gefa mismunandi niðurstöðu. Í þessari könnun var talin ástæða til þess að kanna sérstaklega áhrif texta í raunvísindakafla textasafns Orðtíðnibókarinnar á niðurstöður mörkunar. Valið var að nota TnT-markarann við þessa tilraun þar sem sá markari gefur besta nákvæmni og auk þess er einfaldast að nota þann markara. Búin voru til tvö ný söfn skráa þar sem hvert safn hafði 10 pör þjálfunar- og prófunarskráa. Í fyrra safninu var efnafræðitextum í raunvísindahluta textasafns Orðtíðnibókarinnar sleppt. Við það fækkaði lesmálsorðum í hverri skrá um u.þ.b. 1%. Í seinna safninu var öllum raunvísindakaflanum sleppt og við það fækkaði lesmálsorðum um u.þ.b. 10%. Í töflu 7 er sýnd niðurstaða mörkunarinnar.

Tafla 7. Nákvæmni mörkunar mismunandi texta

	Meðalfjöldi lesmálsorða í prófunarsafni			Meðalfjöldi rétt greindra orða			Meðalnákvæmni			Óþekkt orð
	Öll	Þekkt	Óþekkt	Öll	Þekkt	Óþekkt	Öll orð	Þekkt orð	Óþekkt orð	%
Allur textinn	59.030	54.991	4.039	53.340	50.448	2.892	90,36	91,74	71,62	6,84
Efnafr. texti fjarlægður	58.427	54.433	3.995	52.813	49.952	2.861	90,39	91,77	71,65	6,84
Raunv. texti fjarlægður	53.175	49.553	3.623	48.128	45.546	2.582	90,51	91,91	71,29	6,81

Hlutfall óþekktra orða lækkar við það að öllum raunvísindakaflanum er sleppt. Tölur um nákvæmni sýna meðaltöl fyrir öll 10 pörin. Nákvæmni mörkunar eykst

lítillega fyrir öll orð og einnig þekkt og óþekkt orð þegar efnafræðitexta er sleppt. Þegar öllum raunvísindakaflanum er sleppt eykst heildarnákvæmni og nákvæmni fyrir mörkun þekktra orða en nákvæmni fyrir mörkun óþekktra orða lækkar.

6.4.3 Nánari skoðun á nákvæmni og villum sem markarar gera

Búin var til sérstök skrá til þess að skoða villur sem markarar gera. Hér að framan var greint frá því að markararnir voru þjálfaðir og prófaðir á þeim 10 pörum sem lýst er í 6.1. Þá fást 10 mörkuð prófunarsöfn sem samanlögð hafa að geyma allan texta orðtíðnibókarinnar. Þessi prófunarsöfn voru lögð saman þannig að fyrir hvert orð í textasafninu mætti bera saman rétt mark og þau mörk sem markararnir úthlutuðu því orði.

Fyrir hvern greiningarstreng var reiknuð nákvæmni (*precision*, P), griphlutfall (*recall*, R) og F-gildi. Niðurstöður útreikninganna má sjá hér. P, R og F var einnig reiknað fyrir einfaldaða greiningarstrengi eins og lýst var í 6.4.1 og má sjá niðurstöður hér. Í töflu 8 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.

Tafla 8. Nákvæmni (P), griphlutfall (R) og F-gildi fyrir orðflokka

	Fjöldi	fnTBL			MXPOST			TnT		
Orðflokkar	í safni	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 (töluorð)	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

Í töflu 9 er griphlutfallið 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 orðflokk í safninu.

Tafla 9. 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

		fnTBL			MXPOST			TnT		
Orðflokkur	Fjöldi í safni	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 (töluorð)	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

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 8. 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.

6.4.4 Villur í orðgreiningu

Hér að framan hefur verið bent á að markararnir eiga erfiðara með að greina á milli mismunandi greiningarmynda innan sama orðflokks en að greina orðflokkinn rétt. Þess vegna var lögð nokkur áhersla á að skoða nánar villur sem markarar gera.

Villur má skoða á ýmsan hátt og verða hér sýnd nokkur afbrigði. Forritinu *μ-TBL* (Lager 1999) fylgir kerfi til þess að skoða villur sem markarar gera. Þar er lagt til að fyrst séu villur flokkaðar samkvæmt strengnum „mark er XX, ætti að vera YY“. Tíðnitafla er búin til fyrir hvern markara sem lýsir villunum á þann hátt. Villur sem fnTBL gerir má sjá hér. Villur sem MXPOST gerir má sjá hér og villur sem TnT gerir má sjá hér. Í töflu 10 sjást tuttugu algengustu villur sem hver markari gerir af þessu tagi. Algengustu villurnar sem allir markarar gera er að rugla saman fallstjórn atviksorða (sem eru reyndar forsetningar). Næst kemur ruglingur á milli beygingarmynda nafnorða sem hafa sömu mynd. Má þar nefna þolfall og þágufall kvenkynsorða í eintölu og nefnifall og þolfall

Tafla 10. Tuttugu algengustu villur sem hver markari gerir

fnTBL				MXPOST				TnT			
markari> rétt	Tíðni	%	S.lögð %	markari> rétt	Tíðni	%	S.lögð %	markari> rétt	Tíðni	%	S.lögð %
Alls	66.096	100,00		Samtals	64.434	100,00		Alls	56.894	100,00	
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

hvorugkynsorða í eintölu. Ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum er líka algengur þar sem þessar beygingarmyndir líta eins út. Einnig má nefna nafnhátt og þriðju persónu fleirtölu í nútíð en þessar beygingarmyndir líta eins út.

Eins og sést af töflu 10 gera markararnir misjafnlega margar villur. Þeir gera einnig misjafnlega margar gerðir af villum. TnT gerir 5.373 mismunandi villur, fnTBL gerir 5.897 mismunandi villur og MXPOST gerir 7.115 gerðir af villum. Þar sem markaskrá Orðtíðnibókarinnar er mjög stór er unnt að gera mjög margvíslegar villur. Fræðilega má gera 552*552 = 304.704 gerðir af villum ef tiltekið safn sem á að marka hefur 552 ólík mörk.

Af töflu 10 sést að fyrstu 20 villur sem TnT gerir skýra um 25% af villum sem markarinn gerir en fyrir fnTBL er þessi tala rúmlega 18% og rúmlega 17% fyrir MXPOST.

Einnig var búinn til orðstöðulykill fyrir rangt greind orð fyrir alla markarana. Orðstöðulykil fyrir fnTBL má sjá hér, fyrir MXPOST hér og fyrir TnT hér.

Í töflu 11 eru sýndar 20 algengustu sameiginlegar villur sem allir markarar gera. Þetta yfirlit sýnir hvaða atriði allir markararnir eiga í erfiðleikum með. Markararnir virðast eiga í erfiðleikum með að greina fallstjórn atviksorða (forsetninga) og einnig að greina á milli greiningarmynda sem hafa sömu orðmyndir eins og í ljós kom þegar hver markari var skoðaður sérstaklega. En það vekur athygli hér að næstalgengasta villan 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 kvenkynsnafnorðum. 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. Alla töfluna má skoða hér.

Tafla 11. Allir markarar gera sömu villur. Tuttugu algengustu villurnar

	Tíðni	%	Samanl. %
Samtals	13.055	100,00	100,00
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

Í töflum 12 og 13 er sýnt hvaða orð markararnir eiga auðveldast og erfiðast með að marka rétt. Í töflu 12 eru sýnd 30 algengustu lesmálsorð sem markararnir þrír marka allir rétt. Á þessum lista eru greinarmerki og smáorð (atviksorð (forsetningar) og samtengingar), persónufornöfn og greiningarmyndir af sögninni *vera*. Sum þessi lesmálsorð geta tekið fleiri en einn greiningarstreng eins og t.d. *hann* sem getur verið annað hvort nefnifall eða þolfall eintala af persónufornafninu *hann*. Nefnifallið er algengara en þolfallið, í um 87% tilvika er *hann* nefnifall í textum Orðtíðnibókarinnar.

Tafla 12. Þrjátíu algengustu lesmálsorðin sem allir markarar greindu eins og rétt

Orð	Mark	Tíðni	%	Samanl. %
Samtals		484.294	100,00	
.	.	33.181	6,85	6,85
og	c	22.173	4,58	11,43
,	,	22.083	4,56	15,99
að	cn	11.008	2,27	18,26
í	aþ	8.898	1,84	20,10
var	sfg3eþ	7.400	1,53	21,63
hann	fpken	6.809	1,41	23,03
á	aþ	6.163	1,27	24,31
ég	fp1en	6.040	1,25	25,55
ekki	aa	5.761	1,19	26,74
er	sfg3en	5.517	1,14	27,88
að	c	5.473	1,13	29,01
en	c	5.454	1,13	30,14
sem	ct	5.334	1,10	31,24
hún	fpven	4.854	1,00	32,24
um	ao	4.083	0,84	33,09
til	ae	4.059	0,84	33,92
það	fphen	3.838	0,79	34,72
«	«	3.597	0,74	35,46
»	»	3.567	0,74	36,20
af	aþ	3.252	0,67	36,87
við	ao	3.246	0,67	37,54
í	ao	3.230	0,67	38,20
á	ao	2.949	0,61	38,81
svo	aa	2.278	0,47	39,28
-	-	2.256	0,47	39,75
því	fpheþ	2.130	0,44	40,19
?	?	2.102	0,43	40,62
með	aþ	2.102	0,43	41,06
þegar	c	2.084	0,43	41,49

Í töflu 13 eru sýnd 30 lesmálsorð sem allir markarar greindu eins en rangt. Í þessum hópi eru atviksorð (forsetningar), samtengingar og fornöfn, aðallega persónufornöfn. Ruglingur í greiningu á þessum 30 orðum sýnir rugling í greiningu milli greiningarflokka en ekki rugling milli orðflokka. Í töflu 9 má sjá að markararnir eiga t.d. erfitt með að greina á milli greiningarflokka fornafna og kemur það greinilega fram í töflu 13.

Tafla 13. Þrjátíu algengustu lesmálsorð sem allir markarar greindu eins en rangt

Orð	Rétt mark	Rangt mark	Tíðni	%	Samanl. %
Samtals			13.055	100,00	
í	ao	aþ	250	1,91	1,91
sér	fpveþ	fpkeþ	167	1,28	3,19
sem	c	ct	137	1,05	4,24
í	aþ	ao	117	0,90	5,14
á	ao	aþ	112	0,86	6,00
sig	fpveo	fpkeo	102	0,78	6,78
það	fpheo	fphen	99	0,76	7,54
hann	fpkeo	fpken	88	0,67	8,21
um	aa	ao	77	0,59	8,80
með	ao	aþ	74	0,57	9,37
hvað	fsheo	fshen	72	0,55	9,92
þá	fpkfo	aa	72	0,55	10,47
til	aa	ae	71	0,54	11,01
á	aþ	ao	67	0,51	11,53
þeirra	fphfe	fpkfe	62	0,47	12,00
því	aa	fpheþ	61	0,47	12,47
það	fahen	fphen	56	0,43	12,90
þeim	fpkfþ	fphfþ	55	0,42	13,32
fyrir	aþ	ao	54	0,41	13,73
ekkert	foheo	fohen	47	0,36	14,09
okkur	fp1fo	fp1fþ	47	0,36	14,45
þetta	faheo	fahen	47	0,36	14,81
að	aa	c	44	0,34	15,15
sér	fpkfþ	fpkeþ	42	0,32	15,47
til	ae	aa	41	0,31	15,79
sig	fpkfo	fpkeo	40	0,31	16,09
þeim	fpvfþ	fphfþ	38	0,29	16,38
allt	foheo	fohen	37	0,28	16,67
sér	fpheþ	fpkeþ	37	0,28	16,95
það	fphen	fpheo	37	0,28	17,23

6.4.5 Frekari samanburður á mörkurum

Fundið var hversu oft markararnir voru sammála og fengu annað hvort rétta eða ranga niðurstöðu. Töflur 11, 12 og 13 gefa hugmyndir um í hvaða tilvikum allir markarar komast að réttri eða rangri niðurstöðu. Tafla 14 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 með því að sameina niðurstöður tveggja eða fleiri markara eins og gert veður hér á eftir.

Tafla 14. Hversu margir markarar eru sammála um rétt mark?

	Tíðni	%	Samanl. %
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

Í töflu 15 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 líkar þó að TnT gefi umtalsvert betri niðurstöðu. Það gæti því verið að unnt sé að bæta niðurstöðu TnT með niðurstöðu MXPOST.

Tafla 15. Samanburður á markarapörum

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

6.4.6 Hvernig má bæta niðurstöður mörkunar?

Margvíslegar aðferðir eru notaðar til þess að bæta niðurstöður mörkunar. Fer það eftir aðstæðum hvaða aðferð er vænlegust til árangurs. Í 6.4.1 er greint frá áhrifum marka-skrár á nákvæmni mörkunar. Þar voru sýndar niðurstöður af því að einfalda mörk á tiltekinn hátt.

Eins og sést t.d. af töflu 3 fæst mun minni nákvæmni við mörkun óþekktra orða, þ.e. orða sem koma ekki fyrir í þjálfunarsafni heldur en orða sem markarinn hefur þegar séð. Þess vegna er lögð áhersla á að bæta mörkun óþekktra orða þegar markarar eru búnir til. Það má gera með ýmsum aðferðum. Það er t.d. gert með því að bæta þær aðferðir sem markarar nota við mörkun óþekktra orða. Einnig má nota alls kyns skrár yfir nöfn og heiti fyrirtækja og stofnana, skammstafanir og fleira sem getur orðið til þess að bæta niðurstöðu mörkunar. Þar að auki eru útbúnar orðaskrár þar sem koma fyrir eins margar orðmyndir og kostur er og þeir greiningarstrengir sem geta fylgt hverri orðmynd. Nú er á lokastigi vinna við beygingarlýsingu íslensks nútímamáls. Sú vinna mun skila orðaskrá yfir beygingarmyndir rúmlega 170.000 orða. Ekki gafst kostur á að nota það orðasafn við mörkunartilraunina en gerð var önnur tilraun sem nú verður greint frá.

6.4.6.1 Áhrif orðasafns

Gerð var tilraun til þess að nota orðasafn við mörkun til þess að fá hugmynd um hversu mikið mætti bæta mörkun með því móti. Notuð voru forritin TnT og fnTBL þar sem þau gefa kost á að nota viðbótarorðasafn.

Gerður var listi yfir óþekkt orð í hverju prófunarsafni (koma ekki fyrir í viðkomandi þjálfunarsafni) og þau síðan sameinuð í eina skrá. Síðan var tekið annað hvert orð úr þessu safni og notað sem viðbótarorðasafn við mörkun með TnT og fnTBL. Með því móti ætti óþekktum orðum að fækka um u.þ.b. helming. Þetta viðbótarorðasafn var síðan notað við mörkun á öllum 10 prófunarsöfnunum. Í töflu 16 sést niðurstaða fyrir mörkun án orðasafns og með þessu orðasafni.

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 fækkar villum sem TnT-markarinn gerir um 12%. Þessar niðurstöður sýna því að mörkun mun batna ef unnt er að nota orðasafn.

Tafla 16. Þjálfun og mörkun 10 skráapara með TnT og fnTBL, orðasafn notað

		Óþekkt orð án orðas.	Nákvæmni án orðasafns			Nákvæmni með orðasafni*		
Markari	Par	%	Öll orð	Þekkt orð	Óþekkt orð	Öll orð	Þekkt orð	Óþekkt orð
TnT	01	7,57	89,81	91,44	69,90	91,16	91,63	85,39
	02	6,79	90,20	91,64	70,57	91,42	91,86	85,45
	03	6,88	90,37	91,74	71,92	91,52	91,89	86,49
	04	6,69	90,59	91,85	73,04	91,67	92,02	86,82
	05	6,51	90,74	92,01	72,45	91,80	92,17	86,54
	06	6,70	90,37	91,71	71,67	91,54	91,90	86,48
	07	6,70	90,64	92,03	71,32	91,84	92,25	86,22
	08	6,93	90,24	91,58	72,30	91,41	91,76	86,75
	09	6,80	90,10	91,52	70,68	91,35	91,75	85,87
	10	6,83	90,55	91,88	72,35	91,71	92,05	87,06
Meðaltal		6,84	90,36	91,74	71,62	91,54	91,93	86,31
Viðbót						1,18	0,19	14,69
fnTBL	01	7,57	88,32	91,10	54,40	89,66	91,20	70,86
	02	6,79	88,64	91,22	53,15	89,90	91,39	69,52
	03	6,88	88,79	91,41	53,29	90,13	91,58	70,51
	04	6,69	88,95	91,58	52,30	90,36	91,73	71,16
	05	6,51	89,46	91,86	55,04	90,52	91,85	71,35
	06	6,70	88,67	91,11	54,72	89,93	91,38	69,84
	07	6,70	88,93	91,54	52,59	90,14	91,66	69,00
	08	6,93	88,72	91,20	55,38	89,95	91,31	71,76
	09	6,80	88,57	91,15	53,07	89,89	91,35	69,93
	10	6,83	88,98	91,38	56,31	90,09	91,53	70,44
Meðaltal		6,84	88,80	91,36	54,02	90,06	91,50	70,44
Viðbót						1,25	0,14	16,41

* Orðasafnið er búið til úr helmingi þeirra orða sem álitin eru óþekkt frá sjónarhóli hvers prófunarsafns

6.4.6.2 Sameina niðurstöður markara

Í greinum um mörkun er getið nokkurra aðferða við að sameina niðurstöður tveggja eða fleiri markara til þess að ná fram meiri nákvæmni í mörkun.
Skipta má aðferðunum við að sameina markara í þrjá flokka.

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.

6.4.6.2.1 Kjósa á milli markara

Í (Halteren et al 2001) er viðamikið 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.

Höfundar benda á að mismunandi aðferðir við mörkun leiði til ólíkra villna. Með því að kjósa á milli markara segjast höfundar losna við svokallað „gang effect“. Þannig vonast höfundar til þess að losna við duttlunga og þar með villur sem verða vegna bjaga sem tiltekinn markari hefur. Enn betri leið, segja þeir, er að búa til svo kallað „arbiter effect“. Nýr markari (flokkari) er þjálfaður á grundvelli útkomna hinna markaranna (flokkaranna). Með þessu móti er ekki aðeins verið að vinna gegn bjaganum heldur reynt að nýta hann til þess að finna rétta útkomu. Höfundar setja fram þá tilgátu að báðar aðferðir gefi betri niðurstöðu en besti markarinn einn sér og ef nægilega mikið er til af þjálfunargögnum muni „arbiter“-aðferðin gefa betri niðurstöðu en kosning. Í greininni eru notaðar mismunandi mörkunaraðferðir sem allar eru látnar nota sömu gögn. Notaðir voru TBL-markari Brills, minnismarkarinn MBT, MXPOST-markari Ratnaparkhis og TnT, Markovsmarkari Brants.

Í því verki sem hér er lýst var gerð tilraun með fjögur afbrigði af kosningaaðferðinni. Einfaldasta aðferðin byggir 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 (*recall*) hvers marks. Vegið er með nákvæmni (e. *precision*) sem segir til um hvernig markarinn stendur sig og (*1-griphlutfall*) sem segir til um hve oft markaranum mistekst að finna rétta markið. Hvert mark fær *nákvæmni* þess markara sem leggur markið til og (*1-griphlutfall*) marksins hjá þeim mörkurum sem leggja það ekki til.

Í töflu 17 sést nánari samanburður á mörkurunum þremur.

Tafla 17. Samanburður á mörkurum

	Tíðni	%	Samanl. %
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	

Í töflu 18 sést niðurstaða af því að kjósa með mismunandi aðferðum milli marka sem markararnir þrír úthluta. Hæsta nákvæmni fæst með því að nota heildarnákvæmni sem vog. Er það sama niðurstaða og fékkst í (Halteren et al 2001) fyrir hollenskan texta. 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. 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 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 (precision) þess markara sem leggur markið til og (1-griphlutfall) marksins hjá þeim mörkurum sem leggja það ekki til.

Tafla 18. Nákvæmni þriggja markara og nákvæmni sem fæst með þremur mismunandi aðferðum við að kjósa á milli niðurstöðu markaranna og nákvæmni þegar greiningarstrengir eru einfaldaðir

		Óþekkt orð		Þekkt orð		Öll orð		Framyfir
Mörk	Aðferð	Tíðni	%	Tíðni	%	Tíðni	%	Tnt
	Alls	40.392	100,00	549.905	100,00	590.297	100,00	
Óbreytt	MXPOST	25.246	62,50	500.617	91,04	525.863	89,08	-1,28
Óbreytt	fnTBL	21.823	54,03	502.378	91,36	524.201	88,80	-1,56
Óbreytt	TnT	28.919	71,60	504.484	91,74	533.403	90,36	0,00
Óbreytt	Meirhlutakosning	27.889	69,05	510.903	92,91	538.792	91,27	0,91
Óbreytt	Kosn. með heildarnákv.	29.003	71,80	511.348	92,99	540.351	91,54	1,18
Óbreytt	Kosn. með nákv. marks	27.808	68,85	511.088	92,94	538.896	91,29	0,93
Óbreytt	Kosn. með nákv. og griphl.f.	28.738	71,15	511.440	93,01	540.178	91,51	1,15
Einfölduð	MXPOST	25.255	62,52	508.753	92,52	534.008	90,46	0,10
Einfölduð	fnTBL	21.831	54,05	509.309	92,62	531.140	89,98	-0,38
Einfölduð	TnT	28.925	71,61	513.173	93,32	542.098	91,83	1,47
Einfölduð	Meirhlutakosning	27.896	69,06	516.722	93,97	544.618	92,26	1,90
Einfölduð	Kosn. með heildarnákv.	29.009	71,82	517.283	94,07	546.292	92,55	2,18
Einfölduð	Kosn. með nákv. marks	27.816	68,87	516.956	94,01	544.772	92,29	1,93
Einfölduð	Kosn. með nákv. og griphl.f.	28.741	71,16	517.254	94,06	545.995	92,49	2,13
Einf. f. kosn.	Meirhlutakosning	27.898	69,07	516.778	93,98	544.676	92,27	1,91
Einf. f. kosn.	Kosn. með heildarnákv.	29.010	71,82	517.342	94,08	546.352	**92,56**	2,19
Einf. f. kosn.	Kosn. með nákv. marks	27.818	68,87	516.972	94,01	544.790	92,29	1,93
Einf. f. kosn.	Kosn. með nákv. og griphl.f.	28.743	71,16	517.247	94,06	545.990	92,49	2,13

Einföldun felst í að gera ekki greinarmun á atviksorðum og ekki heldur samtengingum.
Fornöfn eru sett í einn flokk en að öðru leyti er greining þeirra eftir kyni, tölu og falli látin haldast.

Eins og sést af töflu 17 gefur TnT oftar rétt mark en hinir markararnir þegar öll mörk eru ólík. Þegar vegið er með nákvæmni hvers marks fyrir hvern markara verður niðurstaðan ekki sú sama og þegar vegið er með heildarnákvæmni og enn þá önnur með síðustu aðferðinni þegar vegið er með nákvæmni og griphlutfalli hvers marks.

Í töflu 18 sést einnig niðurstaða af því að einfalda mörk. Sýnd er nákvæmni markaranna hvers fyrir sig þegar mörk eru einfölduð eins og lýst var í 6.4.1. Einnig er sýnd niðurstaða af því að einfalda mörk sem eru valin með kosningu og af því að kjósa um einfölduð mörk. Fyrir það textasafn sem hér er til skoðunar fæst best niðurstaða með því að kjósa á milli markaranna eftir að mörk eru einfölduð. Ná má 92,55% nákvæmni með því að kjósa um upprunaleg mörk og vega með heildarnákvæmni og einfalda síðan þau mörk sem þannig eru valin. Ef kosið er um einfölduð mörk með sömu aðferð fæst **92,56%** nákvæmni. Mismunurinn er ekki tölfræðilega marktækur.

6.4.6.2.2 Beita málfræðireglum

Lars Borin (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.

Lars Borin valdi þrjá markara til þessa að prófa. Þessir markarar áttu þegar að vera þjálfaðir á þýskum texta. Markararnir sem voru valdir voru *QTAG*, *TreeTagger* og *Morphy*. QTAG gaf nákvæmni undir 90% og var því ekki skoðaður frekar. Borin skoðaði niðurstöður markaranna tveggja sem eftir voru miðað við tvær markaskrár, upphaflega markaskrá sem hafði um 1000 mörk og minnkaða markaskrá. 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 (LB 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 mætti síðan nota til þess að bæta mörkun með því að sameina niðurstöður markaranna.

Fyrsta tilgátan var ekki prófuð. Í töflu 17 sést 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.

Lars Borin segir að hinar tvær tilgáturnar hafi verið staðfestar. Markararnir sem hann prófaði gerðu ólíkar villur og mismunurinn virtist kerfisbundinn.

Athugað var hvort nota mætti niðurstöðu MXPOST, fnTBL eða TnT til þess að bæta niðurstöðu sem fékkst með kosningu. Hæsta nákvæmni, 92,56%, fékkst þar sem kosið var um niðurstöðu einfaldaðra marka sem þrír markarar höfðu úthlutað. Í töflu 19 sést samanburður á þessari niðurstöðu og niðurstöðum markaranna þriggja eftir einföldun.

Á töflunni sést að niðurstöður MXPOST myndu bæta mestu við niðurstöðu með kosningu og gefa 94,22% 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.

Tafla 19. Samanburður á útkomu markaranna og niðurstöðu þar sem kosið er um einfölduð mörk og vegið með heildarnákvæmni

Alls	590.297	100,00	
Kosning vs. MXPOST	Tíðni	%	Samanl. %
Bæði mörk rétt	524.179	88,80	88,80
Kosning rétt, MXPOST rangt	22.173	3,76	92,56
Kosning röng, MXPOST rétt	9.829	1,67	94,22
Mörk lík og röng	24.267	4,11	98,33
Mörk ólík og röng	9.849	1,67	100,00
Kosning vs. fnTBL			
Bæði mörk rétt	525.756	89,07	89,07
Kosning rétt, fnTBL rangt	20.596	3,49	92,56
Kosning röng, fnTBL rétt	5.384	0,91	93,47
Mörk lík og röng	30.134	5,10	98,57
Mörk ólík og röng	8.427	1,43	100,00
Kosning vs. TnT			
Bæði mörk rétt	537.521	91,06	91,06
Kosning rétt, TnT rangt	8.831	1,50	92,56
Kosning röng, TnT rétt	4.577	0,78	93,33
Mörk lík og röng	37.425	6,34	99,67
Mörk ólík og röng	1.943	0,33	100,00

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.

Tafla 20 sýnir þau tilvik þar sem það að velja mark MXPOST fram yfir útkomu úr kosningu fækkar villum um meira en 8. 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 sem eru dregnar af þessum lista eru í forminu:

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

Í töflu 20 sést hluti af þeim reglum sem var beitt. Breyta má viðmiðun og velja færri eða fleiri reglur. Í aftasta dálkinum sést hver nákvæmni verður með því að beita reglunum í röð ef nákvæmni sem fæst þegar búið er að einfalda mörk og kjósa um niðurstöður markara er 92,56%. Heildarnákvæmni ætti að verða 92,82% ef valdar eru allar reglur sem bæta niðurstöðu um meira en 5 mörk.

Þegar reglunum var beitt og reglur valdar þannig að niðurstaða batnaði um meira en 5 mörk fékkst nákvæmni fyrir öll orð 92,82% eins og búist var við, nákvæmni fyrir óþekkt orð 72,15% og fyrir þekkt orð 94,34%. Þessi niðurstaða er marktækt hærri en 92,56%.

Einnig var gerð tilraun með að beita reglum þegar upprunaleg mörkun með TnT og fnTBL var gerð með aðstoð orðasafns eins og lýst er í 6.4.6.1. Mörk voru einfölduð og kosið um einfölduð mörk með því að vega með heildarnákvæmni markaranna. Niðurstaða úr þeirri kosningu er 93,53% nákvæmni. Fundnar voru reglur til þess að velja mark MXPOST umfram útkomu úr kosningu eins og lýst er hér að ofan. Þegar reglum var beitt fékkst 93,65% nákvæmni þegar reglur voru valdar þannig að

niðurstaða batnaði um meira en 5 mörk þegar þeim var beitt. Marktækur munur er á þessum tveimur niðurstöðum.

Tafla 20. Mismunur á útkomu MXPOST og niðurstöðu þar sem kosið er um einfölduð mörk og vegið með heildarnákvæmni

Kosning_mxp	Tíðni	Kosning rétt	mxp rétt	ólíkir rangir	mxp-kosning	viðbót %	viðbót %	Endanl. nákv.
Nákvæmni áður en reglum er beitt								92,56
nveþ_nveo	421	124	258	39	134	0,0227	0,0227	92,58
nveo_nveþ	471	158	278	35	120	0,0203	0,0430	92,60
nheog_nheng	223	74	146	3	72	0,0122	0,0552	92,61
nheo_nhen	375	134	201	40	67	0,0114	0,0666	92,62
nhen_nheo	316	109	170	37	61	0,0103	0,0769	92,63
nveþ_nven	165	41	98	26	57	0,0097	0,0866	92,64
nveo_nven	155	39	88	28	49	0,0083	0,0949	92,65
sfg1eþ_sfg3eþ	68	9	58	1	49	0,0083	0,1032	92,66
nvfn_nvfo	265	106	154	5	48	0,0081	0,1113	92,67
lheosf_lhensf	156	52	95	9	43	0,0073	0,1186	92,67
nven_nveþ	157	46	85	26	39	0,0066	0,1252	92,68
nven_nveo	163	52	87	24	35	0,0059	0,1311	92,69
lhensf_lheosf	221	85	120	16	35	0,0059	0,1370	92,69
nhfng_nhfog	67	16	49	2	33	0,0056	0,1426	92,70
lvfnsf_lvfosf	85	26	56	3	30	0,0051	0,1477	92,70
nheng_nheog	122	45	75	2	30	0,0051	0,1528	92,71
sþghen_ssg	131	47	77	7	30	0,0051	0,1579	92,71
sfg3eþ_sfg1eþ	97	30	59	8	29	0,0049	0,1628	92,72
nkeo-m_nken-m	82	21	49	12	28	0,0047	0,1675	92,72
nvfog_nvfng	64	18	46	0	28	0,0047	0,1723	92,73
nkeþ_nkeo	225	85	111	29	26	0,0044	0,1767	92,73
nvfng_nvfog	61	18	42	1	24	0,0041	0,1808	92,74
sfg3fn_sng	91	33	56	2	23	0,0039	0,1847	92,74
lvfosf_lvfnsf	73	26	47	0	21	0,0036	0,1882	92,74
nkfn_nkfo	42	10	31	1	21	0,0036	0,1918	92,75
nhfog_nhfng	57	18	39	0	21	0,0036	0,1953	92,75
nveþ_nvee	61	13	34	14	21	0,0036	0,1989	92,75
nvee_nveo	67	18	37	12	19	0,0032	0,2021	92,76
nveo-ö_nveþ-ö	34	8	25	1	17	0,0029	0,2050	92,76
nkfo_nkfn	40	11	28	1	17	0,0029	0,2079	92,76
nvee_nveþ	61	16	32	13	16	0,0027	0,2106	92,77
nhen_nheþ	43	8	24	11	16	0,0027	0,2133	92,77
sfg1en_sng	25	4	18	3	14	0,0024	0,2157	92,77
nhfn_nhfo	141	62	75	4	13	0,0022	0,2179	92,77
lhfnsf_lvensf	143	53	66	24	13	0,0022	0,2201	92,78
ssm_snm	27	5	18	4	13	0,0022	0,2223	92,78
lheþsf_lveovf	33	5	17	11	12	0,0020	0,2243	92,78
lhensf_ssg	48	10	22	16	12	0,0020	0,2263	92,78
nkeo_nkeþ	278	122	133	23	11	0,0019	0,2282	92,78
nhfn_nheo	57	10	21	26	11	0,0019	0,2301	92,79
lheovf_lhenvf	23	6	16	1	10	0,0017	0,2317	92,79
sþgven_ssg	16	2	12	2	10	0,0017	0,2334	92,79
nveo_nvee	66	20	30	16	10	0,0017	0,2351	92,79
nheo-ö_nhen-ö	28	8	17	3	9	0,0015	0,2367	92,79
nken-m_nkeo-m	127	43	52	32	9	0,0015	0,2382	92,79

6.4.7 Niðurstöður og tillögur

Hér á undan hefur verið greint frá tilraunum við að marka íslenskan texta með ýmsum aðferðum sem hafa verið þróaðar fyrir önnur tungumál. Þrír 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. Það virðist skipta máli í hvaða röð aðgerðunum er beitt. Í töflu 21 er gefið yfirlit yfir helstu niðurstöður. Í fyrri hluta töflunnar eru niðurstöður miðað við að hjálparorðasafn sé ekki notað við mörkun. Fyrst er textinn markaður með þremur mismunandi mörkurum. Síðan eru mörk einfölduð eins og greint var frá í 6.4.1. Eins og greint var frá í 6.4.6.2.1 fæst aðeins betri niðurstaða með því að kjósa um einfölduð mörk en að einfalda mörk sem koma út úr kosningu. Síðan er gefin niðurstaða sem fæst með því að beita tilteknum málfræðireglum og velja þannig niðurstöðu MXPOST

Tafla 21. Nákvæmni þriggja markara og nákvæmni sem fæst með þremur mismunandi aðferðum við að kjósa á milli niðurstöðu markaranna og nákvæmni þegar greiningarstrengir eru einfaldaðir. Að lokum er beitt reglum. Einnig niðurstöður miðað við að nota orðasafn.

	Orða-		Óþekkt orð		Þekkt orð		Öll orð		+
Mörk	safn[1]	Aðferð	Tíðni	%	Tíðni	%	Tíðni	%	Tnt
Alls			40.392	100,00	549.905	100,00	590.297	100,00	
Óbreytt	Nei	fnTBL	21.823	54,03	502.378	91,36	524.201	88,80	-1,56
Óbreytt	Nei	MXPOST	25.246	62,50	500.617	91,04	525.863	89,08	-1,28
Óbreytt	Nei	TnT	28.919	71,60	504.484	91,74	533.403	**90,36**	0,00
Óbreytt		Meirhlutakosning	27.889	69,05	510.903	92,91	538.792	91,27	0,91
Óbreytt		Kosn. v. með heildarnákv.	29.003	71,80	511.348	92,99	540.351	91,54	1,18
Óbreytt		Kosn. v. með nákv. marks	27.808	68,85	511.088	92,94	538.896	91,29	0,93
Óbreytt		Kosn. v. með nákv. og griphl.f.	28.738	71,15	511.440	93,01	540.178	91,51	1,15
Einfölduð		fnTBL	21.831	54,05	509.309	92,62	531.140	89,98	-0,38
Einfölduð		MXPOST	25.255	62,52	508.753	92,52	534.008	90,46	0,10
Einfölduð		TnT	28.925	71,61	513.173	93,32	542.098	**91,83**	1,47
Einfölduð		Meirhlutakosning	27.896	69,06	516.722	93,97	544.618	92,26	1,90
Einfölduð		Kosn. v. með heildarnákv.	29.009	71,82	517.283	94,07	546.292	92,55	2,18
Einfölduð		Kosn. v. með nákv. marks	27.816	68,87	516.956	94,01	544.772	92,29	1,93
Einfölduð		Kosn. v. með nákv. og griphl.f.	28.741	71,16	517.254	94,06	545.995	92,49	2,13
Einf. f. kosn.		Meirhlutakosning	27.898	69,07	516.778	93,98	544.676	92,27	1,91
Einf. f. kosn.		Kosn. v. með heildarnákv.	29.010	71,82	517.342	94,08	546.352	**92,56**	2,19
Einf. f. kosn.		Kosn. v. með nákv. marks	27.818	68,87	516.972	94,01	544.790	92,29	1,93
Einf. f. kosn.		Kosn. v. með nákv. og griphl.f.	28.743	71,16	517.247	94,06	545.990	92,49	2,13
		MXPOST fram yfir kosn. m. heildarnkv.	29.141	72,15	518.775	94,34	547.916	**92,82**	2,46
Óbreytt	Já	fnTBL	28.461	70,44	503.142	91,50	531.603	90,06	-0,30
Óbreytt	Nei	MXPOST	25.252	62,50	500.611	91,04	525.863	89,08	-1,28
Óbreytt	Já	TnT	34.859	86,28	505.511	91,93	540.370	91,54	1,18
Óbreytt		Meirhlutakosning	33.126	81,99	511.546	93,03	544.672	92,27	1,91
Óbreytt		Kosn. v. með heildarnákv.	34.331	84,97	512.044	93,12	546.375	92,56	2,20
Óbreytt		Kosn. v. með nákv. marks	33.374	82,60	511.771	93,07	545.145	92,35	1,99
Óbreytt		Kosn. v. með nákv. og griphl.f.	34.020	84,20	512.074	93,12	546.094	92,51	2,15
Einfölduð		fnTBL	28.467	70,46	509.788	92,71	538.255	91,18	0,82
Einfölduð		MXPOST	25.261	62,52	508.747	92,52	534.008	90,46	0,10
Einfölduð		TnT	34.863	86,29	513.797	93,44	548.660	92,95	2,58
Einfölduð		Meirhlutakosning	33.132	82,00	517.141	94,04	550.273	93,22	2,86
Einfölduð		Kosn. v. með heildarnákv.	34.336	84,98	517.720	94,15	552.056	93,52	3,16
Einfölduð		Kosn. v. með nákv. marks	33.380	82,62	517.395	94,09	550.775	93,30	2,94
Einfölduð		Kosn. v. með nákv. og griphl.f.	34.022	84,21	517.685	94,14	551.707	93,46	3,10
Einf. f. kosn.		Meirhlutakosning	33.131	82,00	517.189	94,05	550.320	93,23	2,87
Einf. f. kosn.		Kosn. v. með heildarnákv.	34.336	84,98	517.773	94,16	552.109	**93,53**	3,17
Einf. f. kosn.		Kosn. v. með nákv. marks	33.366	82,58	517.420	94,09	550.786	93,31	2,94
Einf. f. kosn.		Kosn. v. með nákv. og griphl.f.	34.023	84,21	517.699	94,15	551.722	93,47	3,10
		MXPOST fram yfir kosn. m. heildarnkv.	34.013	84,18	518.818	94,35	552.831	**93,65**	3,29

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.

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

umfram niðurstöðu kosningar. Í seinni hluta töflunnar eru sýndar sambærilegar niðurstöður miðað við að notað sé orðasafnið, sem var búið til, þegar markað er með TnT og fnTBL. Aftasti dálkurinn í töflunni sýnir viðbót í nákvæmni umfram niðurstöðu sem fæst með því að nota TnT án orðasafns. Hæsta nákvæmni sem fæst þegar búið er að einfalda mörk markaranna sem fást án orðasafns, kjósa á milli þeirra og beita reglum er 92,82%. Villum fækkar um 25% miðað við að nota TnT eingöngu. Ef notað er orðasafnið við mörkun með TnT og fnTBL og síðan beitt sömu aðferðum fæst 93,65% nákvæmni og villum fækkar um 34%.

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 stærðar 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.

7. 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 aðskilin lítil textasöfn voru notuð. 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 13 skáldverkum frá því eftir 1980, samtals 5.672 lesmálsorð að meðtöldum greinarmerkjum. Í þriðja safninu eru textar um tölvur og tækni sem eru teknir af vefsetri 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 fjármálráðuneytis og Morgunblaðinu (viðskipti), alls 2.776 lesmálsorð að meðtöldum greinarmerkjum. Aðalsteinn Eyþórsson tók þessa texta saman úr textasafni Orðabókar Háskólans og af vefsíðum.

Reynt var að fylgja reglum Orðtíðnibókar þegar lesmálsorð voru skilgreind. Það var tiltölulega einfalt fyrir bókmenntatextana en oft komu upp álitamál í textunum um tölvur og tækni og lögfræði og viðskipti. Við greiningu texta í lesmálsorð var notað forrit sem Geir Gunnarsson, nemi í tungutækni, hafði gert. Greining í lesmálsorð var síðan leiðrétt handvirkt. Einnig var notað forrit til þess að skipta textanum í setningar sem Auður Þórunn Rögnvaldsdóttir hafði gert fyrir texta Orðtíðnibókarinnar.

Textinn var ekki leiðréttur að öðru leyti en því að í stað *z* var sett *s* þar sem það átti við í fyrsta safninu sem hefur að geyma texta frá 19. öld og fyrri hluta 20. aldar. Þess vegna koma sennilega upp fleiri vandamál og óþekkt orð en í texta sem er leiðréttur.

Markararnir þrír voru þjálfaðir á öllum texta Orðtíðnibókarinnar. Textarnir fjórir voru síðan markaðir með því að nota niðurstöðu þjálfunarinnar. Ekki var notað viðbótarorðasafn. Einnig var kosið um mörkin með því að nota þá aðferð sem besta niðurstöðu gaf í tilraunum með texta Orðtíðnibókarinnar, þ.e. vega með heildarnákvæmni markaranna. Að lokum var beitt þeim málfræðireglum sem voru fundnar við tilraunir með texta Orðtíðnibókarinnar til þess að velja mark sem MXPOST úthlutar fram yfir það mark sem fæst með kosningu.

Síðan var fundin „rétt“ mörkun til þess að unnt væri að reikna út nákvæmni mörkunar með hinum ýmsu aðferðum. Reynt var að fylgja þeim reglum sem notaðar voru við greiningu á lesmálsorðum í textum Orðtíðnibókarinnar.

7.1 Mörkun bókmenntatexta frá 19. öld og fyrri hluta 20. aldar

Í töflu 22 sést niðurstaða mörkunar bókmenntatexta frá 19. öld og fyrri hluta 20. aldar. Fundin var nákvæmni mörkunar fyrir óþekkt orð, þekkt orð og öll orð fyrir alla markarana. Þá kemur í ljós að TnT gefur áberandi besta niðurstöðu eða 93,03% fyrir öll orð. Hlutfall óþekktra orða er nokkuð hærra en fannst fyrir texta Orðtíðnibókarinnar eða 8,70% á móti 6,84% að meðaltali fyrir prófunarsöfn sem búin voru til úr textum Orðtíðnibókarinnar. TnT-markarinn nær samt betri nákvæmni fyrir þessa texta en fyrir texta orðtíðnibókarinnar. Niðurstaðan versnar þegar kosið er um markarana. Ástæðan gæti verið sú að MXPOST og fnTBL fá mun verri niðurstöðu en TnT þannig að þeir geta ekki bætt niðurstöðu TnT. Niðurstaða versnar enn þegar reglum er beitt á

niðurstöðu kosningar. Ástæðan er sennilega sú að MXPOST nær mjög slökum árangri við mörkun textanna og getur því ekki bætt aðrar niðurstöður. Hins vegar fæst betri niðurstaða þegar mörk sem TnT úthlutar eru einfölduð. Einnig er sýnd nákvæmni markaranna þegar aðeins er litið til orðflokks.

Tafla 22. Nákvæmni sem fæst við mörkun gamals bókmenntatexta

	Óþ. orð	Óþekkt orð			Þekkt orð			Öll orð		
Aðferð	%	n	rétt	%	n	rétt		n	rétt	%
fnTBL	8,70	524	279	53,24	5.498	4.985	90,67	6.022	5.264	87,41
MXP	8,70	524	334	63,74	5.498	4.935	89,76	6.022	5.269	87,50
TnT	8,70	524	393	75,00	5.498	5.209	94,74	6.022	5.602	93,03
Kosning	8,70	524	378	72,14	5.498	5.222	94,98	6.022	5.600	92,99
Reglur	8,70	524	378	72,14	5.498	5.219	94,93	6.022	5.597	92,94
TnT, einf.	8,70	524	393	75,00	5.498	5.218	94,91	6.022	5.611	93,18
fnTBL, orðfl.	8,70	524	409	78,05	5.498	5.374	97,74	6.022	5.783	96,03
MXP, orðfl.	8,70	524	458	87,40	5.498	5.326	96,87	6.022	5.784	96,05
TnT, orðfl.	8,70	524	472	90,08	5.498	5.430	98,76	6.022	5.902	98,01

7.2 Mörkun bókmenntatexta frá því eftir 1980

Þessi texti er settur saman úr brotum úr 13 skáldverkum, samtals 5.672 lesmálsorð að meðtöldum greinarmerkjum. Reynt var að fylgja reglum Orðtíðnibókar um skilgreiningu lesmálsorða. Það var tiltölulega einfalt í þessu tilviki þar sem bókmenntatexti af þessari gerð hefur aðallega venjuleg orð. Í ljós kom að fjögur af textabrotunum voru líka í textum Orðtíðnibókarinnar. Niðurstöður í töflu 23 eru því sýndar fyrir allan textann, fyrir þau textabrot sem voru hluti af textasafni Orðtíðnibókarinnar og fyrir þá texta sem markararnir höfðu ekki séð áður. Gefin er nákvæmni fyrir þekkt og óþekkt orð og öll orð samanlagt.

TnT-markarinn náði bestum árangri eða 94,73% fyrir alla textana, 78,93% fyrir óþekkt orð og 95,55% fyrir þekkt orð. Nákvæmni versnaði þegar kosið var á milli markaranna og þegar reglum var beitt á niðurstöðu kosningar. Aðeins betri niðurstaða, 94,75% nákvæmni, fékkst þegar mörk sem TnT úthlutaði voru einfölduð. Það getur verið athyglisvert að skoða hvaða niðurstöðu markararnir ná við mörkun texta sem þeir hafa séð áður. TnT náði 98,66% nákvæmni en MXPOST aðeins 94,75% nákvæmni eða álíka nákvæmni og TnT náði fyrir alla textana. Niðurstaða fyrir óséða texta er nokkuð verri eða 92,91% fyrir TnT og hækkar ekki þó að mörk séu einfölduð.

Mikill breytileiki er milli markara. Markararnir raða sér einnig ólíkt því sem gerðist í tilrauninni með texta Orðtíðnibókarinnar. Í þessari tilraun sýnir TnT langbestu niðurstöðuna, síðan fnTBL og þá MXPOST. Þegar textar Orðtíðnibókar voru markaðir náðist betri niðurstaða með MXPOST en fnTBL. Geta markaranna til þess að marka texta sem þeir hafa séð áður getur gefið vísbendingu um hversu mikilli nákvæmni þeir nái þegar unnt er að finna leiðir til þess að fækka óþekktum orðum. TnT nær bestum árangri og fer aðeins í einu tilviki niður fyrir 98% nákvæmni.

Til gamans var nákvæmni reiknuð út sérstaklega fyrir hvert textabrot þó að það sé ekki sýnt hér. Bestum árangri ná allir markararnir við að marka texta Vigdísar Grímsdóttur (textabútur úr *Kaldaljósi*), TnT nær þar 99,20% nákvæmni. Versta niðurstöðu fá allir markararnir við mörkun texta Þórunnar Valdimarsdóttur (*Júlía*). TnT nær þar aðeins 89,74% nákvæmni sem er eina tilvikið þar sem nákvæmni fer niður fyrir 90%. TnT nær umtalsvert betri árangir við mörkun þeirra texta sem ekki koma fyrir í Orðtíðnibókinni, 92,91%, en við mörkun texta Orðtíðnibókarinnar sjálfrar (90,36%, mörkun án orðasafns). MXPOST og fnTBL fá hins vegar verri niðurstöðu, MXPOST 87,58% á móti 89,08% fyrir Orðtíðnibókina og fnTBL 88,10% á móti

88,80% (án orðasafns) fyrir texta Orðtíðnibókarinnar. Í töflu 23 er sýnt hvaða nákvæmni markararnir þrír ná fyrir óséða texta ef aðeins er litið á orðflokka. Bent skal á að hér var ekki notað viðbótarorðasafn en það gæti bætt niðurstöðuna umtalsvert.

Tafla 24. Nákvæmni sem fæst við mörkun texta um tölvur og tækni

		Óþ. orð	Óþekkt orð			Þekkt orð			Öll orð		
Safn	Aðferð	%	n	rétt	%	n	rétt	%	n	rétt	%
I	fnTBL	15,11	442	169	38,24	2484	2190	88,16	2.926	2.359	80,62
I	MXPOST	15,11	442	186	42,08	2484	2191	88,20	2.926	2.377	81,24
I	TnT	15,11	442	222	50,23	2484	2317	93,28	2.926	2.539	86,77
I	Kosning	15,11	442	218	49,32	2484	2323	93,52	2.926	2.541	86,84
I	Reglur	15,11	442	219	49,55	2484	2328	93,72	2.926	2.547	87,05
I	TnT einf.	15,11	442	222	50,23	2484	2317	93,28	2.926	2.539	86,77
I, Orðfl	fnTBL	15,11	442	356	80,54	2484	2437	98,11	2.926	2.793	95,45
I, Orðfl	MXPOST	15,11	442	364	82,35	2484	2410	97,02	2.926	2.774	94,81
I, Orðfl	TnT	15,11	442	395	89,37	2484	2453	98,75	2.926	2.848	97,33
II	fnTBL	14,76	426	165	38,73	2461	2182	88,66	2.887	2.347	81,30
II	MXPOST	14,76	426	178	41,78	2461	2179	88,54	2.887	2.357	81,64
II	TnT	14,76	426	211	49,53	2461	2303	93,58	2.887	2.514	87,08
II	Kosning	14,76	426	207	48,59	2461	2309	93,82	2.887	2.516	87,15
II	Reglur	14,76	426	208	48,83	2461	2314	94,03	2.887	2.522	87,36
II	TnT einf.	14,76	426	211	49,53	2461	2303	93,58	2.887	2.514	87,08
III	fnTBL	14,33	408	165	40,44	2439	2161	88,60	2.847	2.326	81,70
III	MXPOST	14,33	408	178	43,63	2439	2162	88,64	2.847	2.340	82,19
III	TnT	14,33	408	209	51,23	2439	2285	93,69	2.847	2.494	87,60
III	Kosning	14,33	408	205	50,25	2439	2289	93,85	2.847	2.494	87,60
III	Reglur	14,33	408	206	50,49	2439	2294	94,05	2.847	2.500	87,81
III	TnT einf.	14,33	408	209	51,23	2439	2285	93,69	2.847	2.494	87,60

I: öll orð
II: að frádregnum tölum
III: að frádregnum tölum og skammstöfunum

7.3 Mörkun texta um tölvur og tækni

Í töflu 24 sést niðurstaða mörkunar texta um tölvur og tækni. Fundin var nákvæmni fyrir öll orð, þekkt orð og óþekkt orð fyrir alla markarana fyrir öll lesmálsorð í textunum. Einnig er sýnd niðurstaða kosningar og þess að beita reglum á niðurstöðu kosningar. Enn fremur er sýnd nákvæmni fyrir einfölduð mörk sem TnT úthlutar. Einföldun hefur þar engin áhrif. Að lokum er sýnd nákvæmni markaranna ef aðeins er skoðaður orðflokkur (fyrsti stafur í greiningu). Nákvæmniútreikningar voru endurteknir með því að sleppa tölum (ritaðar með tölustöfum) og sleppa tölum og skammstöfunum.

Hlutfall óþekktra orða er frekar hátt, eða 15,11% í öllum textanum. Það skýrir að hluta lélega niðurstöðu. En hér eins fyrir bókmenntatextana stendur TnT-markarinn sig langbest. Nákvæmni sem TnT-markarinn nær við mörkun þekktra orða gæti bent til þess að sá markari næði viðunandi árangri ef unnt væri að fækka óþekktum orðum. Það að kjósa um markarana bætir niðurstöðu og það að beita reglunum bætir niðurstöðuna lítillega. Betri árangur næst ef tölum er sleppt og enn frekar ef skammstöfunum er sleppt.

Tafla 25. Nákvæmni sem fæst við mörkun texta um lögfræði og viðskipti

		Óþ. orð	Óþekkt orð			Þekkt orð			Öll orð		
Safn	Aðferð	%	n	rétt	%	n	rétt	%	n	rétt	%
I	fnTBL	14,05	390	213	54,62	2.386	2.041	85,54	2.776	2.254	81,20
I	MXPOST	14,05	390	236	60,51	2.386	2.042	85,58	2.776	2.278	82,06
I	TnT	14,05	390	284	72,82	2.386	2.174	91,11	2.776	2.458	88,54
I	Kosning	14,05	390	277	71,03	2.386	2.167	90,82	2.776	2.444	88,04
I	Reglur	14,05	390	279	71,54	2.386	2.176	91,20	2.776	2.455	88,44
I	TnT einf.	14,05	390	284	72,82	2.386	2.176	91,20	2.776	2.460	88,62
I	fnTBL, orðfl.	14,05	390	336	86,15	2.386	2.309	96,77	2.776	2.645	95,28
I	MXPOST, orðfl.	14,05	390	348	89,23	2.386	2.301	96,44	2.776	2.649	95,43
I	TnT, orðfl.	14,05	390	366	93,85	2.386	2.337	97,95	2.776	2.703	97,37
II	fnTBL	13,89	377	205	54,38	2.338	2.029	86,78	2.715	2.234	82,28
II	MXPOST	13,89	377	226	59,95	2.338	2.015	86,18	2.715	2.241	82,54
II	TnT	13,89	377	273	72,41	2.338	2.148	91,87	2.715	2.421	89,17
II	Kosning	13,89	377	266	70,56	2.338	2.144	91,70	2.715	2.410	88,77
II	Reglur	13,89	377	268	71,09	2.338	2.152	92,04	2.715	2.420	89,13
II	TnT einf.	13,89	377	273	72,41	2.338	2.150	91,96	2.715	2.423	89,24
II	fnTBL, orðfl.	13,89	377	326	86,47	2.338	2.273	97,22	2.715	2.599	95,73
II	MXPOST, orðfl.	13,89	377	336	89,12	2.338	2.265	96,88	2.715	2.601	95,80
II	TnT, orðfl.	13,89	377	353	93,63	2.338	2.297	98,25	2.715	2.650	97,61
III	fnTBL	14,09	376	205	54,52	2.293	2.002	87,31	2.669	2.207	82,69
III	MXPOST	14,09	376	226	60,11	2.293	1.978	86,26	2.669	2.204	82,58
III	TnT	14,09	376	272	72,34	2.293	2.117	92,32	2.669	2.389	89,51
III	Kosning	14,09	376	265	70,48	2.293	2.111	92,06	2.669	2.376	89,02
III	Reglur	14,09	376	267	71,01	2.293	2.119	92,41	2.669	2.386	89,40
III	TnT einf.	14,09	376	272	72,34	2.293	2.119	92,41	2.669	2.391	89,58
III	fnTBL, orðfl.	14,09	376	325	86,44	2.293	2.228	97,17	2.669	2.553	95,65
III	MXPOST, orðfl.	14,09	376	335	89,10	2.293	2.221	96,86	2.669	2.556	95,77
III	TnT, orðfl.	14,09	376	352	93,62	2.293	2.252	98,21	2.669	2.604	97,56

I: öll orð
II: að frá dregnum tölum
III: að frádregnum tölum og skammstöfunum

7.4 Mörkun texta um lögfræði og viðskipti

Í töflu 25 sést niðurstaða mörkunar á texta um lögfræði og viðskipti. Fundin var nákvæmni fyrir öll orð, þekkt orð og óþekkt orð fyrir alla markarana fyrir öll lesmálsorð í textunum. Einnig er sýnd niðurstaða kosningar og þess að beita reglum á niðurstöðu kosningar. Enn fremur er sýnd nákvæmni fyrir einfölduð mörk sem TnT úthlutar. Að lokum er sýnd nákvæmni markaranna ef aðeins er skoðaður orðflokkur (fyrsti stafur í greiningu). Nákvæmniútreikningar voru endurteknir með því að sleppa tölum (ritaðar með tölustöfum) og sleppa tölum og skammstöfunum (t.d., hf....).

Hlutfall óþekktra orða er frekar hátt, eða 14,05% í öllum textanum. Það skýrir að hluta lélega niðurstöðu. En hér eins og áður stendur TnT-markarinn sig langbest. Kosning um einfölduð mörk gefur verri niðurstöðu en TnT einn gefur. Það bætir hins vegar niðurstöðuna að beita reglum. Það vekur athygli að allir markararnir fá betri heildarniðurstöðu við mörkun lögfræði- og viðskiptatextans en við mörkun tölvutextans en lélegri niðurstöðu við mörkun þekktra orða. Skoða þarf nánar hvernig á þessu stendur. Betri árangur næst ef tölum er sleppt og enn frekar ef skammstöfunum er sleppt.

Tafla 25. Nákvæmni sem fæst við mörkun texta um lögfræði og viðskipti

		Óþ. orð	Þekkt orð		Þekkt orð				Öll orð		
Safn	Aðferð	%	n	rétt	%	n	rétt	%	n	rétt	%
I	TnT	14,05	390	284	72,82	2.386	2.174	91,11	2.776	2.458	88,54
I	MXPOST	14,05	390	236	60,51	2.386	2.042	85,58	2.776	2.278	82,06
I	fnTBL	14,05	390	213	54,62	2.386	2.041	85,54	2.776	2.254	81,20
I	Kosning	14,05	390	277	71,03	2.386	2.167	90,82	2.776	2.444	88,04
I	Reglur	14,05	390	279	71,54	2.386	2.176	91,20	2.776	2.455	88,44
I	TnT einf.	14,05	390	284	72,82	2.386	2.176	91,20	2.776	2.460	88,62
I	TnT, orðfl.	14,05	390	366	93,85	2.386	2.337	97,95	2.776	2.703	97,37
I	MXPOST, orðfl.	14,05	390	348	89,23	2.386	2.301	96,44	2.776	2.649	95,43
I	fnTBL, orðfl.	14,05	390	336	86,15	2.386	2.309	96,77	2.776	2.645	95,28
II	TnT	13,89	377	273	72,41	2.338	2.148	91,87	2.715	2.421	89,17
II	MXPOST	13,89	377	226	59,95	2.338	2.015	86,18	2.715	2.241	82,54
II	fnTBL	13,89	377	205	54,38	2.338	2.029	86,78	2.715	2.234	82,28
II	Kosning	13,89	377	266	70,56	2.338	2.144	91,70	2.715	2.410	88,77
II	Reglur	13,89	377	268	71,09	2.338	2.152	92,04	2.715	2.420	89,13
II	TnT einf.	13,89	377	273	72,41	2.338	2.150	91,96	2.715	2.423	89,24
II	TnT, orðfl.	13,89	377	353	93,63	2.338	2.297	98,25	2.715	2.650	97,61
II	MXPOST, orðfl.	13,89	377	336	89,12	2.338	2.265	96,88	2.715	2.601	95,80
II	fnTBL, orðfl.	13,89	377	326	86,47	2.338	2.273	97,22	2.715	2.599	95,73
III	TnT	14,09	376	272	72,34	2.293	2.117	92,32	2.669	2.389	89,51
III	MXPOST	14,09	376	226	60,11	2.293	1.978	86,26	2.669	2.204	82,58
III	fnTBL	14,09	376	205	54,52	2.293	2.002	87,31	2.669	2.207	82,69
III	Kosning	14,09	376	265	70,48	2.293	2.111	92,06	2.669	2.376	89,02
III	Reglur	14,09	376	267	71,01	2.293	2.119	92,41	2.669	2.386	89,40
III	TnT einf.	14,09	376	272	72,34	2.293	2.119	92,41	2.669	2.391	89,58
III	TnT, orðfl.	14,09	376	352	93,62	2.293	2.252	98,21	2.669	2.604	97,56
III	MXPOST, orðfl.	14,09	376	335	89,10	2.293	2.221	96,86	2.669	2.556	95,77
III	fnTBL, orðfl.	14,09	376	325	86,44	2.293	2.228	97,17	2.669	2.553	95,65

I: öll orð
II: að frá dregnum tölum
III: að frádregnum tölum og skammstöfunum

8. Framhald verkefnis

Lagt er til að þær niðurstöður, sem hér hafa verið kynntar, verði notaðar til þess að koma á fót stóru textasafni þar sem lesmálsorð eru greind í orðflokka og eftir beygingu. Slíkt textasafn er nauðsynlegt til þess að geta þróað margvísleg tungutækni-tól og til rannsókna á texta. Má þar m.a. nefna forrit fyrir setningagreiningu (*parsing*), orðabókagerð, leit í texta, talkennsl og talgervingu og þýðingarforrit. Stórt markað textasafn er einnig forsenda fyrir margvíslegum tíðnirannsóknum á texta.

Aðferðunum sem hér voru notaðar má beita á nýjan texta, leiðrétta niðurstöður og bæta nýja textanum síðan við það textasafn sem fyrir er. Aðferðirnar eru síðan endurbættar með aðstoð nýs textasafns og þeim beitt á nýjan texta og síðan koll af kolli. Einnig er nauðsynlegt að þróa aðferðir og tól til margvíslegrar forvinnslu textans, s.s. að greina hann í lesmálsorð, greina að málsgreinar, lesa úr skammstöfunum og tölum og greina margs konar sérnöfn. Einnig þarf að koma í notkun stóru orðasafni sem yrði búið til upp úr beygingarlýsingu íslensks máls sem gerð hefur verið fyrir styrk frá tungutækniverkefninu. Beygingarlýsingin hefur beygingarmyndir um 170.000 orða.

Með því verkefni sem nú er að ljúka telja höfundar að tekist hafi að sýna fram á að unnt er að greina íslenskan texta vélrænt. Það fer síðan eftir notkun textasafnsins hversu nákvæmlega textinn þarf að vera greindur og er þar bæði átt við hversu stór markaskráin þarf að vera og hversu mikillar nákvæmni í greiningu er krafist.

Hér á eftir er listi yfir verkþætti sem talið er að þurfi að vinna að til þess að koma megi upp slíku mörkuðu textasafni. Á þessu stigi er ekki gerð nein tilraun til þess að meta hversu mikinn tíma hver verkþáttur tekur.

8.1 Forvinnsla texta

Undirbúa þarf textann á margvíslega hátt fyrir mörkun. Textann þarf að greina í lesmálsorð og málsgreinar og aðgreina þarf fyrirsagnir. Þetta gæti virst tiltölulega einfalt og auðvelt og er það ef um er að ræða samfelldan bókmenntatexta. Fyrir margs konar nytjatexta, t.d. lagatexta og ýmsar skýrslur og tæknilegan texta, er ekki alltaf augljóst hvað eru lesmálsorð og hvar er upphaf og lok málsgreina. Við undirbúning orðtíðnibókarinnar voru lesmálsorð skilgreind en ekki málsgreinar. Skilgreining á lesmálsorðum sem notuð var í Orðtíðnibókinni var notuð við mörkun tilraunatexta í markaraverkefninu. Fara þarf vandlega yfir þær skilgreiningar og athuga hvort þær eigi við þegar stórt textasafn verður undirbúið. Má t.d. benda á að ekki er alltaf augljóst hvernig fara á með skammstafanir. Í Orðtíðnibókinni eru skammstafanir greindar í frumeiningar og hvert orð greint sérstaklega. Ekki er víst að þetta sé alltaf heppileg aðferð. Í tengslum við markaraverkefnið var fengin stór skammstafanaskrá frá Íslenskri málstöð og hún gerð tölvutæk. Sú skrá mun nýtast við gerð textasafnsins. Einnig getur orkað tvímælis hvernig skal fara með tölur. Í Orðtíðnibókinni eru tölur greindar í samræmi við það hvernig er lesið úr þeim. Til þess að það sé unnt við gerð stórs textasafns þarf að vera til forrit sem getur lesið úr tölum á þann hátt.

Við undirbúning forvinnslu þarf bæði að skilgreina nákvæmlega það sem á að gera og síðan að útbúa forrit til þess að framkvæma verkin.

8.2 Greining sérnafna

Markararnir sem verða notaðir styðjast í fyrstu við líkön sem eru búin til út frá texta Orðtíðnibókarinnar. Orðmyndir sem ekki koma fyrir þar eru því „óþekktar“ frá sjónarhóli markaranna. Þar á meðal eru ýmis sérnöfn sem geta verið mannanöfn, staðanöfn, nöfn á fyrirtækjum og félögum o.þ.h. Nauðsynlegt er að hafa yfir að ráða skrá yfir margvísleg slík heiti og allar beygingarmyndir þeirra. Einnig er stundum gripið til þess ráðs að bera kennsl á sérnöfn með tölfræðilegum aðferðum.

8.3 Greining óþekktra orða

Markarar þekkja aðeins þær orðmyndir sem koma fyrir í þeim þjálfunartexta sem þeir voru þjálfaðir á. Þess vegna er nauðsynlegt að hafa einnig yfir að ráða stóru orðasafni þar sem fyrir koma eins mörg orð og kostur er og allar beygingarmyndir þeirra. Á undaförnum mánuðum hefur verið unnið að beygingarlýsingu íslensks nútímamáls hjá Orðabók háskólans fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins. Ekki gafst í því verkefni, sem hér er lýst, tækifæri til þess að búa til úr lýsingunni orðasafn til þess að nota með mörkurum. En það er eitt af þeim verkefnum sem þarf að vinna til þess að unnt sé að marka stórt textasafn. Í beygingarlýsingunni eru beygingarmyndir um 170.000 orða, nafnorða, lýsingarorða og sagna og allra fornafna. En slík beyging-arlýsing getur aldrei orðið tæmandi. Orð bætast stöðugt við hina svo nefndu opnu orðflokka, þ.e. nafnorð, lýsingarorð og sagnir. Þess vegna er einnig nauðsynlegt að hafa aðferðir og forrit til þess að finna og greina samsett orð.

8.4 Finna texta

Gera þarf yfirlit yfir hvers konar textar þurfi að vera í textasafni eins og hér um ræðir. Finna þarf slíka texta og semja við eigendur þeirra um afnotarétt af þeim. Textarnir

þurfa að vera um margvísleg efni, skrifaðir í ýmsum stílum og frá ýmsum tímum. Ekki er alveg ljóst hversu stórt textasafnið þarf að vera. Stór textasöfn sem gerð hafa verið erlendis eru frá einni milljón til 100 milljón lesmálsorð. Venjulega er byrjað með lítið textasafn sem er handmarkað og síðan byggðar á því aðferðir til vélrænnar mörkunar eins og ráðgert er að gera hér. Líklega þarf að marka nokkuð stórt textasafn með þeim aðferðum sem þegar hafa verið þróaðar, leiðrétta mörkunina og byggja síðan endurbættar aðferðir á því textasafni sem þá verður til. Þetta er síðan endurtekið nokkrum sinnum.

Þau verk sem hér hefur verið lýst þarf að vinna hversu stórt textasafn sem ákveðið er að búa til. Stærð textasafnsins mun síðan ráðast af því hvaða textar eru aðgengilegir og hversu mikið fjármagn er tiltækt.

Heimildir

Auður Þórunn Rögnvaldsdóttir. 2002. The Icelandic µ-TBL Experiment: Templates for Icelandic. Term paper in NLP 1, GSLT.

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. 21-26.

Brants, Thorsten, 2000a. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000,* 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, E. 1994. Some Advances in Rule-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94).* 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.

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, Available from http://ilk.uvt.nl/downloads/pub/papers/ilk.0313.pdf

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, Available from http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf

Eiríkur Rögnvaldsson. 2002. The Icelandic µ-TBL Experiment: µ-TBL Rules for Icelandic Compared to English Rules. Term paper in NLP 1, GSLT.

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.

Jurafsky, Daniel, & James H. Martin. 2000. *Speech and Language Processing*. Prenitce-Hall, New Jersey.

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

Kristín Bjarnadóttir. 2002. The Icelandic µ-TBL Experiment: Preparing the Corpus. Term paper in NLP 1, GSLT.

Lager, Torbjörn. 1999. The µ-TBL System. User's manual. Version 0.9. http://wwww.ling.gu.se/~lager/mutbl.html

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

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics. 19(2),* 313–220.

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

Ngai, G. and Florian, R. 2001. Transformation -Based Learning in the Fast Lane. In *Proceedings of North American Chapter of ACL (NAACL-01),* pp 20–47. Carnegie Mellon University. Pittsburgh, PA, USA. June. ACL.

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). Philadelphia. PA, USA.

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.

Samuelson, Christer. 1993. Morphological tagging based entirely on Bayesian inference. In *9th Nordic Conference on Computational Linguistics NODALITA-93*. Stockholm University, Stockholm, Sweden.

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.

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

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

Voutilainen, A. (1995). Morphological disambiguation. In Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (Eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text,* pp. 165–285. Mouton de Gruyter, Berlin.

Sjöbergh, Jonas. 2003. Combining POS-taggers for improved accuracy on Swedish text. NoDaLiDa 03. Reykjavík. Óbirt handrit.

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

Viðauki A

Sniðmát fyrir samhengi orða og marka, notuð af fnTBL-markara.

pos_0 word_0 word_1 word_2 => pos
pos_0 word_-1 word_0 word_1 => pos
pos_0 word_0 word_-1 => pos
pos_0 word_0 word_1 => pos
pos_0 word_0 word_2 => pos
pos_0 word_0 word_-2 => pos
pos_0 word:[1,2] => pos
pos_0 word:[-2,-1] => pos
pos_0 word:[1,3] => pos
pos_0 word:[-3,-1] => pos
pos_0 word_0 pos_2 => pos
pos_0 word_0 pos_-2 => pos
pos_0 word_0 pos_1 => pos
pos_0 word_0 pos_-1 => pos
pos_0 word_0 => pos
pos_0 word_-2 => pos
pos_0 word_2 => pos
pos_0 word_1 => pos
pos_0 word_-1 => pos
pos_0 pos_-1 pos_1 => pos
pos_0 pos_1 pos_2 => pos
pos_0 pos_-1 pos_-2 => pos
pos_0 pos_1 => pos
pos_0 pos_-1 => pos
pos_0 pos_-2 => pos
pos_0 pos_2 => pos
pos_0 pos:[1,3] => pos
pos_0 pos:[1,2] => pos
pos_0 pos:[-3,-1] => pos
pos_0 pos:[-2,-1] => pos
pos_0 pos_1 word_0 word_1 => pos
pos_0 pos_1 word_0 word_-1 => pos
pos_-1 pos_0 word_-1 word_0 => pos
pos_-1 pos_0 word_0 word_1 => pos
pos_-2 pos_-1 pos_0 => pos
pos_-2 pos_-1 word_0 => pos
pos_1 word_0 word_1 => pos
pos_1 word_0 word_-1 => pos
pos_0 pos_1 pos_2 => pos
pos_0 pos_1 pos_2 word_1 => pos

Sniðmát til þess að greina óþekkt orð.

word => pos
pos word::~~5 => pos
pos word::5~~ => pos
pos word::~~4 => pos

pos word::4~~ => pos
pos word::4++ => pos
pos word::++4 => pos
pos word::--4 => pos
pos word::4-- => pos
pos word::~~3 => pos
pos word::3~~ => pos
pos word::++3 => pos
pos word::3++ => pos
pos word::3-- => pos
pos word::--3 => pos
pos word::~~2 => pos
pos word::2~~ => pos
pos word::--2 => pos
pos word::2-- => pos
pos word::++2 => pos
pos word::2++ => pos
pos word::~~1 => pos
pos word::1~~ => pos
pos word::++1 => pos
pos word::1++ => pos
pos word::--1 => pos
pos word::1-- => pos
pos word::1<> => pos
pos word^^1 => pos
pos word^^-1 => pos
word::5~~ => pos
word::~~5 => pos
word::4~~ => pos
word::~~4 => pos
word::4++ => pos
word::++4 => pos
word::--4 => pos
word::4-- => pos
word::~~3 => pos
word::3~~ => pos
word::++3 => pos
word::3++ => pos
word::3-- => pos
word::--3 => pos
word::~~2 => pos
word::2~~ => pos
word::--2 => pos
word::2-- => pos
word::++2 => pos
word::2++ => pos
word::~~1 => pos
word::1~~ => pos
word::1<> => pos
word::++1 => pos

word::1++ => pos
word::--1 => pos
word::1-- => pos
word^^-1 => pos
word^^1 => pos

Viðauki B

Skýring skammstafana í greiningarstrengjum 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

Viðauki C

Haustið 2005 var gerð tilraun til þess að þjálfa MBT-markarann á textum Orðtíðnibókarinnar (Sigrún Helgadóttir 2005). MBT er svokallaður minnismarkari (Memory-Based Tagger) (Daelemans *et al*. 2003) sem notar Tilburg Memory-Based Learner (*TiMBL*, Daelemans *et al*. 2004). Minnisaðferðin er ein gerð af vélrænu námi.

Námfúst forrit sem byggist í minnisaðferð lærir af mengi dæma sem eru geymd í gagnasafni og hvert dæmi hefur verið flokkað á tiltekinn hátt og hefur verið 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 ver 1 en í tilraunum með minnisaðferð er það eitt af markmiðunum að finna besta gildi fyrir *k*.

MBT-markarinn býr til aðskilin gagnasöfn fyrir þekkt orð og óþekkt orð. MBT-markarinn notar *TiMBL* kerfið sem notar svokallað IGTREE algrím fyrir þekkt orð og IB1 algrím fyrir óþekkt orð.

Gerðar voru tilraunir með mismunandi gildi á stikum sem MBT-forritið notar og val á mismunandi sérkennum (e. *features*). Besta niðurstaðan varð þessi:

Nákvæmni fyrir öll orð	87,19%
Nákvæmni fyrir þekkt orð	89,25%
Nákvæmni fyrir óþekkt orð	59,22%

Sérkenni sem voru skoðuð fyrir þekkt orð voru: ddwfa (mark orðsin sem á að skoða, tvö mörk til vinstri við orðið sem er skoðað, næsta orð til vinstri og markið næst til hægri). Fyrir óþekkt orð var skoðað: dFapsssssss (eitt mark til vinstri við orð sem er skoðað, eitt mark til hægri, fyrsti stafur í orði og sjö öftustu stafir í orði).

Þessi gildi á stikum voru valin:

	Þekkt orð	Óþekkt orð
Algrím	IGTREE	IB1
Distance metric	Overlap	Modified value difference metric
k	–	5
Feature-weighting	Shared variance	Gain Ratio